1.2. Limpieza y estandarización de datos

May 21, 2025

Creado por:

Isabel Maniega

1 1.2 Limpieza y estandarización de datos

1.1 1.2.1 Comprenda los datos estructurados y no estructurados y sus implicaciones en el análisis de datos (ídem 1.1.4)

Los datos **estructurados** (o "limpios") son los datos clásicos de una hoja de cálculo: todo está bien y limpio (no quiere decir que no pueda haber datos faltantes o un formato incorrecto) y los datos están organizados en una estructura similar a una tabla. Las bases de datos que almacenan este tipo de datos se denominan bases de datos relacionales: usamos SQL para administrar los datos en esas bases de datos. Ejemplos de datos estructurados son los archivos .csv o Excel.

Los **datos semiestructurados**, como sugiere el nombre, incorporan algunos elementos de datos estructurados, aunque no están organizados en una estructura tabular. Sin embargo, contienen etiquetas y elementos para organizar los datos de una manera significativa y crear jerarquías. Las bases de datos que almacenan datos semiestructurados se denominan bases de datos no relacionales, como MongoDB.

Datos no estructurados, que son la mayoría de los datos del mundo, son datos sin procesar que no siguen ningún esquema. Son los más ricos en información, pero en la mayoría de los casos deben limpiarse para que sean significativos. Algunos ejemplos de datos no estructurados son los archivos de video y audio, así como las fotos.

1.1.1 Limpieza y preprocesamiento de datos

La limpieza y el preprocesamiento de datos son pasos esenciales para garantizar que sus datos sean precisos y estén listos para el análisis.

- **Manejo de datos faltantes:** Desarrolle estrategias para lidiar con los datos faltantes, como la imputación o la eliminación, según la naturaleza de sus datos y los objetivos de la investigación.
- **Detección de valores atípicos:** Identifique y aborde los valores atípicos que pueden sesgar los resultados del análisis. Considere si los valores atípicos deben corregirse, eliminarse o conservarse según su importancia.
- **Normalización y escalamiento:** Normalice o escale los datos para que estén dentro de un rango común, haciéndolos adecuados para ciertos algoritmos y modelos.
- **Transformación de datos:** Aplique transformaciones de datos, como escalamiento logarítmico o codificación categórica, para preparar los datos para tipos específicos de análisis.

- **Desequilibrio de datos:** Aborde problemas de desequilibrio de clases en conjuntos de datos, en particular aplicaciones de aprendizaje automático, para evitar un entrenamiento de modelos sesgado.

1.1.2 Análisis exploratorio de datos (EDA)

EDA es el proceso de explorar visual y estadísticamente sus datos para descubrir patrones, tendencias y posibles perspectivas.

- **Estadísticas descriptivas:** Calcule estadísticas básicas como la media, la mediana y la desviación estándar para resumir las distribuciones de datos.
- **Visualización de datos:** Cree visualizaciones como histogramas, diagramas de dispersión y mapas de calor para revelar relaciones y patrones dentro de los datos.
- **Análisis de correlación:** Examine las correlaciones entre variables para comprender cómo se influyen entre sí.
- **Prueba de hipótesis:** Realice pruebas de hipótesis para evaluar la importancia de las diferencias o relaciones observadas en sus datos.

1.1.3 Técnicas de análisis estadístico

Elija las técnicas de análisis estadístico adecuadas en función de sus preguntas de investigación y tipos de datos.

- **Estadística descriptiva:** Utilice la estadística descriptiva para resumir y describir sus datos, proporcionando una descripción general inicial de las características clave.
- **Estadística inferencial:** Aplique la estadística inferencial, incluidas las pruebas t, ANOVA o el análisis de regresión, para probar hipótesis y extraer conclusiones sobre los parámetros de la población.
- **Pruebas no paramétricas:** Emplee pruebas no paramétricas cuando no se cumplan los supuestos de normalidad o cuando trabaje con datos ordinales o nominales.
- **Análisis de series temporales:** Analice datos de series temporales para descubrir tendencias, estacionalidad y patrones temporales.

1.1.4 Visualización de datos

La visualización de datos es una herramienta poderosa para transmitir información compleja en un formato digerible.

- **Gráficos y gráficos:** Utilice varios gráficos y gráficos, como gráficos de barras, gráficos de líneas, gráficos circulares y mapas de calor, para representar los datos visualmente.
- **Paneles interactivos:** Cree paneles interactivos con herramientas como Tableau, Power BI o aplicaciones web personalizadas para permitir que las partes interesadas exploren los datos de forma dinámica.
- **Narrativa:** Utilice la visualización de datos para contar una historia convincente basada en datos, destacando los hallazgos y las perspectivas clave.
- **Accesibilidad:** Asegúrese de que las visualizaciones de datos sean accesibles para todos los públicos, incluidos aquellos con discapacidades, siguiendo las pautas de accesibilidad.

1.1.5 Extraer conclusiones y perspectivas

Por último, extraer conclusiones y perspectivas de su análisis de datos es el objetivo final.

- **Interpretación contextual:** Interprete sus hallazgos en el contexto de sus objetivos de investigación y el panorama empresarial o de investigación más amplio.
- **Perspectivas prácticas:** Identifique perspectivas prácticas que puedan informar la toma de decisiones, el desarrollo de estrategias o las direcciones futuras de investigación.
- **Generación de informes:** Cree informes o presentaciones integrales que comuniquen sus hallazgos de forma clara y concisa a las partes interesadas.
- **Validación:** Verifique sus conclusiones con expertos en el área o especialistas en la materia para garantizar la precisión y la relevancia.

Si sigue estos pasos en el análisis e interpretación de datos, podrá transformar los datos sin procesar en información valiosa que impulse decisiones informadas, optimice los procesos y cree nuevas oportunidades para su organización.

1.2 ¿Cómo informar y presentar datos?

Ahora, exploremos los pasos cruciales para informar y presentar datos de manera eficaz, asegurándose de que sus hallazgos se comuniquen de manera clara y significativa a las partes interesadas.

1.2.1 1. Cree informes de datos

Los informes de datos son la culminación de sus esfuerzos de análisis de datos, ya que presentan sus hallazgos de una manera estructurada y comprensible.

- **Estructura del informe:** Organice su informe con una estructura clara, que incluya una introducción, metodología, resultados, discusión y conclusiones.
- **Integración de visualización:** Incorpore visualizaciones de datos, cuadros y gráficos para ilustrar los puntos y tendencias clave.
- **Claridad y concisión:** Utilice un lenguaje claro y conciso, evitando la jerga técnica, para que su informe sea accesible para una audiencia diversa.
- **Información procesable:** Resalte la información procesable y las recomendaciones que las partes interesadas pueden usar para tomar decisiones informadas.
- **Apéndices:** Incluya apéndices con la metodología detallada, las fuentes de datos y cualquier información adicional que respalde sus hallazgos.

1.2.2 2. Aproveche las herramientas de visualización de datos

Las herramientas de visualización de datos pueden mejorar significativamente su capacidad para transmitir información compleja de manera eficaz. Las principales herramientas de visualización de datos incluyen:

- **Tableau:** Tableau ofrece una amplia gama de opciones de visualización y paneles interactivos, lo que lo convierte en una opción popular para los profesionales de datos.
- **Power BI:** Power BI de Microsoft ofrece potentes capacidades de visualización de datos e inteligencia empresarial, adecuadas para crear informes y paneles dinámicos.
- **Bibliotecas de Python:** Utilice bibliotecas de Python como Matplotlib, Seaborn y Plotly para realizar visualizaciones y análisis de datos personalizados.
- **Excel:** Microsoft Excel sigue siendo una herramienta versátil para crear gráficos básicos, en particular para conjuntos de datos más pequeños.
- **Desarrollo personalizado:** Considere el desarrollo personalizado para necesidades de visualización especializadas o cuando las herramientas existentes no cumplan con sus requisitos.

1.2.3 3. Comunicar los hallazgos a las partes interesadas

Comunicar eficazmente los hallazgos a las partes interesadas es esencial para impulsar la acción y la toma de decisiones.

- **Comprensión de la audiencia:** Adapte su comunicación a las necesidades específicas y al conocimiento previo de su audiencia. Evite la jerga técnica cuando hable con partes interesadas no técnicas.
- **Narración visual:** Elabore una narrativa que guíe a las partes interesadas a través de los datos, destacando los conocimientos clave y sus implicaciones.
- **Participación:** Utilice presentaciones o informes atractivos e interactivos para mantener el interés de la audiencia y alentar la participación.
- **Manejo de preguntas:** Esté preparado para responder preguntas y brindar aclaraciones durante las presentaciones o debates. Anticipe posibles inquietudes u objeciones.
- **Ciclo de retroalimentación:** Fomente la retroalimentación y el diálogo abierto con las partes interesadas para garantizar que sus hallazgos se alineen con sus objetivos y expectativas.

1.3 Ejemplos de recopilación de datos

Para comprender mejor la aplicación práctica de la recopilación de datos en varios dominios, exploremos algunos ejemplos del mundo real, incluidos aquellos en el contexto empresarial. Estos ejemplos ilustran cómo la recopilación de datos puede impulsar la toma de decisiones informada y generar información significativa.

1.3.1 Encuestas de opinión de clientes comerciales

Escenario: una empresa minorista desea mejorar la experiencia de sus clientes y mejorar la oferta de productos. Para lograrlo, inicia encuestas de opinión de clientes.

Enfoque de recopilación de datos:

- **Creación de encuestas:** La empresa diseña una encuesta con preguntas específicas sobre las preferencias de los clientes, las experiencias de compra y la satisfacción con el producto.
- **Distribución:** Las encuestas se distribuyen a través de varios canales, incluido el correo electrónico, los quioscos en las tiendas y el sitio web de la empresa.
- **Recopilación de datos:** Las respuestas de miles de clientes se recopilan y almacenan en una base de datos centralizada.

Análisis de datos y perspectivas:

- **Análisis de sentimiento del cliente:** Mediante técnicas de procesamiento del lenguaje natural (PLN), la empresa analiza las respuestas abiertas para medir el sentimiento del cliente.
- **Rendimiento del producto:** Al analizar los datos de la encuesta, la empresa identifica qué productos reciben las calificaciones más altas y más bajas, lo que lleva a tomar decisiones sobre qué productos mejorar o discontinuar.
- **Optimización del diseño de la tienda:** Al examinar los comentarios relacionados con las experiencias en la tienda, la empresa puede ajustar el diseño y la señalización de la tienda para mejorar el flujo y la comodidad de los clientes.

1.3.2 Digitalización de registros de pacientes en el sector sanitario

Escenario: Un centro sanitario pretende pasar de registros de pacientes en papel a registros digitales para mejorar la eficiencia y la atención al paciente.

Enfoque de recopilación de datos:

- **Escaneo e ingreso de datos:** Los registros en papel existentes se escanean y el personal de ingreso de datos los convierte a formato digital.
- **Implementación de registros médicos electrónicos (EHR):** El centro adopta un sistema EHR para almacenar y gestionar los datos de los pacientes de forma segura.
- **Ingreso de datos continuo:** A medida que se recopila nueva información del paciente, se ingresa directamente en el sistema EHR.

Análisis de datos e información:

- **Acceso al historial del paciente:** Los médicos y las enfermeras obtienen acceso instantáneo a los registros de los pacientes, lo que mejora la precisión del diagnóstico y el tratamiento.
- **Análisis de datos:** Los datos agregados de los pacientes se pueden analizar para identificar tendencias en enfermedades, resultados del tratamiento y utilización de recursos sanitarios.
- **Optimización de recursos:** El análisis de los datos de los pacientes permite que el centro asigne recursos de manera más eficiente, como la programación del personal en función de los patrones de admisión de los pacientes.

1.3.3 Monitoreo de la participación en las redes sociales

Escenario: Una agencia de marketing digital administra campañas en las redes sociales para varios clientes y desea realizar un seguimiento del rendimiento de la campaña y la participación de la audiencia.

Enfoque de recopilación de datos:

- **Herramientas de monitoreo de redes sociales:** La agencia emplea herramientas de monitoreo de redes sociales para recopilar datos sobre la participación en las publicaciones, el alcance, los "Me gusta", las publicaciones compartidas y los comentarios.
- **Enlaces de seguimiento personalizados:** Se crean enlaces de seguimiento únicos para cada campaña para monitorear el tráfico y las conversiones.
- **Datos demográficos de la audiencia:** Los datos sobre los datos demográficos de los usuarios que participan se recopilan a partir de los análisis de la plataforma.

