CRISP-DM aplicado a la representación estructurada de Municipios en Boyacá

Introducción

El presente documento detalla el proceso llevado a cabo bajo la metodología CRISP-DM para caracterizar los municipios del departamento de Boyacá, Colombia. El objetivo principal es utilizar variables escalares para representar de manera completa a dichos municipios. Se busca identificar patrones y relaciones entre estas variables para obtener una comprensión profunda de la diversidad y características de los municipios seleccionados.

Comprensión del Negocio

El objetivo del proyecto es caracterizar los municipios de Boyacá utilizando variables escalares como población, altitud, área, actividad económica, educación, entre otras. Esto permitirá tener una comprensión más completa y detallada de los municipios, lo que podría llegar a ser útil para la toma de decisiones en áreas como el desarrollo socioeconómico, la planificación urbana y la asignación de recursos.

Comprensión de los Datos

Origen de los Datos

Los datos utilizados fueron obtenidos de diversas fuentes, incluyendo bases de datos de la Gobernación de Boyacá, el Departamento Administrativo Nacional de Estadística (DANE) y datos abiertos del Gobierno de Colombia.

Descripción de las Variables

Se dispone de una serie de variables que incluyen información demográfica, económica, educativa y geográfica de los municipios. Entre las variables relevantes se encuentran:

- Población total
- Altitud
- Área territorial
- Actividades económicas (primarias, secundarias y terciarias)
- Número de instituciones educativas
- Matriculados en educación pública y privada
- Promedio de puntaje global en pruebas educativas

Preparación de los Datos

Limpieza y Curación de Datos

Antes de proceder con el análisis, se llevó a cabo un proceso de limpieza y curación de datos para garantizar la calidad y coherencia de los mismos. Esto incluyó la identificación y tratamiento de valores faltantes, errores de formato y estandarización de los datos.

Modelado

El modelo utilizado para este proyecto es el Análisis de Componentes Principales, una técnica de reducción de dimensionalidad que permite identificar patrones y relaciones en un conjunto de datos mediante la transformación de variables originales en un nuevo conjunto de variables no correlacionadas llamadas componentes principales.

Evaluación

Resultados del PCA

Tras aplicar PCA a los datos, se obtuvieron nuevos componentes principales que capturan la variabilidad de las variables originales. Estos componentes pueden interpretarse en términos del contexto de las variables originales y ayudan a reducir la dimensionalidad del conjunto de datos.

Despliegue

Interpretación de los Componentes Principales

Los nuevos componentes principales pueden ser interpretados en función de las variables originales que contribuyen más a su formación. Se realizaron análisis y visualizaciones adicionales para comprender mejor la estructura de los datos y posibles agrupaciones de municipios.

Conclusión

El proceso de caracterización de municipios mediante variables escalares ha permitido obtener una representación detallada y comprensiva de la diversidad y características de los municipios de Boyacá. Los resultados del PCA proporcionan insights valiosos que pueden ser utilizados por las autoridades locales y otros actores interesados en la toma de decisiones relacionadas con el desarrollo y la planificación de la región.

CRISP-DM aplicado a la Representación No Estructurada de Datos de Celebridades

Introducción

El análisis de datos no estructurados, especialmente el texto, juega un papel vital en el entendimiento de las percepciones y las influencias culturales de figuras públicas y celebridades. La representación no estructurada permite transformar texto en formatos analizables, facilitando la extracción de insights significativos. Este documento se enfoca en la aplicación de técnicas de representación no estructurada en el análisis de textos sobre personajes latinoamericanos, siguiendo el marco del proceso CRISP DM.

Comprensión del Negocio

El propósito de este análisis es entender las relaciones, las diferencias y las similitudes entre diferentes artistas latinoamericanos basados en su representación textual. Esto puede apoyar en la identificación de patrones culturales, influencias y la popularidad relativa de los artistas.

Comprensión de los Datos

Los datos consisten en textos relacionados con 40 artistas latinoamericanos. La fuente de estos textos es Wikipedia. Esta fase implica explorar y entender la naturaleza del contenido textual disponible para cada artistas.

Preparación de los Datos

La preparación de los datos en este contexto implica la limpieza de texto (eliminación de marcadores de formato, caracteres especiales) y la transformación del texto a una forma adecuada para el análisis. Esto incluye la creación de representaciones TF-IDF, que reflejan la importancia de las palabras dentro de los documentos en relación con una colección de documentos, y la creación de representaciones basadas en embeddings, que capturan el contexto y las relaciones semánticas entre palabras.

Modelado

Utilizando las representaciones preparadas, se aplica t-SNE (t-distributed Stochastic Neighbor Embedding) para reducir la dimensionalidad de los datos y visualizar las relaciones entre los personajes en un espacio bidimensional o tridimensional. Este modelado ayuda a identificar agrupaciones y patrones en las representaciones textuales de los personajes.

Evaluación

La evaluación se centra en interpretar el mapa generado por t-SNE, analizando cómo los personajes se agrupan según sus similitudes textuales. Esto puede revelar insights sobre las conexiones culturales, profesionales o ideológicas entre los artistas, así como identificar posibles outliers o personajes únicos en el conjunto de datos.

Despliegue

Los insights generados se pueden utilizar para informar estrategias de contenido, marketing cultural, o incluso estudios académicos sobre la influencia cultural de figuras públicas latinoamericanas. La visualización resultante ofrece una herramienta poderosa para explorar las relaciones entre estos personajes de una manera intuitiva y accesible.

Conclusión

La aplicación de técnicas de representación no estructurada, enmarcadas dentro del proceso CRIS DM, al análisis de textos sobre personajes latinoamericanos ofrece una metodología robusta para desentrañar las complejidades y las riquezas de las influencias culturales y sociales. Este enfoque no solo mejora nuestra comprensión de las figuras públicas, sino que también destaca la importancia de los métodos avanzados de análisis de datos en la interpretación de la cultura y la sociedad.