Introducción a la Ciencia de Datos

Conceptos básicos, flujo de trabajo y ejemplo práctico con K-means

hack(io)

Definición y Alcance

Definición

La ciencia de datos es la disciplina que utiliza métodos científicos, procesos, algoritmos y sistemas para extraer conocimiento y patrones significativos a partir de datos estructurados y no estructurados.

Alcance

El alcance de la ciencia de datos abarca análisis estadístico, aprendizaje automático, visualización de datos, minería de datos, Big Data y otras técnicas para comprender y analizar datos complejos.

Importancia

La ciencia de datos es crucial en la toma de decisiones empresariales, la predicción de tendencias, la optimización de procesos y la mejora de la eficiencia operativa en diversas industrias.

Aplicaciones

La ciencia de datos se aplica en la personalización de experiencias de usuario, la detección de fraudes, la optimización de campañas publicitarias, la medicina predictiva, la ciencia climática y muchos otros campos.

El reto: Segmentación Basada en el Comportamiento de Reclamaciones

Agrupar a los clientes en base a su historial de reclamaciones para identificar patrones de riesgo y mejorar la gestión de reclamaciones.
Para esto tendréis que seguir los siguientes pasos:

uno

Decidir que columnas de todas las que tenemos son importantes para resolver el problema. Justificar porque seleccionasteis esas columnas.

dos

Estandarizar las variables para hacerlas comparables.

tres

Aplicar el modelo de clustering a través de la página web compartida con vosotros. El link lo tenéis aquí.

cuatro

Interpretación de Clústeres.

¿Qué es el Clustering?

Concepto

El clustering, o agrupamiento, es una técnica de análisis de datos que busca clasificar un conjunto de datos en subgrupos o 'clusters' que compartan características similares entre sí, en base a la distancia entre ellos en un espacio multidimensional.

Importancia

El clustering es fundamental en el análisis de datos, ya que permite identificar patrones, tendencias o grupos específicos dentro de los datos, lo que facilita la toma de decisiones estratégicas y la segmentación de mercados, entre otros usos.

Aplicaciones

El clustering se utiliza en diversas áreas, como marketing, bioinformática, reconocimiento de patrones, entre otras, para comprender la estructura subyacente de los datos y descubrir información útil.

Estandarización

La estandarización es un proceso que ajusta los datos para que tengan una media de cero y una desviación estándar de uno. Esto se logra restando la media y dividiendo por la desviación estándar.

Es decir, la estandarización transforma diferentes tipos de datos a una misma escala. Así, valores grandes y pequeños se comparan fácilmente. Como ajustar el volumen de la música para que todas las canciones suenen igual.

En modelos de clustering, la estandarización es crucial para que las variables con diferentes escalas tengan un impacto equitativo en el resultado final. En modelos de clustering, la estandarización es esencial para asegurar que las variables con escalas distintas contribuyan de manera equitativa a la agrupación de datos. Utilizar datos sin estandarizar puede llevar a que la influencia de variables con escalas más grandes domine el proceso de agrupación, generando clusters sesgados e inexactos.

Fórmula de Estandarización

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

Donde:

Z es el valor estandarizado.

x es el valor original del dato.

μ es la media del conjunto de datos.

σ es la desviación estándar del conjunto de datos.

Ideas para visualización

(podéis usar la plataforma que queráis, Power BI, Excel, Looker, Google Sheets, etc)

· Gráfico de Barras para Cobertura: Comparar la distribución de tipos de cobertura (básica, extendida, premium) entre clústeres.

· Histograma de Meses Desde la Última Reclamación: Mostrar cómo se distribuyen los clientes en cada clúster según el tiempo desde su última reclamación.

· Gráfico de Barras para Número de Quejas Abiertas: Comparar la frecuencia de diferentes números de quejas abiertas entre clústeres.

· Gráfico de Dispersión Total Claim Amount vs. Months Since Last Claim: Ver la relación entre el monto total reclamado y los meses desde la última reclamación en cada clúster.

Algunos recursos interesantes para saber que tipo de grafico usar :

- Data to Viz: https://www.data-to-viz.com/
- Data Viz Project: https://datavizproject.com/