Análisis de datos e información:

- **Efectividad de la campaña:** La agencia evalúa qué campañas son más efectivas en términos de participación y tasas de conversión.
- **Segmentación de la audiencia:** La información sobre los datos demográficos de la audiencia ayuda a adaptar las campañas futuras a los datos demográficos objetivo específicos.
- **Estrategia de contenido:** Analizar qué tipos de contenido (por ejemplo, videos, infografías) generan más interacción es una herramienta que permite tomar decisiones estratégicas.

Estos ejemplos muestran cómo la recopilación de datos sirve como base para la toma de decisiones informada y el desarrollo de estrategias en diversos sectores. Ya sea para mejorar las experiencias de los clientes, optimizar los servicios de atención médica u optimizar los esfuerzos de market-

ing, la recopilación de datos permite a las organizaciones aprovechar información valiosa para el crecimiento y la mejora.

1.3.4 Análisis de datos no estructurados

El análisis y la interpretación adecuados de distintos tipos de datos, como audio, imágenes, texto y vídeo, implican el uso de tecnologías avanzadas: aprendizaje automático e inteligencia artificial. Las técnicas impulsadas por el aprendizaje automático, como el procesamiento del lenguaje natural (PLN), el análisis de audio y el reconocimiento de imágenes, son fundamentales para descubrir conocimientos y perspectivas ocultas.

El **procesamiento del lenguaje natural (NLP)**, un subcampo de la inteligencia artificial, es una técnica que facilita la comprensión, interpretación y generación del lenguaje humano por parte de las computadoras. Se utiliza principalmente para analizar datos no estructurados basados en texto, como correos electrónicos, publicaciones en redes sociales y reseñas de clientes.

La clasificación de texto, una técnica central del NLP, simplifica la organización y categorización del texto para facilitar su comprensión y utilización. Esta técnica permite realizar tareas como etiquetar la importancia o identificar comentarios negativos en los comentarios. El análisis de sentimientos, una aplicación común de clasificación de texto, categoriza el texto en función de los sentimientos, juicios u opiniones del autor. Esto permite a las marcas comprender la percepción de la audiencia, priorizar las tareas de servicio al cliente e identificar las tendencias de la industria.

Otro enfoque del NLP para manejar datos de texto no estructurados es la extracción de información (IE). IE recupera información predefinida, como nombres, fechas de eventos o números de teléfono, y la organiza en una base de datos. IE, un componente vital del procesamiento inteligente de documentos, emplea procesamiento del lenguaje natural y visión artificial para extraer automáticamente datos de varios documentos, clasificarlos y transformarlos en un formato de salida estandarizado.

El **reconocimiento de imágenes** identifica objetos, personas y escenas dentro de imágenes. Es muy beneficioso para analizar datos visuales como fotografías e ilustraciones. Las técnicas de reconocimiento de imágenes, como la detección de objetos, permiten a las organizaciones reconocer contenido generado por el usuario, analizar imágenes de productos y extraer textos de documentos escaneados para su posterior análisis.

El **análisis de video** implica la extracción de información significativa de los datos de video, como la identificación de patrones, objetos o actividades dentro del metraje. Esta tecnología puede cumplir numerosos propósitos, incluidos la seguridad y la vigilancia, el análisis del comportamiento del cliente y el control de calidad en la fabricación. Las técnicas, como la detección de movimiento, el seguimiento de objetos y el reconocimiento de actividades, permiten a las organizaciones obtener información sobre sus operaciones, clientes y amenazas potenciales.

Las herramientas de **análisis de audio** pueden procesar y analizar datos de audio, incluidas grabaciones de voz, música y sonidos ambientales, para extraer información útil o identificar patrones. Las técnicas de análisis de audio, como el reconocimiento de voz, la detección de emociones y la identificación de hablantes, se utilizan en múltiples industrias, como el entretenimiento (generación de contenido, recomendación musical), la atención al cliente (análisis de centros de llamadas, asistentes de voz) y la seguridad (biometría de voz, detección de eventos acústicos).

Si su proyecto de datos requiere la creación de modelos de ML personalizados, puede optar por plataformas específicas para cada tarea que lo ayuden a descubrir de manera eficaz patrones, ten-

dencias y relaciones a partir de datos no estructurados. Muchas plataformas de inteligencia artificial y aprendizaje automático brindan capacidades para procesar y analizar varios tipos de datos no estructurados, como texto, audio e imágenes, que se pueden utilizar para crear e implementar modelos de inteligencia artificial. Por ejemplo, puede crear o entrenar sus propios modelos de ML con los que se enumeran a continuación. Sin embargo, requieren contar con un equipo de ciencia de datos para entrenar modelos en sus datos.

- TensorFlow es un marco de aprendizaje automático de código abierto que admite muchos algoritmos de aprendizaje automático y profundo. Tiene la capacidad de procesar tipos de datos no estructurados y ofrece un amplio conjunto de bibliotecas y herramientas para crear, entrenar e implementar modelos de IA.
- IBM Watson es una colección de servicios y herramientas de IA que incluyen procesamiento de lenguaje natural, análisis de sentimientos y reconocimiento de imágenes, entre otros, para manejar datos no estructurados. Proporciona una variedad de modelos y API prediseñados, así como herramientas para crear modelos personalizados, lo que facilita la integración de capacidades de IA en sistemas ya existentes.

Por último, pero no por ello menos importante, es posible que deba aprovechar el etiquetado de datos si entrena modelos para tareas personalizadas. En un sentido práctico, el etiquetado de datos implica anotar o etiquetar datos sin procesar, como texto, imágenes, video o audio, con información relevante que ayuda a los modelos de aprendizaje automático a aprender patrones y realizar tareas específicas con precisión.

Por ejemplo, al entrenar modelos de NLP para el análisis de sentimientos, los anotadores humanos etiquetan muestras de texto con su sentimiento correspondiente, como positivo, negativo o neutral. De manera similar, los anotadores etiquetan objetos o regiones dentro de imágenes en el reconocimiento de imágenes para ayudar a los modelos a aprender a detectarlos y clasificarlos correctamente. En el análisis de video, el etiquetado de datos puede implicar etiquetar objetos, rastrear su movimiento o identificar actividades específicas. Finalmente, para el análisis de audio, el etiquetado puede incluir la transcripción del habla, la identificación de hablantes o el marcado de eventos específicos dentro del audio.

1.4 1.2.2 Identificar, rectificar o eliminar datos erróneos

1.4.1 Introducción a las técnicas avanzadas de limpieza de datos

La limpieza de datos es un paso fundamental en el proceso de análisis de datos. Implica detectar y corregir errores y anomalías para mejorar la calidad de los datos. Las técnicas avanzadas como la corrección de errores y la detección de anomalías pueden mejorar significativamente los resultados del análisis.

Los errores de datos se refieren a imprecisiones como valores faltantes, duplicados, problemas de formato, valores atípicos, etc. Identificar y corregir estos errores es clave para un análisis confiable. Los métodos comunes de corrección de errores incluyen:

- Validación de datos para verificar el formato, los rangos, las relaciones, etc.
- Manejo de datos faltantes mediante eliminación o imputación.
- Identificación y eliminación de registros duplicados. Detección y tratamiento de valores atípicos.

Estas técnicas mejoran la consistencia y precisión de los datos para tareas como el modelado de

aprendizaje automático.

1.4.2 El papel de la detección de anomalías en el análisis de datos

Una anomalía se refiere a un punto de datos que se desvía significativamente de los patrones esperados. La detección de anomalías permite identificar posibles problemas y eventos inusuales.

Técnicas como las puntuaciones Z y la agrupación ayudan a detectar valores atípicos. La experiencia en el dominio guía la evaluación de si las anomalías son errores o hallazgos significativos que vale la pena investigar más a fondo mediante análisis predictivo.

La detección eficaz de anomalías mejora la fiabilidad del análisis al permitir el tratamiento adecuado de los valores atípicos.

1.4.3 Dimensiones de la calidad de los datos y su impacto en el análisis

La integridad, la validez, la precisión, la coherencia, etc. son dimensiones clave de la calidad de los datos. Los problemas en estas áreas pueden socavar el análisis.

Por ejemplo, los datos incompletos conducen a modelos sesgados. Los tipos de datos no válidos crean errores de procesamiento. Los datos inexactos provocan información incorrecta.

Por lo tanto, la aplicación de controles de calidad antes del análisis es fundamental para evitar resultados erróneos y malas decisiones.

1.4.4 Aplicaciones del mundo real de la limpieza avanzada de datos

La limpieza sofisticada de datos permitió la predicción temprana de brotes de enfermedades al corregir errores de notificación en los datos de casos de infección. Las estrategias de negociación algorítmica mejoraron enormemente mediante la detección de anomalías en las métricas financieras y la corrección automática de datos.

Las técnicas avanzadas permiten realizar análisis de alta calidad en ámbitos como la atención sanitaria, las finanzas, el transporte, el comercio minorista, etc. Los conocimientos obtenidos en última instancia impulsan decisiones y políticas impactantes.

1.4.5 ¿Qué tipos de errores se deben limpiar en el paso de limpieza de datos?

La limpieza de datos tiene como objetivo identificar y corregir los siguientes errores de datos comunes:

- **Valores faltantes:** Campos que están en blanco o contienen valores nulos. Estos pueden distorsionar los resultados del análisis.
- **Valores atípicos:** Puntos de datos que son extremadamente altos o bajos en comparación con el resto del conjunto de datos. Estos pueden ser legítimos, pero a menudo indican errores.
- **Duplicados:** El mismo registro de datos que aparece varias veces, lo que puede sobrerrepresentar esos datos.
- **Formato inconsistente:** Datos que tienen un formato diferente en los registros, como fechas escritas en varios formatos. Esto dificulta el análisis.
- **Datos incorrectos:** Valores que son claramente incorrectos, como texto ingresado en un campo numérico.

- **Datos irrelevantes:** Información que no se aplica a los objetivos del análisis. Esto se debe eliminar.

Una limpieza de datos exhaustiva para corregir estos errores es crucial antes de continuar con el procesamiento y el análisis de datos, ya que los datos sucios pueden generar un entrenamiento del modelo de aprendizaje automático poco confiable y perspectivas analíticas defectuosas. El proceso de limpieza debe utilizar métodos automatizados como la identificación de valores atípicos y la normalización de texto combinada con la verificación manual para detectar todos los problemas. Los datos limpios conducen a un análisis de mayor calidad.

1.4.6 ¿Cuáles son los mejores métodos para la limpieza de datos?

La limpieza de datos es un paso fundamental en el análisis de datos para garantizar resultados precisos y confiables. Algunas de las principales técnicas de limpieza de datos incluyen:

Corrección de errores Corregir errores en los datos, como errores tipográficos, problemas de formato, valores fuera de rango, etc. Esto mejora la calidad de los datos. Los métodos incluyen:

- Coincidencia de patrones para identificar anomalías
- Establecer reglas de validación
- Revisiones manuales

Detección de anomalías

Identificar valores atípicos que caen fuera de los rangos de valores esperados. Los métodos clave son:

- Calcular la desviación estándar con puntuaciones Z para detectar valores atípicos
- Usar algoritmos de clasificación para detectar anomalías
- Aplicar análisis de agrupamiento para revelar puntos de datos anormales

Imputación de valores faltantes

Reemplazar los valores faltantes con sustitutos adecuados para permitir un análisis completo. Las tácticas incluyen:

- Sustitución de media/moda
- Imputación de regresión
- Predicciones de aprendizaje automático

Escalado de características

Transformar atributos en un rango estándar y comparable de valores. Técnicas populares:

- Escalado mínimo-máximo en un rango de 0 a 1
- -Estandarización con puntuaciones z
- Transformaciones logarítmicas para datos sesgados

Duplicación

Eliminar entradas duplicadas en un conjunto de datos. Esto evita el sesgo estadístico. Los métodos incluyen:

- Ordenar y escanear en busca de duplicados adyacentes
- Comparar valores en todos los campos

- Asignar identificadores únicos

La aplicación frecuente de estos métodos avanzados garantiza datos limpios y precisos para un modelado predictivo y un análisis de datos confiables.

1.4.7 ¿Cuáles son las herramientas correctas para el paso de detección de discrepancias en la limpieza de datos?

La limpieza y preparación de datos es un paso crucial antes de analizar y modelar los datos. Detectar discrepancias en los datos es clave para identificar problemas que deben abordarse. Algunas herramientas útiles para la detección de discrepancias incluyen:

Herramientas de auditoría de datos

Analizar datos para descubrir reglas, relaciones y anomalías. A menudo, se utilizan análisis estadísticos para encontrar correlaciones o agrupaciones para detectar valores atípicos. Ayudan a descubrir problemas de calidad de los datos. Ejemplos: DataCleaner, Talend Open Studio.

Herramientas de visualización Las representaciones visuales hacen que los valores atípicos, las brechas y los errores sean más evidentes. Los gráficos interactivos son útiles para explorar los datos. Ejemplos: Tableau, Power BI, Apache Superset.

SQL Escribir consultas SQL para analizar datos, resumir estadísticas, identificar valores NULL, duplicados y valores atípicos. Útil para evaluar la calidad de los datos.

Hojas de cálculo Ordenar, filtrar, utilizar formato condicional para resaltar problemas. Los gráficos y las tablas dinámicas permiten visualizar áreas problemáticas. Son sencillos pero eficaces para la auditoría básica de datos.

Las herramientas adecuadas dependen de las habilidades con los datos y de los problemas que se deben resolver. Sin embargo, la combinación de enfoques programáticos y visuales suele ser la más eficaz para detectar discrepancias que requieren corrección antes del análisis.

1.4.8 ¿Cómo se limpian los datos para el análisis de regresión?

Limpiar los datos es un paso fundamental antes de realizar un análisis de regresión. Estos son los pasos clave:

Identificar las variables Revisar cuidadosamente los datos e identificar las variables predictoras y de destino que utilizará en el modelo de regresión. Comprender el papel de cada variable guiará las decisiones de limpieza de los datos.

Manejar los valores faltantes Examinar las variables en busca de valores faltantes. Puede eliminar filas o casos con valores faltantes o imputar valores según la cantidad y el patrón de valores faltantes.

Detectar y eliminar valores atípicos Observar la distribución de cada variable predictora e identificar los valores atípicos extremos. Considere eliminar o transformar estos valores para que no influyan demasiado en el ajuste del modelo.

Verificar y transformar la distribución Evaluar si las variables predictoras continuas se distribuyen normalmente. Aplicar transformaciones como el logaritmo o la raíz cuadrada si las distribuciones están muy sesgadas. La normalidad ayuda a mejorar la precisión del modelo.

Codificar variables categóricas Para cualquier predictor categórico, cree variables ficticias para usar en la regresión. Evite dejar las variables categóricas como texto o cadenas.

Escalar y normalizar variables Estandarizar las variables continuas para que estén en una escala común con una media de 0 y una desviación estándar de 1. Esto ayuda a la interpretación de los coeficientes de regresión.

Esto es lo que hay que tener en cuenta Verificar la multicolinealidad entre los predictores Verificar los supuestos del modelo, como la linealidad y la homocedasticidad Evaluar la capacidad predictiva utilizando conjuntos de datos de prueba y de entrenamiento Seguir las mejores prácticas para limpiar las entradas de regresión conduce a modelos más precisos y robustos. Pruebe diferentes técnicas de preparación de datos para optimizar el rendimiento del modelo.

1.4.9 Fundamentos de la corrección de errores y la detección de anomalías

La corrección de errores y la detección de anomalías son procesos críticos en la ingeniería de datos que ayudan a mejorar la calidad de los datos y el rendimiento del modelo. Esta sección explora algunas técnicas clave.

Puntuación Z estándar y detección de valores atípicos La puntuación Z estándar mide cuántas desviaciones estándar se encuentra una observación con respecto a la media. Los valores fuera de -3 a +3 desviaciones estándar son valores atípicos potenciales. La detección de valores atípicos es una parte importante de la detección de anomalías, ya que pueden indicar errores o eventos inusuales. Algunas formas clave de detectar valores atípicos con puntuaciones Z:

- Calcular la puntuación Z para cada punto de datos
- Establecer un umbral (p. ej., -3 a +3)
- Marcar las observaciones con una puntuación Z por encima del umbral como valores atípicos potenciales
- Investigar más a fondo las observaciones marcadas

Las puntuaciones Z permiten detectar anomalías incluso cuando se desconoce la distribución de datos subyacente. Esto las convierte en una técnica versátil para explorar conjuntos de datos.

Escalado de características y normalización de datos Muchos algoritmos de aprendizaje automático funcionan mejor cuando las características están en una escala similar. El escalado de características transforma los datos para que tengan una media de 0 y una desviación estándar de 1. Las técnicas comunes incluyen:

- Escalado mínimo-máximo
- Estandarización (puntuaciones z)
- Transformaciones logarítmicas

La normalización ajusta los datos para que se ajusten a una forma de distribución específica. Esto puede ayudar a abordar distribuciones sesgadas durante el entrenamiento del modelo de detección de anomalías.

Aprovechamiento del aprendizaje automático para el análisis predictivo Los modelos de aprendizaje automático, como la regresión y las redes neuronales, pueden analizar datos históricos para predecir valores esperados. Las diferencias significativas entre los valores previstos y los reales pueden indicar anomalías o errores.

Algunos modelos comunes utilizados:

- Regresión lineal
- Regresores de bosque aleatorio
- Autocodificadores

Los modelos pueden puntuar los datos entrantes y marcar anomalías. Esto permite el monitoreo en tiempo real de los flujos de datos.

Inteligencia artificial (IA) en la detección de anomalías Las técnicas de IA están haciendo avanzar la detección de anomalías en conjuntos de datos grandes y complejos. Los métodos de aprendizaje no supervisado pueden perfilar automáticamente patrones de datos normales para identificar desviaciones. Los métodos clave incluyen:

- Algoritmos de agrupamiento
- Redes neuronales
- Modelos de aprendizaje profundo

La IA proporciona capacidades predictivas para corregir anomalías. Esto puede mejorar la calidad de los datos a través de la identificación de errores y ajustes por sesgo.

1.4.10 Técnicas avanzadas de limpieza de datos para la calidad de los macrodatos

Mantener datos de alta calidad es fundamental para las organizaciones que aprovechan el análisis de macrodatos y la IA. Sin embargo, el volumen, la variedad y la velocidad de los macrodatos pueden introducir errores y anomalías que socaven el análisis. Las técnicas avanzadas como el modelado predictivo, los algoritmos de detección de anomalías, los mecanismos de corrección automática y los procesos de limpieza de datos continuos permiten a las organizaciones mejorar la calidad de los macrodatos.

Modelado predictivo para la identificación de errores El modelado predictivo analiza los datos históricos para identificar patrones y relaciones. Estos modelos pueden evaluar los nuevos datos para detectar posibles errores o valores atípicos en función de lo esperado. Por ejemplo, los modelos predictivos pueden marcar un aumento repentino en el tráfico del sitio web como anormal en función de los patrones de tráfico típicos. Esto permite identificar rápidamente los problemas de datos incluso en conjuntos de datos masivos.

Algoritmos de detección de anomalías para macrodatos Los algoritmos especializados de detección de anomalías están diseñados para manejar la escala y la complejidad de los macrodatos. Las técnicas de aprendizaje automático no supervisadas pueden modelar el comportamiento normal de los datos y detectar valores atípicos sin una costosa supervisión manual. Algoritmos como bosques de aislamiento, factores atípicos locales y covarianza robusta aprovechan métodos estadísticos avanzados para descubrir anomalías. Estas técnicas ayudan a descubrir errores e irregularidades que de otro modo pasarían desapercibidos.

Mecanismos de corrección de anomalías automatizados Una vez que se han detectado anomalías, los sistemas automatizados pueden iniciar acciones correctivas sin intervención humana. Las reglas predefinidas pueden activar eventos como filtrado, imputación de valores faltantes o suavizado de valores atípicos. La corrección más avanzada puede emplear técnicas de aprendizaje supervisado para recomendar acciones basadas en ejemplos corregidos anteriores. Esta automatización permite la mejora continua de la calidad de los datos a escala de big data.

Garantizar la integridad de los datos con la limpieza continua de datos Con la limpieza continua de datos, las canalizaciones automatizadas ejecutan controles periódicos utilizando reglas

de validación, restricciones de integridad y pruebas de calidad. Los problemas se registran, las anomalías se marcan y los problemas pueden activar alertas para una inspección más exhaustiva. Los procesos continuos permiten el monitoreo constante de la calidad de los datos, lo que garantiza que la precisión y la confiabilidad se mantengan a lo largo del tiempo incluso a medida que ingresan nuevos datos. Este enfoque proactivo es esencial para la integridad de los datos a largo plazo.

Implementación práctica en flujos de trabajo de análisis de datos La calidad de los datos es crucial para un análisis preciso de los datos y una toma de decisiones eficaz. Las técnicas avanzadas de limpieza de datos, como la corrección de errores y la detección de anomalías, pueden mejorar en gran medida la calidad de los datos cuando se implementan correctamente en los flujos de trabajo y las canalizaciones de datos.

Incorporación de la corrección de errores en las canalizaciones de datos - Realice la validación de datos en la ingesta para detectar problemas de forma temprana Aproveche las reglas, las búsquedas y los datos maestros para identificar errores comunes - Utilice el reconocimiento de patrones para señalar los errores probables para que los revise un humano - Incorpore la corrección automática de errores en el proceso ETL utilizando la lógica de validación - Supervise continuamente las métricas clave para refinar las reglas de detección de errores

Diseño de sistemas de detección de anomalías para análisis en tiempo real - Centrarse en indicadores clave de rendimiento y métricas críticos que indiquen la salud del negocio - Establecer líneas de base dinámicas adaptadas a patrones de métricas - Habilitar alertas en tiempo real cuando se produzcan anomalías - Priorizar anomalías para investigación en función de la gravedad - Reentrenar modelos periódicamente a medida que surjan nuevos patrones de datos

Mecanismos de retroalimentación para la corrección de anomalías - Registrar todas las detecciones de anomalías y las correcciones realizadas - Etiquetar anomalías verificadas como verdaderas o falsas positivas - Usar registros y etiquetas para mejorar la precisión de la detección - Permitir que los analistas proporcionen contexto sobre anomalías - Incorporar la retroalimentación de los analistas en la lógica de corrección

Estudios de caso: corrección de errores y detección de anomalías en acción

Empresa de comercio electrónico

- Detectar errores de carga de productos de forma temprana, evitando que ingresen datos incorrectos al sistema
- Identificar picos anómalos en ventas y reembolsos debido a problemas del sitio
- Las soluciones rápidas limitaron el impacto en los ingresos y las experiencias negativas de los clientes

Aplicación de viajes compartidos

- Marcado Anomalías en el comportamiento del conductor y en los detalles del viaje
- Ayudó a identificar posibles casos de fraude y abuso para su revisión
- El sistema de retroalimentación mejora continuamente los patrones de detección de anomalías

La incorporación cuidadosa de procesos de corrección de errores y detección de anomalías en los flujos de trabajo de análisis y los canales de datos es clave para permitir una toma de decisiones impactante basada en datos. Las estrategias de implementación adecuadas pueden generar mejoras significativas en la calidad de los datos y el rendimiento empresarial.

1.5 Datos faltantes | Tipos, explicación e imputación

Los datos faltantes, o valores faltantes, ocurren cuando no tiene datos almacenados para ciertas variables o participantes. Los datos pueden perderse debido a una entrada de datos incompleta, mal funcionamiento del equipo, pérdida de archivos y muchas otras razones.

En cualquier conjunto de datos, suele haber algunos datos faltantes. En la investigación cuantitativa, los valores faltantes aparecen como celdas en blanco en su hoja de cálculo.

Tipos de datos faltantes Los datos faltantes son errores porque sus datos no representan los valores verdaderos de lo que se propuso medir.

Es importante considerar el motivo de los datos faltantes, ya que lo ayuda a determinar el tipo de datos faltantes y lo que debe hacer al respecto.

Hay tres tipos principales de datos faltantes.

Tipo	Definición
Datos faltantes completamente aleatorios (MCAR)	Los datos faltantes se distribuyen aleatoriamente en la variable y no están relacionados con otras variables.
Datos faltantes aleatorios (MAR)	Los datos faltantes no se distribuyen aleatoriamente, sino que se explican por otras variables observadas.
Datos faltantes no aleatorios (MNAR)	Los datos faltantes difieren sistemáticamente de los valores observados.

Ejemplo: Proyecto de investigación Recopila datos sobre patrones de gastos de fin de año. Encuesta a adultos sobre cuánto gastan anualmente en regalos para familiares y amigos en cantidades en dólares.

Datos faltantes completamente aleatorios Cuando los datos faltan completamente aleatorios (MCAR), la probabilidad de que falte un valor en particular en su conjunto de datos no está relacionada con nada más.

Los valores faltantes se distribuyen aleatoriamente, por lo que pueden provenir de cualquier parte de la distribución completa de sus valores. Estos datos MCAR tampoco están relacionados con otras variables no observadas.

Ejemplo: datos MCAR Observa que hay algunos valores faltantes en su conjunto de datos de gastos de vacaciones. Algunas personas comenzaron a responder su encuesta, pero abandonaron o se saltaron una pregunta. Sin embargo, observa que tiene puntos de datos de una amplia distribución, que van desde valores bajos a altos.

Por lo tanto, concluye que los valores faltantes no están relacionados con ningún rango específico de gasto en vacaciones.

A menudo, los datos se consideran MCAR si parecen no estar relacionados con valores específicos u otras variables. En la práctica, es difícil cumplir con este supuesto porque la “aleatoriedad real” es poco común.

Cuando faltan datos debido a fallas en el equipo o muestras perdidas, se consideran MCAR.

Datos faltantes al azar Los datos faltantes al azar (MAR, por sus siglas en inglés) no son datos faltantes al azar; este término es un poco inapropiado.

Este tipo de datos faltantes difiere sistemáticamente de los datos que ha recopilado, pero se puede explicar completamente mediante otras variables observadas.

La probabilidad de que falte un punto de datos está relacionada con otra variable observada, pero no con el valor específico de ese punto de datos en sí.

Ejemplo: datos MAR Repite su recopilación de datos con un nuevo grupo. Observa que hay más valores faltantes para adultos de 18 a 25 años que para otros grupos de edad. Pero al observar los datos observados para adultos de 18 a 25 años, observa que los valores están muy dispersos. Es poco probable que los datos faltantes se deban a los valores específicos en sí.

En cambio, algunos adultos más jóvenes pueden estar menos inclinados a revelar los montos de sus gastos de vacaciones por razones no relacionadas (por ejemplo, mayor protección de su privacidad).

Datos faltantes no aleatorios Los datos faltantes no aleatorios (MNAR) faltan por razones relacionadas con los valores mismos.

Ejemplo: datos MNAR En el nuevo conjunto de datos, también observa que hay menos valores bajos. Algunos participantes con ingresos bajos evitan informar los montos de sus gastos de vacaciones porque son bajos. Es importante buscar este tipo de datos faltantes porque puede faltar información de subgrupos clave dentro de su muestra. Es posible que su muestra no sea representativa de su población.

Sesgo de deserción En estudios longitudinales, el sesgo de deserción puede ser una forma de datos MNAR. El sesgo de deserción significa que algunos participantes tienen más probabilidades de abandonar que otros.

Por ejemplo, en estudios médicos a largo plazo, algunos participantes pueden abandonar porque se enferman cada vez más a medida que avanza el estudio. Sus datos son MNAR porque sus resultados de salud son peores, por lo que su conjunto de datos final puede incluir solo personas sanas y perder datos importantes.

1.5.1 ¿Los datos faltantes son problemáticos?

Los datos faltantes son problemáticos porque, según el tipo, a veces pueden causar sesgo de muestreo. Esto significa que sus resultados pueden no ser generalizables fuera de su estudio porque sus datos provienen de una muestra no representativa.

En la práctica, a menudo puede considerar que dos tipos de datos faltantes son ignorables porque los datos faltantes no difieren sistemáticamente de sus valores observados:

- Datos MCAR
- Datos MAR

Para estos dos tipos de datos, la probabilidad de que falte un punto de datos no tiene nada que ver con el valor en sí. Por lo tanto, es poco probable que sus valores faltantes sean significativamente diferentes de sus valores observados.

Por otro lado, tiene un conjunto de datos sesgado si los datos faltantes difieren sistemáticamente de sus datos observados. Los datos que son MNAR se denominan no ignorables por esta razón.

1.5.2 Cómo prevenir los datos faltantes

Los datos faltantes a menudo provienen de sesgo de deserción, falta de respuesta o protocolos de investigación mal diseñados. Al diseñar su estudio, es una buena práctica facilitar a sus participantes la provisión de datos.

A continuación, se ofrecen algunos consejos que le ayudarán a minimizar la falta de datos:

- Limite la cantidad de seguimientos
- Minimice la cantidad de datos recopilados
- Haga que los formularios de recopilación de datos sean fáciles de usar
- Utilice técnicas de validación de datos
- Ofrezca incentivos

Una vez que haya recopilado los datos, es importante almacenarlos con cuidado y realizar varias copias de seguridad.

1.5.3 Cómo lidiar con los valores faltantes

Para ordenar sus datos, sus opciones generalmente incluyen aceptar, eliminar o recrear los datos faltantes.

Debe considerar cómo lidiar con cada caso de datos faltantes según su evaluación de por qué faltan los datos.

- ¿Estos datos faltan por razones aleatorias o no aleatorias?
- ¿Los datos faltan porque representan valores cero o nulos?
- ¿La pregunta o la medida estaban mal diseñadas?

Sus datos pueden aceptarse o dejarse como están si son MCAR o MAR. Sin embargo, los datos MNAR pueden necesitar un tratamiento más complejo.

Aceptación

La opción más conservadora implica aceptar sus datos faltantes: simplemente deje estas celdas en blanco.

Es mejor hacer esto cuando crea que está tratando con valores MCAR o MAR. Cuando tenga una muestra pequeña, querrá conservar la mayor cantidad de datos posible porque cualquier eliminación de datos puede afectar su poder estadístico.

También puede recodificar todos los valores faltantes con etiquetas de "N/A" (abreviatura de "no aplicable") para que sean coherentes en todo el conjunto de datos.

Estas acciones le ayudan a conservar los datos de la mayor cantidad posible de sujetos de investigación con pocos o ningún cambio.

Eliminación

Puede eliminar los datos faltantes de los análisis estadísticos mediante la eliminación por lista o por pares.

Eliminación por lista

La eliminación por lista significa eliminar los datos de todos los casos (participantes) que tienen datos faltantes para cualquier variable en su conjunto de datos. Tendrá un conjunto de datos que está completo para todos los participantes incluidos en él.

Una desventaja de esta técnica es que puede terminar con una muestra mucho más pequeña y/o sesgada con la que trabajar. Si faltan cantidades significativas de datos de algunas variables o medidas en particular, los participantes que proporcionan esos datos pueden diferir significativamente de los que no lo hacen.

Su muestra podría estar sesgada porque no representa adecuadamente a la población.

Ejemplo: eliminación por lista Decide eliminar todos los participantes con datos faltantes de su conjunto de datos de encuesta. Esto reduce la muestra de 114 a 77 participantes. Observa que la mayoría de los participantes con datos faltantes dejaron sin responder una pregunta específica sobre sus opiniones. Muchos de esos participantes también eran mujeres, por lo que ahora su muestra está compuesta principalmente por hombres.

Eliminación por pares

La eliminación por pares le permite conservar más datos al eliminar únicamente los puntos de datos que faltan en los análisis. Conserva más datos porque se incluyen todos los datos disponibles de los casos.

También significa que tiene un tamaño de muestra desigual para cada una de las variables. Pero es útil cuando tiene una muestra pequeña o una gran proporción de valores faltantes para algunas variables.

Cuando realiza análisis con múltiples variables, como una correlación, solo se incluyen los casos (participantes) con datos completos para cada variable.

Ejemplo: eliminación por pares Decide eliminar únicamente los valores faltantes, mientras conserva los otros puntos de datos para estos participantes. Esto no reduce el tamaño general de la muestra. 12 personas no respondieron una pregunta sobre su género, lo que reduce el tamaño de la muestra de 114 a 102 participantes para la variable "género". 3 personas no respondieron una pregunta sobre su edad, lo que reduce el tamaño de la muestra de 114 a 111 participantes para la variable "edad". De esta manera, puede conservar más valores, pero el tamaño de la muestra ahora difiere entre las variables.

Imputación

Imputación significa reemplazar un valor faltante con otro valor basado en una estimación razonable. Utiliza otros datos para recrear el valor faltante y obtener un conjunto de datos más completo.

Puede elegir entre varios métodos de imputación.

El método de imputación más sencillo implica reemplazar los valores faltantes con el valor medio o mediano de esa variable.

Imputación hot-deck

En la imputación hot-deck, reemplaza cada valor faltante con un valor existente de un caso o participante similar dentro de su conjunto de datos. Para cada caso con valores faltantes, el valor faltante se reemplaza por un valor de un denominado "donante" que es similar a ese caso según los datos de otras variables.

Ejemplo: imputación hot-deck En una encuesta, pide a los participantes que respondan preguntas sobre cómo califican una nueva aplicación de compras del 1 al 5. Observa que dos participantes se saltaron la pregunta 3, por lo que estas celdas están vacías. Ordena los datos en función de otras variables y busca participantes que respondieron de manera similar a otras preguntas en comparación con los participantes con valores faltantes.

Toma la respuesta a la pregunta 3 de un donante y úsala para completar la celda en blanco para cada valor faltante.

Imputación de base fría

Alternativamente, en la imputación de base fría, reemplaza los valores faltantes con valores existentes de casos similares de otros conjuntos de datos. Los nuevos valores provienen de una muestra no relacionada.

Ejemplo: imputación de base fría En lugar de reemplazar los valores faltantes con respuestas de participantes de la misma muestra, abre un conjunto de datos diferente de un compañero de trabajo. Realizó una encuesta similar pero utilizó una muestra diferente. Busca participantes que respondieron de manera similar a otras preguntas en comparación con los participantes con valores faltantes.

Toma la respuesta a la pregunta 3 del otro conjunto de datos y la usa para completar la celda en blanco para cada valor faltante.

1.6 1.2.3 Comprender la normalización y el escalamiento de datos.

Este será una de las principales tareas que realicemos en el preprocesamiento de datasets. Los principales algoritmos de machine learning que existen no funcionan muy bien cuando existen una gran diferencia entre los valores de una columna. Si tenemos una columna «L» que contiene valores muy distante, por ejemplo, tiene un valor mínimo de 0 y máximo de 849820. Esto es algo que debemos evitar, ya que los algoritmos de aprendizaje no funcionan bien en estos casos.

Existen dos formas claras de solucionar estos problemas de escalas: la normalización de valores y la estandarización.

Normalización

El escalado min-max o normalización, es una técnica común a la hora de solucionar el problema de tener diferentes escalas en los valores de una columna. El objetivo que se consigue con esta técnica es que todos los valores de una columna estén comprendido en el intervalo [0-1]. De forma matemática, lo que estamos haciendo es a cada valor le restamos el mínimo y lo dividimos entre el valor máximo.

Otras maneras de utilizar la normalización, además de min-max son:

- **Normalización min-max:** se calcula de la siguiente fórmula

$$x_s = \frac{x - x_{min}}{x_{max} - x_{min}} * (max - min) + min$$

  donde:

- x es la variable en su escala original
- xmax y xmin son el máximo y el mínimo

- *max* y *min* son el máximo y el mínimo pre-definidos (es decir que queremos obtener tras el escalamiento)
- xs es la variable obtenida tras el escalamiento

¿Cuándo NO usar MinMaxScaler?

- Cuando la distribución tiene un sesgo Una distribución con sesgo es cuando su forma se aleja demasiado de una distribución normal o gaussiana (campana simétrica).

En este caso NO se recomienda el uso de MinMaxScaler pues al escalar los datos comprimimos la distribución de los datos a un rango más pequeño que el original. Es decir que desaprovechamos todo el rango de valores disponible tras el escalamiento

- Cuando los datos tienen valores extremos (outliers) Los valores extremos (u outliers) son simplemente datos cuyos valores se encuentran excepcionalmente fuera del rango de valores de la mayoría de nuestros datos.

En este caso tampoco se recomienda el uso de minmaxscaler. Y esto se debe a que por tratarse de valores extremos, estos outliers generalmente corresponderán a los valores xmax y xmin que aparecen en la ecuación de escalamiento de nuestra variable original.

Así que al hacer el escalamiento estos valores extremos quedarán mapeados al máximo y mínimo en el rango resultante, mientras que los valores no extremos (que es la mayoría de los datos) quedarán mapeados dentro de un rango resultante muchísimo menor.

Es decir que si usamos minmaxscaler cuando tenemos valores extremos en nuestros datos hace que la mayoría de nuestros datos (que no son extremos) queden como resultado comprimidos a un rango de valores muy pequeño.

¿Cuándo podemos usar MinMaxScaler? Teniendo en cuenta lo anterior, se sugiere el uso de MinMaxScaler en estas situaciones:

1. Cuando tenemos claro el rango de valores esperado a la salida del escalamiento
2. Cuando la distribución de los datos NO tiene demasiado sesgo
3. Cuando los datos NO contienen outliers

Estandarización

La otra forma de realizar el escalado de valores que vamos a ver se denomina estandarización. Matemáticamente hablando, lo que estamos haciendo en este proceso es restar la media de los valores y dividir por la desviación estándar de los mismos. De esta forma los valores obtenidos tendrán una media de cero y una varianza de uno.

Existen algunas diferencias notables con respecto a la normalización. En primer lugar los valores obtenidos por la estandarización no están acotados en ningún rango ([0-1] por ejemplo).

En segundo lugar, este método consigue solucionar el problema de valores atípicos u outliers que presenta la normalización. Por ejemplo, supongamos que un atributo de temperatura contiene valores entre 0 y 100 por regla general. Por un error de medición, tenemos un valor de 10000 que se consideraría un outlier. Si aplicamos la normalización, la mayor parte de valores estarían en el rango 0-0.1. Sin embargo, esto no se da con la estandarización.

En sklearn tenemos el método StandardScaler() para calcularlo usamos los siguiente:

- **Normalización Z-score:** sigue la siguiente fórmula:

$$z = (x - u)/s$$

donde:

- x el valor de la muestra
- u es la media de los datos
- s es la desviación estandar de los datos.

Vamos a ver algunos ejemplos prácticos:

1.7 Ejemplo 1

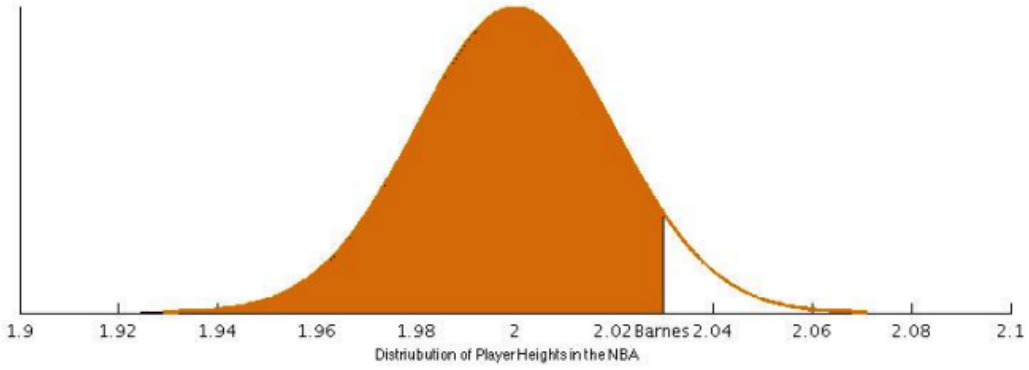
La altura promedio de un jugador de baloncesto profesional fue de 2.00 metros con una desviación estándar de 0.02 metros. Harrison Barnes es un jugador de baloncesto que mide 2.03. ¿ Cuántas desviaciones estándar de la media es la altura de Barnes?

Primero debemos dibujar la curva normal que representa la distribución de las alturas de los jugadores de baloncesto.

[1]:
```
from IPython import display
```

[2]:
```
display.Image("ejer_1.png")
```

[2]:



Observa que colocamos la altura media 2.00 justo en el medio de la distribución y hacemos marcas que están a 1 desviación estandar o 0.02 metros de distancia en ambas direcciones

A continuación, debemos calcular la puntuación estándar (es decir, la puntuación Z) para la altura de Barnes. Ya que μ= 2.00, s= 0.02 y x = 2.03 podemos encontrar el puntaje Z.

[3]:
```
# Z score = x - μ / s
Zscore = round((2.03-2)/0.02, 2)
Zscore
```

[3]: 1.5

[4]:
```
# Otra forma

μ = 2.0
x = 2.03
s = 0.02
```

[5]:
```
Zscore = round((x - μ)/s, 2)
Zscore
```

[5]: 1.5

Encontrarmos 1.5 como la puntuación Z, nos dice que la altura de Barnes está 1.5 desviaciones estándar de la media, es decir, 1.5 * s + μ = Altura de Barnes

[6]:
```
Altura_Barnes = (Zscore * s) + μ
Altura_Barnes
```

[6]: 2.03

1.8 Ejemplo 2

La altura de promedio de un jugador de hockey profesional es de 1.86 metros con una desviación estándar de 0.06 metros. Tyler Myers, un profesional de hockey, tiene la misma altura que Harrison Barnes. ¿Cuál de los dos es más alto en su respectiva liga?

Para encontrar la puntuación estándar de Tyler Myers podemos usar la información: μ= 1.86, s=0.06, x= 2.03. Esto da como resultado la puntuación estándar:

[7]:
```
μ = 1.86
s = 0.06
x = 2.03
```

[8]:
```
Zscore = round((x - μ)/ s, 2)
Zscore
```

[8]: 2.83

Comparando las dos puntuaciones Z, vemos que la puntuación de Tyler Myers es de 2.83 es mayor que la puntuación de Barnes 1.5. Esto nos dice que hay más jugadores de hockey más bajos que Myers que jugadores de baloncesto más bajos que Barnes.

Así pues podemos encontrarnos con distintas etapas del preprocesamiento:

- **Data cleaning:** la limpieza de datos elimina ruido y resuelve las inconsistencias en los datos.
- **Data integration:** con la Integración de datos se migran datos de varias fuentes a una fuente coherente como un Data Warehouse.
- **Data transformation:** la transformación de datos sirve para normalizar datos de cualquier tipo.
- **Data reduction:** la reducción de datos reduce el tamaño de los datos agregandolos.

1.8.1 Explique la codificación de variables categóricas para el análisis cuantitativo, incluidos los métodos de codificación one-hot y codificación de etiquetas.

En general, los algoritmos de machine learning no trabajan bien con datos con valores que no sean numéricos. Es por ello, que aquellos datos con valores que contengan un texto o sean una categoría, debemos transformarlos a valores numéricos.

Ejemplo de columnas no numéricas: «Color» que nos indica el color en inglés (Red, Blue, etc) y «Spectral_Class» que es otro dato categórico (M, O, A, etc).

Tipos de datos categóricos

Los datos categóricos se pueden clasificar en términos generales en dos tipos:

- **Datos nominales:** este tipo de datos representan categorías sin ningún orden inherente. Los ejemplos incluyen género (masculino, femenino), color (rojo, azul, verde) y país (EE. UU., India, Reino Unido).
- **Datos ordinales:** este tipo de datos representa categorías con un orden o clasificación significativo. Los ejemplos incluyen el nivel educativo (escuela secundaria, licenciatura, maestría, doctorado) y la satisfacción del cliente (baja, media, alta).

Técnicas de codificación Exploremos las técnicas más utilizadas:

1. Codificación de etiquetas (Label encoding)

La codificación de etiquetas es un método simple y directo que asigna un número entero único a cada categoría. Este método es adecuado para datos ordinales donde el orden de las categorías es significativo. Caso de Uso: Aplicable para casos en los que el ordenamiento de categorías tiene relevancia analítica.

```
data = ['red', 'blue', 'green', 'blue', 'red']

0 => 'blue'
1 => 'green'
2 => 'red'

data = [2 0 1 0 2]
```

2. Codificación en caliente (One-hot Encoding)

One-Hot Encoding convierte datos categóricos en una matriz binaria, donde cada categoría está representada por un vector binario. Este método es adecuado para datos nominales. Caso de uso: Más apropiado para aquellas situaciones en las que las categorías no tienen un orden inherente o existe una distinción clara entre ellas.

red	blue	green
0	1	0
1	0	0
0	0	1

```
data = ['red', 'blue', 'green', 'blue', 'red']
```

```
data = [[0. 1. 0.]
        [1. 0. 0.]
        [0. 0. 1.]
        [1. 0. 0.]
        [0. 1. 0.]]
```

Ventajas y desventajas de cada técnica de codificación

Técnica de codificación	Ventajas	Desventajas
Codificación de etiquetas	- Simple y fácil de implementar - Adecuado para datos ordinales	- Introduce relaciones ordinales arbitrarias para datos nominales- Puede no funcionar bien con valores atípicos
Codificación en caliente	- Adecuado para datos nominales - Evita introducir relaciones ordinales- Mantiene información sobre los valores de cada variable	- Puede provocar un aumento de la dimensionalidad y la escasez. - Puede provocar un sobreajuste, especialmente con muchas categorías y tamaños de muestra pequeños.

1.8.2 Explique los pros y contras de la reducción de datos (reducir el número de variables bajo consideración o simplificar los modelos versus pérdida de explicabilidad de los datos).

La reducción de datos es una técnica utilizada en la minería de datos para reducir el tamaño de un conjunto de datos y al mismo tiempo preservar la información más importante. Esto puede resultar beneficioso en situaciones en las que el conjunto de datos es demasiado grande para procesarlo de forma eficiente o en las que el conjunto de datos contiene una gran cantidad de información irrelevante o redundante.

Existen varias técnicas diferentes de reducción de datos que se pueden utilizar en la minería de datos, que incluyen:

- **Muestreo de datos:** esta técnica implica seleccionar un subconjunto de datos con los que trabajar, en lugar de utilizar todo el conjunto de datos. Esto puede resultar útil para reducir el tamaño de un conjunto de datos y al mismo tiempo preservar las tendencias y patrones generales de los datos.
- **Reducción de dimensionalidad:** esta técnica implica reducir la cantidad de funciones en el conjunto de datos, ya sea eliminando funciones que no son relevantes o combinando varias funciones en una sola.
- **Compresión de datos:** esta técnica implica el uso de técnicas como la compresión con o sin pérdidas para reducir el tamaño de un conjunto de datos.
- **Discretización de datos:** esta técnica implica convertir datos continuos en datos discretos dividiendo el rango de valores posibles en intervalos o contenedores.
- **Selección de características:** esta técnica implica seleccionar un subconjunto de características del conjunto de datos que sean más relevantes para la tarea en cuestión.

Es importante tener en cuenta que la reducción de datos puede tener un equilibrio entre la precisión y el tamaño de los datos. Cuantos más datos se reduzcan, menos preciso será el modelo y menos

generalizable.

En conclusión, la reducción de datos es un paso importante en la minería de datos, ya que puede ayudar a mejorar la eficiencia y el rendimiento de los algoritmos de aprendizaje automático al reducir el tamaño del conjunto de datos. Sin embargo, es importante ser consciente del equilibrio entre el tamaño y la precisión de los datos y evaluar cuidadosamente los riesgos y beneficios antes de implementarlos.

Ventajas o desventajas de la Reducción de Datos en Minería de Datos:

La reducción de datos en la minería de datos puede tener una serie de ventajas y desventajas.

Ventajas: - **Eficiencia mejorada:** la reducción de datos puede ayudar a mejorar la eficiencia de los algoritmos de aprendizaje automático al reducir el tamaño del conjunto de datos. Esto puede hacer que trabajar con grandes conjuntos de datos sea más rápido y práctico.

- **Rendimiento mejorado:** la reducción de datos puede ayudar a mejorar el rendimiento de los algoritmos de aprendizaje automático al eliminar información irrelevante o redundante del conjunto de datos. Esto puede ayudar a que el modelo sea más preciso y robusto.
- **Costos de almacenamiento reducidos:** la reducción de datos puede ayudar a reducir los costos de almacenamiento asociados con grandes conjuntos de datos al reducir el tamaño de los datos.
- **Interpretabilidad mejorada:** la reducción de datos puede ayudar a mejorar la interpretabilidad de los resultados al eliminar información irrelevante o redundante del conjunto de datos.

Desventajas: - **Pérdida de información:** la reducción de datos puede resultar en una pérdida de información si se eliminan datos importantes durante el proceso de reducción.

- **Impacto en la precisión:** la reducción de datos puede afectar la precisión de un modelo, ya que reducir el tamaño del conjunto de datos también puede eliminar información importante que se necesita para realizar predicciones precisas.
- **Impacto en la interpretabilidad:** la reducción de datos puede dificultar la interpretación de los resultados, ya que eliminar información irrelevante o redundante también puede eliminar el contexto necesario para comprender los resultados.
- **Costos computacionales adicionales:** la reducción de datos puede agregar costos computacionales adicionales al proceso de minería de datos, ya que requiere tiempo de procesamiento adicional para reducir los datos.

En conclusión, la reducción de datos puede tener ventajas y desventajas. Puede mejorar la eficiencia y el rendimiento de los algoritmos de aprendizaje automático al reducir el tamaño del conjunto de datos. Sin embargo, también puede provocar una pérdida de información y dificultar la interpretación de los resultados. Es importante sopesar los pros y los contras de la reducción de datos y evaluar cuidadosamente los riesgos y beneficios antes de implementarla.

Las técnicas empleadas para la reducción de los datos son PCA y LDA.

Análisis de Componentes Principales (PCA)

El análisis de componentes principales (PCA) es una técnica de reducción de dimensionalidad lineal que se puede utilizar para extraer información de un espacio de alta dimensión proyectándola en un

subespacio de menor dimensión. Intenta preservar las partes esenciales que tienen más variación de los datos y eliminar las partes no esenciales con menos variación.

Una cosa importante a tener en cuenta sobre PCA es que es una técnica de reducción de dimensionalidad no supervisada , puede agrupar los puntos de datos similares en función de la correlación de características entre ellos sin supervisión (o etiquetas).

es un procedimiento estadístico que utiliza una transformación ortogonal para convertir un conjunto de observaciones de variables posiblemente correlacionadas (entidades cada una de las cuales toma varios valores numéricos) en un conjunto de valores de variables linealmente no correlacionadas llamadas componentes principales.

Pero, ¿dónde se puede aplicar PCA?

Visualización de datos: cuando se trabaja en cualquier problema relacionado con datos, el desafío en el mundo actual es el gran volumen de datos y las variables/características que definen esos datos. Para resolver un problema donde los datos son la clave, necesita una amplia exploración de datos, como descubrir cómo se correlacionan las variables o comprender la distribución de algunas variables. Teniendo en cuenta que hay una gran cantidad de variables o dimensiones a lo largo de las cuales se distribuyen los datos, la visualización puede ser un desafío y casi imposible.

Por lo tanto, PCA permite visualizar los datos en un espacio 2D o 3D a simple vista.

Aceleración del algoritmo de aprendizaje automático (ML): dado que la idea principal de PCA es la reducción de la dimensionalidad, puede aprovecharla para acelerar el entrenamiento y el tiempo de prueba de su algoritmo de aprendizaje automático, teniendo en cuenta que sus datos tienen muchas características y que el aprendizaje del algoritmo ML es demasiado lento.

En un nivel abstracto, toma un conjunto de datos que tiene muchas características y simplifica ese conjunto de datos seleccionando algunas Principal Componentes de las características originales.

¿Qué es un componente principal?

Los componentes principales son la clave de PCA. En términos sencillos, cuando los datos se proyectan en una dimensión más baja (suponga tres dimensiones) desde un espacio más alto, las tres dimensiones no son más que los tres componentes principales que capturan (o contienen) la mayor parte de la variación (información) de sus datos .

Los componentes principales tienen dirección y magnitud. La dirección representa a través de qué ejes principales se distribuyen principalmente los datos o tienen la mayor variación y la magnitud indica la cantidad de variación que el Componente principal captura de los datos cuando se proyecta en ese eje. Los componentes principales son una línea recta y el primer componente principal tiene la mayor variación en los datos. Cada componente principal posterior es ortogonal al último y tiene una varianza menor. De esta forma, dado un conjunto de “x” variables correlacionadas sobre “y” muestras, se obtiene un conjunto de “u” componentes principales no correlacionados sobre las mismas “y” muestras.

La razón por la que obtiene componentes principales no correlacionados de las características originales es que las características correlacionadas contribuyen al mismo componente principal, reduciendo así las características de datos originales a componentes principales no correlacionados; cada uno representa un conjunto diferente de características correlacionadas con diferentes cantidades de variación.

Cada componente principal representa un porcentaje de la variación total capturada de los datos.

1.8.3 Explicar los métodos para manejar valores atípicos, incluidas las técnicas de detección y tratamiento para garantizar la calidad de los datos.

Identificar y manejar valores atípicos es un aspecto crucial pero desafiante del análisis de datos.

El imperativo de la detección de valores atípicos en la ciencia de datos

La detección de valores atípicos es un paso esencial en el proceso de análisis de datos. Cuando los datos contienen anomalías o valores atípicos, pueden sesgar los resultados y generar conocimientos inexactos. Identificar y tratar valores atípicos mejora la calidad de los datos y permite un modelado y una toma de decisiones más precisos.

Los valores atípicos son puntos de datos que difieren significativamente de la mayoría de las observaciones. Pueden ser el resultado de errores experimentales, errores en la entrada de datos o desviaciones naturales. Si no se controlan, los valores atípicos pueden tener un efecto enorme en el análisis. Incluso unos pocos valores atípicos pueden sesgar los promedios, influir en los coeficientes de correlación y distorsionar los modelos de aprendizaje automático.

Afortunadamente, los estadísticos han desarrollado métodos sólidos para detectar valores atípicos en conjuntos de datos:

- Inspección visual mediante diagramas de dispersión y diagramas de caja.
- Enfoques estadísticos como puntuaciones z, rangos intercuartiles y regresión robusta
- Técnicas de aprendizaje automático, como bosques aislados y factores atípicos locales.

Una vez detectados, los valores atípicos deben abordarse mediante un análisis cuidadoso basado en el contexto. La imputación, eliminación o transformación puede estar justificada según el motivo de la anomalía. Es necesaria experiencia en el campo para determinar el tratamiento adecuado.

El imperativo es claro: la detección y gestión de valores atípicos debe ser una parte rutinaria del flujo de trabajo de la ciencia de datos. Aunque a menudo se pasan por alto, estas técnicas son cruciales para garantizar insumos de calidad y resultados confiables. Invertir recursos en el manejo de valores atípicos produce dividendos a través de una mayor integridad y una reducción de errores. Con datos precisos como base, las organizaciones pueden extraer conocimientos más precisos para tomar mejores decisiones.

¿Cómo se manejan los valores atípicos en la limpieza de datos?

Los valores atípicos pueden afectar significativamente los resultados del análisis si no se manejan adecuadamente. A continuación se muestran algunos métodos comunes para detectar y tratar valores atípicos durante la limpieza de datos:

Detectar valores atípicos - Visualice distribuciones de datos con histogramas o diagramas de caja para identificar posibles valores atípicos. Los valores que quedan fuera del patrón general pueden ser valores atípicos. - Calcule estadísticas resumidas como la media y la desviación estándar para identificar valores numéricamente alejados del centro. - Herramientas como Tableau, MATLAB y R brindan buenas capacidades de visualización. Utilice métodos estadísticos como puntuaciones z, rangos intercuartílicos (IQR) o regresión sólida para señalar sistemáticamente posibles valores atípicos. - Considere el contexto al identificar valores atípicos. Un valor puede ser inusualmente alto o bajo según otras mediciones, pero aún así es válido por razones de dominio. - Los modelos de aprendizaje automático como los bosques de aislamiento, los factores de valores atípicos locales (LOF) y el SVM de una clase pueden modelar datos normales y detectar anomalías. Útil para la detección de valores atípicos complejos.

Tratamiento de valores atípicos detectados Una vez que haya identificado posibles valores atípicos, tiene algunas opciones:

- Elimine completamente las filas atípicas del análisis. Este enfoque sencillo elimina su influencia directa.
- Transforme los valores atípicos tapando, agrupando o aplicando suavizado. Esto retiene los valores atípicos sin distorsión.
- Modele los valores atípicos por separado con técnicas especializadas como bosques de aislamiento. Esto divide su efecto.
- Mantenga los valores atípicos sin cambios cuando sea apropiado para sus objetivos de análisis. Su singularidad puede contener ideas.

El mejor enfoque depende del caso de uso, la distribución de los datos y el impacto de eliminar o cambiar los valores atípicos. Se deben tener en cuenta tanto el contexto como la solidez estadística.

En general, a menudo se requiere una combinación de técnicas para una gestión sólida de los valores atípicos. El objetivo es generar conjuntos de datos más limpios y representativos para el análisis y el modelado posteriores.

1.8.4 Comprender la importancia de la estandarización del formato de datos en diferentes conjuntos de datos para lograr coherencia, especialmente cuando se trabaja con formatos de fecha y hora y valores numéricos.

¿Por qué es importante el formato y la estandarización de los datos?

El formato y la estandarización de los datos son importantes por varias razones. En primer lugar, le ayudan a evitar errores e inconsistencias que pueden afectar los resultados y conclusiones de su análisis. Por ejemplo, si los datos contienen diferentes formatos de fecha, símbolos de moneda o variaciones ortográficas, es posible que obtenga cálculos, agregaciones o comparaciones incorrectos. En segundo lugar, le ayudan a mejorar la legibilidad y la usabilidad de sus datos. Por ejemplo, si los datos están formateados y estandarizados de acuerdo con un esquema, convención o estilo común, puede comprender, interpretar y comunicar fácilmente sus datos a otros usuarios. En tercer lugar, le ayudan a mejorar la interoperabilidad y la integración de sus datos. Por ejemplo, si sus datos están formateados y estandarizados de acuerdo con un estándar o protocolo específico, puede intercambiar, compartir y combinar fácilmente sus datos con otras fuentes o sistemas.

¿Cuáles son algunos de los desafíos comunes de formato y estandarización de datos?

Los desafíos de formato y estandarización de datos pueden surgir de diversas fuentes y situaciones, como errores de entrada de datos, problemas de calidad de datos, problemas de diversidad de datos y problemas de complejidad de datos. - Los errores de entrada de datos pueden incluir errores tipográficos, valores faltantes, valores incorrectos, valores duplicados o valores incoherentes. - Los problemas de calidad de los datos pueden incluir valores obsoletos, valores inexactos, valores no válidos o valores faltantes. - Los problemas de diversidad de datos pueden implicar diferentes formatos de fecha, símbolos de moneda, unidades de medida, esquemas de codificación o idiomas. - Por último, los problemas de complejidad de los datos podrían abarcar el big data, los datos no estructurados, los datos de streaming o los datos en tiempo real.

¿Cómo puedes formatear y estandarizar tus datos?

Hay una variedad de herramientas y técnicas que puede utilizar para formatear y estandarizar sus datos, según el tipo, la fuente y el propósito. - La limpieza de datos es el proceso de detección y

corrección de errores e incoherencias, como las reglas de validación de datos, las comprobaciones de calidad de datos, la deduplicación de datos o la imputación de datos. - La transformación de datos implica la conversión o modificación de los datos de un formato o estructura a otro, como el análisis de datos, la extracción de datos, la asignación de datos o la conversión de datos. - La normalización de datos significa ajustar o escalar los datos a un rango o escala común, como la normalización mínima-máxima, la normalización de puntuación z o la normalización de escala decimal. - Por último, la estandarización de datos aplica una regla o método común a los datos, como las reglas de formato de datos, las convenciones de nomenclatura de datos, los esquemas de codificación de datos o los estándares de datos.

¿Cuáles son algunas de las mejores prácticas para el formato y la estandarización de datos?

El formato y la estandarización de los datos no son tareas de una sola vez, sino procesos continuos que requieren planificación, supervisión y evaluación. Para garantizar el éxito, lo mejor es definir los requisitos y expectativas de los datos antes de formatear y estandarizar; elegir las herramientas y técnicas adecuadas; y revisar y actualizar los datos periódicamente. La documentación de estos requisitos y expectativas debe comunicarse a las fuentes de datos, proveedores o socios. Las herramientas y técnicas apropiadas deben seleccionarse en función de los requisitos y expectativas de datos, con pruebas y verificación de los resultados. Por último, es necesario realizar un seguimiento periódico de los datos para garantizar la exactitud, integridad, validez y actualidad. Las actualizaciones también deben realizarse cuando se producen cambios o actualizaciones en las fuentes de datos, formatos, estructuras, calidades o estándares.

La estandarización de datos se logra mediante los siguientes métodos:

Ajuste de escala

Garantizar que los datos numéricos estén en una escala consistente. Ejemplo: si un conjunto de datos mide el peso en kilogramos y otro en libras, la estandarización puede implicar convertir todos los valores a kilogramos. Otra técnica común es el escalamiento de características, donde los valores se transforman para que se encuentren entre un rango específico, a menudo 0 y 1.

Importancia: el uso de datos en diferentes escalas puede sesgar los algoritmos, en particular los sensibles a la magnitud de entrada, como el descenso de gradiente en redes neuronales o los cálculos de distancia en la agrupación.

Consistencia categórica

Garantizar la nomenclatura o el etiquetado consistentes de las categorías. Ejemplo: estandarizar las variaciones de los nombres de países, como “EE. UU.”, “EE. UU.” y “Estados Unidos” en una única etiqueta consistente, como “EE. UU.” Importancia: tener múltiples etiquetas para la misma categoría puede generar análisis fragmentados o sesgados, lo que dificulta la obtención de información precisa.

Formato de fecha

Garantizar que los valores de fecha se adhieran a un formato consistente. Ejemplo: Convertir distintos formatos de fecha como “DD-MM-AAAA”, “MM/DD/AAAA” a un único formato, como “AAAA-MM-DD”. Importancia: Un formato de fecha uniforme garantiza una clasificación, filtrado y análisis de series temporales más sencillos. Los distintos formatos pueden dar lugar a interpretaciones erróneas o errores en los análisis cronológicos.

Manejo de valores nulos

Decidir e implementar un enfoque uniforme para tratar los valores faltantes o indefinidos. Ejemplo: En algunos contextos, puede resultar adecuado reemplazar los valores nulos por la media o la mediana de una columna. En otros escenarios, puede ser preferible la eliminación o imputación mediante técnicas más avanzadas. Importancia: Los valores nulos pueden distorsionar las medidas estadísticas e interferir con los algoritmos. Una estrategia uniforme garantiza que el impacto de estos valores faltantes se aborde de forma sistemática.

La estandarización de datos, aunque parece un paso básico, sienta las bases para un análisis de datos sólido y confiable. Ya sea que se trate de preparar datos para un modelo de aprendizaje automático, un análisis estadístico o un informe comercial, garantizar que los datos estén estandarizados puede mejorar significativamente la calidad y la confiabilidad de la información derivada de esos datos.

1.9 1.2.4 Aplicar técnicas de limpieza y estandarización de datos.

¿Qué es la imputación de datos?

La imputación de datos es un método para retener la mayoría de los datos y la información del conjunto de datos sustituyendo los datos faltantes por un valor diferente. Estos métodos se emplean porque sería poco práctico eliminar datos de un conjunto de datos cada vez. Además, al hacerlo se reduciría sustancialmente el tamaño del conjunto de datos, lo que plantearía preguntas sobre sesgo y perjudicaría el análisis.

Empleamos la imputación porque los datos faltantes pueden provocar los siguientes problemas:

- Distorsiona el conjunto de datos: grandes cantidades de datos faltantes pueden provocar anomalías en la distribución de variables, lo que puede cambiar la importancia relativa de diferentes categorías en el conjunto de datos.
- No se puede trabajar con la mayoría de las bibliotecas de Python relacionadas con el aprendizaje automático: al utilizar bibliotecas de aprendizaje automático (SkLearn es la más popular), pueden ocurrir errores porque no hay un manejo automático de estos datos faltantes.
- Impactos en el modelo final: los datos faltantes pueden generar sesgos en el conjunto de datos, lo que podría afectar el análisis del modelo final.
- Deseo de restaurar todo el conjunto de datos: esto suele ocurrir cuando no queremos perder ninguno (o más) de los datos de nuestro conjunto de datos porque todos son cruciales. Además, si bien el conjunto de datos no es muy grande, eliminar una parte de él podría tener un efecto sustancial en el modelo final.

Dado que hemos explorado la importancia, aprenderemos sobre las diversas técnicas y métodos de imputación de datos.

Técnicas de imputación de datos

Después de aprender qué es la imputación de datos y su importancia, ahora aprenderemos sobre algunas de las diversas técnicas de imputación de datos. Estas son algunas de las técnicas de imputación de datos que analizaremos en profundidad:

- Valor siguiente o anterior
- K vecinos más cercanos
- Valor máximo o mínimo
- Predicción de valor faltante

- Valor más frecuente
- Interpolación promedio o lineal
- Promedio (redondeado) o valor medio o promedio móvil
- Valor fijo

A continuación, exploraremos cada una de estas técnicas en detalle.

1. Valor siguiente o anterior Para los datos de series temporales o datos ordenados, existen técnicas de imputación específicas. Estas técnicas tienen en cuenta la estructura ordenada del conjunto de datos, en la que los valores cercanos son probablemente más comparables que los lejanos. El valor siguiente o anterior dentro de la serie temporal se sustituye normalmente por el valor faltante como parte de un método común para los datos incompletos imputados en la serie temporal. Esta estrategia es eficaz tanto para valores nominales como numéricos.

2. K vecinos más próximos El objetivo es encontrar los k ejemplos más próximos en los datos donde el valor de la característica relevante no esté ausente y luego sustituir el valor de la característica que se produce con mayor frecuencia en el grupo.

3. Valor máximo o mínimo Puede utilizar el mínimo o máximo del rango como el costo de reemplazo de los valores faltantes si sabe que los datos deben encajar dentro de un rango específico [mínimo, máximo] y si sabe, a partir del proceso de recopilación de datos, que el instrumento de medición deja de registrar y el mensaje se satura más allá de uno de dichos límites. Por ejemplo, si se ha alcanzado un precio máximo en un intercambio financiero y el procedimiento de intercambio se ha detenido, el precio faltante se puede sustituir por el valor mínimo del límite del intercambio.

4. Predicción de valores faltantes El uso de un modelo de aprendizaje automático para determinar el valor de imputación final para la característica x en función de otras características es otro método popular para la imputación simple. El modelo se entrena utilizando los valores de las columnas restantes y las filas de la característica x sin valores faltantes se utilizan como conjunto de entrenamiento. Dependiendo del tipo de característica, podemos emplear cualquier modelo de regresión o clasificación en esta situación. En el entrenamiento de resistencia, el algoritmo se utiliza para pronosticar el valor más probable de cada valor faltante en todas las muestras. Se utiliza un enfoque de imputación básico, como el valor medio, para imputar temporalmente todos los valores faltantes cuando hay datos faltantes en más de un campo de característica. Luego, los valores de una columna se restauran a faltantes. Después del entrenamiento, el modelo se utiliza para completar las variables faltantes. De esta manera, se entrena un modelo para cada característica que tenga un valor faltante hasta que un modelo pueda imputar todos los valores faltantes.

5. Valor más frecuente El valor más frecuente en la columna se utiliza para reemplazar los valores faltantes en otra técnica popular que es efectiva tanto para características nominales como numéricas.

6. Interpolación promedio o lineal La interpolación promedio o lineal, que calcula entre el valor accesible anterior y el siguiente y sustituye el valor faltante, es similar a la imputación del valor anterior/siguiente pero solo se aplica a datos numéricos. Por supuesto, al igual que con otras operaciones en datos ordenados, es crucial ordenar con precisión los datos de antemano, por ejemplo, en el caso de datos de series temporales, de acuerdo con una marca de tiempo.

7. Media (redondeada) o promedio móvil o valor mediano La mediana, la media o la media redondeada son otras técnicas de imputación populares para características numéricas. La

técnica, en este caso, reemplaza los valores nulos con valores de media, media redondeada o mediana determinados para esa característica en todo el conjunto de datos. Se recomienda utilizar la mediana en lugar de la media cuando el conjunto de datos tiene una cantidad significativa de valores atípicos.

8. Valor fijo La imputación de valor fijo es una técnica universal que reemplaza los datos nulos con un valor fijo y es aplicable a todos los tipos de datos. Puede imputar los valores nulos en una encuesta utilizando "no respondido" como un ejemplo de uso de imputación fija en características nominales.

Como hemos explorado la imputación simple, su importancia y sus técnicas, aprendamos ahora sobre las imputaciones múltiples.

Manipulación de cadenas

Existen algunas herramientas básicas de manipulación de cadenas que usamos mucho cuando trabajamos con texto:

- Transformar caracteres en mayúsculas a minúsculas (o viceversa).
- Reemplazar una subcadena por otra o eliminar la subcadena.
- Dividir una cadena en partes en un carácter particular.
- Cortar una cadena en ubicaciones específicas.

Mostramos cómo podemos combinar estas operaciones básicas para limpiar los datos de nombres de condados. Recuerde que tenemos dos tablas que queremos unir, pero los nombres de los condados están escritos de manera inconsistente.

Comencemos por convertir los nombres de los condados a un formato estándar.

Conversión de texto a un formato estándar con métodos de cadena de Python

Necesitamos solucionar las siguientes inconsistencias entre los nombres encontrados en las dos tablas:

- Uso de mayúsculas: qui versus Qui.
- Omisión de palabras: County y Parish no aparecen en la tabla del censo.
- Diferentes convenciones de abreviaturas: & versus and.
- Diferentes convenciones de puntuación: St versus St.
- Uso de espacios en blanco: DeWitt versus De Witt.

Cuando limpiamos un texto, suele ser más fácil convertir primero todos los caracteres a minúsculas. Es más fácil trabajar completamente con caracteres en minúscula que intentar rastrear combinaciones de mayúsculas y minúsculas. A continuación, queremos corregir las palabras inconsistentes reemplazando & con and y eliminando County y Parish. Por último, necesitamos corregir las inconsistencias de puntuación y espacios en blanco.

La importancia de evaluar las habilidades de manipulación de cadenas

Evaluar la capacidad de un candidato para manipular cadenas es crucial para las empresas que buscan contratar programadores capacitados. A continuación, se explica el motivo:

- Procesamiento eficiente de datos: las habilidades de manipulación de cadenas permiten a los programadores procesar y manipular datos de texto de manera eficiente. En un mundo impulsado por los datos, la capacidad de trabajar con cadenas de manera eficaz permite un manejo y análisis de datos más fluidos.
- Operaciones sin errores: la competencia en la manipulación de cadenas ayuda a garantizar operaciones precisas y sin errores. Al evaluar las habilidades de manipulación de cadenas de un candidato, las organizaciones pueden identificar a las personas que pueden realizar tareas con precisión y minimizar las posibilidades de introducir errores en el código.
- Resolución flexible de problemas: la manipulación de cadenas a menudo implica dividir problemas complejos en tareas más pequeñas y manejables. Los candidatos que se destacan en esta habilidad demuestran su capacidad de pensar de manera crítica y abordar los problemas de manera estructurada, lo que da como resultado capacidades de resolución de problemas más efectivas.
- Rendimiento de código optimizado: la competencia en la manipulación de cadenas permite a los programadores optimizar el rendimiento del código, mejorando la eficiencia y reduciendo el consumo de recursos. Evaluar las habilidades de manipulación de cadenas de los candidatos garantiza que las personas contratadas puedan entregar un código optimizado y de alta calidad.
- Funcionalidad de aplicación mejorada: muchas aplicaciones dependen en gran medida de la manipulación de datos de texto. Evaluar las habilidades de manipulación de cadenas de los candidatos garantiza que puedan desarrollar aplicaciones con una funcionalidad mejorada, como análisis de datos, formato de texto y extracción de datos, para brindar una experiencia de usuario perfecta.

Al evaluar las habilidades de manipulación de cadenas de un candidato, las organizaciones pueden identificar a las personas que poseen la experiencia necesaria para manejar datos de texto de manera efectiva y contribuir al éxito de sus proyectos de programación.

Aplicaciones prácticas de la manipulación de cadenas

La manipulación de cadenas es un concepto fundamental en la programación que se aplica en diversos dominios. A continuación, se muestran algunos casos de uso prácticos en los que se emplea comúnmente la manipulación de cadenas:

- Limpieza y validación de datos: las técnicas de manipulación de cadenas desempeñan un papel fundamental en la limpieza y validación de datos. Esto incluye eliminar espacios innecesarios, corregir formatos no válidos o filtrar caracteres no deseados. Al manipular cadenas, los programadores pueden garantizar la integridad y precisión de los datos.
- Análisis y extracción de texto: la manipulación de cadenas es esencial para analizar y extraer información de datos textuales. Por ejemplo, en el web scraping, los programadores pueden usar operaciones de manipulación de cadenas para extraer datos específicos de etiquetas HTML o documentos estructurados, lo que les permite recopilar información relevante de manera eficiente.
- Validación de formularios y limpieza de entradas: en el desarrollo web, la manipulación de cadenas es crucial para validar las entradas del usuario. Desde garantizar formatos de correo electrónico correctos hasta limitar la longitud de las entradas, los programadores pueden

emplear técnicas de manipulación de cadenas para limpiar y validar los datos proporcionados por el usuario, mejorando la seguridad y previniendo posibles vulnerabilidades.

- Formato y visualización de cadenas: la manipulación de cadenas se utiliza a menudo para dar formato y mostrar información a los usuarios. Esto puede incluir la personalización de la visualización de fechas, el formato de números o la generación de mensajes dinámicos mediante la incorporación de valores variables en cadenas. Al manipular cadenas, los programadores pueden presentar datos de una manera intuitiva y fácil de usar.
- Operaciones de búsqueda y reemplazo: la manipulación de cadenas es fundamental en las operaciones de búsqueda y reemplazo, lo que permite modificaciones de texto eficientes. Esto puede implicar encontrar palabras o frases específicas dentro de una cadena y reemplazarlas con valores alternativos, lo que lo hace ideal para tareas como generar funciones de autocorrección o realizar reemplazos masivos en grandes conjuntos de datos de texto.
- Manejo y cifrado de contraseñas: las técnicas de manipulación de cadenas están involucradas en el manejo de contraseñas y flujos de trabajo de cifrado. Los programadores pueden manipular cadenas para aplicar políticas de contraseñas, realizar algoritmos de hash o cifrado y transformar información confidencial para garantizar la seguridad y proteger los datos del usuario.

Al comprender y aplicar estrategias de manipulación de cadenas, los programadores pueden optimizar el procesamiento de datos, mejorar las interacciones de los usuarios y garantizar la integridad y seguridad de la información textual en diversas aplicaciones e industrias.

¿Qué son los datos estandarizados?

Los datos estandarizados son datos de diferentes fuentes que se han transformado en un formato uniforme basado en estándares. El proceso de estandarización de datos implica armonizar los datos de modo que todas las entradas de diferentes conjuntos de datos que se relacionan con los mismos términos sigan el mismo formato, lo que permite compararlos de manera significativa.

Los ejemplos de los tipos de formatos de datos que requieren estandarización incluyen:

- Cómo se registran y muestran las direcciones
- Uso de mayúsculas (o no) en los cargos
- Formatos de datos (por ejemplo, elegir entre DD/MM/AA y MM/DD/AA)
- Formatos de hora y zonas horarias utilizados
- Cómo se registran las direcciones de correo electrónico
- Cómo se registran las direcciones de sitios web (por ejemplo, si incluyen o no https://)
- Cómo se registran y muestran los números de teléfono (por ejemplo, con o sin códigos de país)
- Cómo se registran los nombres de los estados (ya sea completos o abreviados)

1.9.1 Explicar el concepto de codificación One-Hot y su aplicación en la transformación de variables categóricas en un formato binario y la preparación de datos para algoritmos de aprendizaje automático

La mayoría de los conjuntos de datos de la vida real que encontramos durante el desarrollo de nuestro proyecto de ciencia de datos tienen columnas de tipo de datos mixto. Estos conjuntos de datos constan de columnas tanto categóricas como numéricas. Sin embargo, varios modelos de aprendizaje automático no funcionan con datos categóricos y para adaptar estos datos al modelo de aprendizaje automático, deben convertirse en datos numéricos. Por ejemplo, supongamos que

un conjunto de datos tiene una columna de género con elementos categóricos como masculino y femenino. Estas etiquetas no tienen un orden de preferencia específico y, además, dado que los datos son etiquetas de cadena, los modelos de aprendizaje automático malinterpretan que existe algún tipo de jerarquía en ellas.

Un enfoque para resolver este problema puede ser la codificación de etiquetas, donde asignaremos un valor numérico a estas etiquetas, por ejemplo, masculino y femenino asignados a 0 y 1. Pero esto puede agregar sesgo en nuestro modelo, ya que comenzará a dar mayor preferencia al parámetro femenino como 1>0, pero idealmente, ambas etiquetas son igualmente importantes en el conjunto de datos. Para solucionar este problema, utilizaremos la técnica de codificación One Hot.

Codificación One Hot

La codificación One Hot es una técnica que utilizamos para representar variables categóricas como valores numéricos en un modelo de aprendizaje automático.

Las ventajas de utilizar la codificación One Hot incluyen:

- Permite el uso de variables categóricas en modelos que requieren una entrada numérica.
- Puede mejorar el rendimiento del modelo al proporcionar más información al modelo sobre la variable categórica.
- Puede ayudar a evitar el problema de la ordinalidad, que puede ocurrir cuando una variable categórica tiene un orden natural (por ejemplo, “pequeño”, “mediano”, “grande”).

Las desventajas de utilizar la codificación One Hot incluyen:

- Puede generar una mayor dimensionalidad, ya que se crea una columna separada para cada categoría en la variable. Esto puede hacer que el modelo sea más complejo y lento de entrenar.
- Puede generar datos dispersos, ya que la mayoría de las observaciones tendrán un valor de 0 en la mayoría de las columnas codificadas One Hot. - Puede provocar un sobreajuste, especialmente si hay muchas categorías en la variable y el tamaño de la muestra es relativamente pequeño.

La codificación one-hot es una técnica eficaz para tratar datos categóricos, pero puede provocar un aumento de la dimensionalidad, escasez y sobreajuste. Es importante utilizarla con cautela y considerar otros métodos, como la codificación ordinal o la codificación binaria.

1.9.2 Explicar el concepto de clasificación y su aplicación en la transformación de variables continuas en variables categóricas

Los **datos continuos** pueden definirse como cualquier dato que pueda tomar valores infinitos dentro de un rango definido, y las diferencias entre los valores son significativas.

Este tipo de datos generalmente representa mediciones, donde la precisión depende del instrumento utilizado.

Algunos ejemplos de datos continuos incluyen:

Tiempo: La duración de un evento, como el tiempo que lleva correr una maratón, se puede medir en horas, minutos, segundos o incluso milisegundos.

Peso: El peso de un objeto se puede medir en kilogramos, gramos o incluso miligramos.

Temperatura: La temperatura se puede medir en grados Celsius o Fahrenheit con distintos niveles de precisión.

Altura: La altura de una persona o un objeto se puede medir en metros, centímetros o milímetros.

Los datos continuos se pueden medir utilizando dos escalas principales:

Escala de intervalo: La escala de intervalo mide datos continuos con intervalos iguales entre los valores, pero no tiene un punto cero verdadero. Esto significa que el valor de cero en esta escala no representa la ausencia del atributo que se está midiendo. Algunos ejemplos de datos de escala de intervalo incluyen la temperatura medida en grados Celsius, donde cero no indica la ausencia de temperatura.

Escala de razón: La escala de razón también mide datos continuos con intervalos iguales entre los valores, pero a diferencia de la escala de intervalo, tiene un punto cero verdadero. Esto significa que el valor de cero en esta escala representa la ausencia del atributo que se está midiendo. Los datos de escala de razón se pueden multiplicar o dividir de manera significativa. Algunos ejemplos de datos de escala de razón incluyen el peso, la altura y el tiempo.

Algunas de las características clave de los datos continuos incluyen:

El número infinito de valores: Los datos continuos pueden adoptar valores infinitos dentro de un rango definido.

Diferencias significativas: Las diferencias entre los valores en los datos continuos son significativas y se pueden utilizar para un análisis posterior.

Representación decimal: Los datos continuos se pueden representar mediante decimales o fracciones, según el nivel de precisión requerido.

Agrupamiento de datos continuos en categorías

El agrupamiento es una técnica que se utiliza para convertir datos continuos en datos categóricos dividiendo el rango de datos en una serie de intervalos o bins y luego asignando cada punto de datos a su bin respectivo. Esto puede ayudar a simplificar los datos y facilitar su análisis y visualización.

Por ejemplo, supongamos que tiene un conjunto de datos de edades de personas, que son datos continuos. Para crear datos categóricos, podría agrupar los datos en grupos de edad (por ejemplo, 0-9, 10-19, 20-29, etc.).

Esto le permitiría analizar la distribución de edades en diferentes grupos de edad y visualizar los datos mediante un gráfico de barras.

Conversión de datos ordinales en datos numéricos

Los datos ordinales, un subtipo de datos categóricos, tienen un orden o clasificación inherente, lo que hace posible convertirlos en datos numéricos. Esto se puede lograr asignando un valor numérico a cada categoría, que representa su clasificación en orden.

Por ejemplo, si tiene datos ordinales que representan los niveles educativos (primaria, secundaria, universidad y posgrado), puede asignar los valores 1, 2, 3 y 4, respectivamente.

Al convertir datos ordinales en datos numéricos, puede aplicar una gama más amplia de métodos estadísticos y capturar mejor las relaciones entre las variables.

Consideraciones al convertir tipos de datos

Al convertir entre datos continuos y categóricos o viceversa, es esencial considerar los siguientes factores para garantizar la precisión y validez de su análisis:

Pérdida de información: La conversión de datos continuos en datos categóricos mediante la clasificación puede resultar en una pérdida de información, ya que categorías más amplias reemplazan los valores precisos. Esto puede afectar la granularidad del análisis y potencialmente enmascarar patrones sutiles en los datos.

Elección de los intervalos o valores numéricos: La elección de los intervalos para clasificar los datos continuos o los valores numéricos para convertir los datos ordinales puede afectar significativamente los resultados del análisis. Es fundamental seleccionar contenedores o valores numéricos que representen con precisión los datos y preserven su estructura inherente.

Supuestos y limitaciones: Los distintos métodos estadísticos tienen supuestos y limitaciones específicos según el tipo de datos que están diseñados para manejar. Al convertir los tipos de datos, asegúrese de que los nuevos datos sigan cumpliendo los supuestos de los métodos estadísticos que planea utilizar.

Interpretabilidad: Al convertir los tipos de datos, es esencial mantener la interpretabilidad de los resultados. Asegúrese de que los datos convertidos sigan representando con precisión los datos originales y de que los resultados se puedan comunicar claramente a su audiencia.

pandas.cut

Agrupa valores en intervalos discretos.

Utiliza cut cuando necesites segmentar y ordenar valores de datos en intervalos. Esta función también es útil para pasar de una variable continua a una variable categórica. Por ejemplo, cut podría convertir edades en grupos de rangos de edad. Admite la agrupación en un número igual de intervalos o en una matriz de intervalos preestablecida.

[9]:
```
import pandas as pd
import numpy as np

y = pd.cut(np.array([1, 7, 5, 4, 6, 3]), 3)
y
```

[9]:
```
[(0.994, 3.0], (5.0, 7.0], (3.0, 5.0], (3.0, 5.0], (5.0, 7.0], (0.994, 3.0]]
Categories (3, interval[float64, right]): [(0.994, 3.0] < (3.0, 5.0] < (5.0,
7.0]]
```

[10]:
```
# Discovers the same bins, but assign them specific labels. Notice that the␣
 ↪returned Categorical's categories are labels and is ordered.

x = pd.cut(np.array([1, 7, 5, 4, 6, 3]), 3, labels=["bad", "medium", "good"])
x
```

[10]:
```
['bad', 'good', 'medium', 'medium', 'good', 'bad']
Categories (3, object): ['bad' < 'medium' < 'good']
```

Creado por: