Análisis de Componentes Principales con Web Mining

Centro de Ciencias Básicas

Asignatura:

"Inteligencia Artificial"

Maestro:

Francisco Javier Luna Rosas

Alumnos:

- José Alfredo Díaz Robledo
- Luis Pablo Esparza Terrones
- Luis Manuel Flores Jiménez
- Juan Francisco Gallo Ramírez

Ingeniería en Computación Inteligente

3er Semestre

Introducción

El presente proyecto tiene como objetivo la aplicación del Análisis de Componentes Principales (ACP) a un conjunto de datos que contiene información detallada sobre los sitios del patrimonio mundial de la UNESCO que se encuentran en peligro de desaparecer. La UNESCO, como entidad líder en la protección y preservación del patrimonio cultural y natural, mantiene una lista de sitios de relevancia global que enfrentan diversos tipos de amenazas.

El ACP es una herramienta estadística poderosa que nos permite explorar la estructura subyacente de los datos, identificar patrones y relaciones ocultas, y comprender mejor la situación de los sitios en peligro. Este análisis contribuirá a la evaluación y comprensión de las amenazas que enfrentan estos sitios y, potencialmente, aportará información valiosa para su protección y preservación.

A lo largo de este documento, se presentarán conceptos clave relacionados con el ACP, como el Plano de Similitud, el Círculo de Correlaciones y el teorema de Dualidad. Estos conceptos son esenciales para comprender y contextualizar el análisis.

Además, se mostrarán los resultados del ACP aplicado a los datos específicos de los sitios en peligro de la UNESCO. Se presentarán hallazgos, patrones y tendencias que ayudarán a una comprensión más profunda de la situación actual de estos sitios.

Este proyecto se desarrollará siguiendo las pautas establecidas y se estructurará de manera clara y concisa, incluyendo una sección de evidencias de la implementación del ACP, conclusiones que resumen los resultados más significativos y una sección de referencias de acuerdo con el formato APA.

El análisis propuesto tiene como objetivo contribuir a la misión de la UNESCO de preservar el valioso patrimonio mundial y proporcionar información valiosa para la toma de decisiones relacionadas con la conservación de estos sitios en peligro.

Evidencias

Conceptos básicos del ACP:

El Análisis de Componentes Principales es una técnica de reducción de dimensionalidad utilizada en estadísticas y análisis multivariados. Su objetivo principal es transformar un conjunto de variables originales en un nuevo conjunto de variables no correlacionadas llamadas "componentes principales". Estos componentes principales capturan la máxima variabilidad en los datos y permiten una representación más compacta de la información.

Plano de Similitud:

El Plano de Similitud es una representación gráfica que muestra la proximidad entre individuos (observaciones) en función de sus similitudes o diferencias en el espacio de los componentes principales. En este plano, los individuos que están cerca uno del otro tienen perfiles de variables similares, mientras que los individuos que están alejados tienen perfiles de variables diferentes. El Plano de Similitud es útil para identificar patrones de agrupación o dispersión en los datos.

Círculo de Correlaciones:

El Círculo de Correlaciones es un gráfico que ilustra la relación entre las variables originales y los componentes principales. Cada variable se representa como un vector que parte del origen y apunta en la dirección en la que esa variable tiene una mayor influencia en el componente principal. La longitud del vector indica la fuerza de esa influencia. Este gráfico ayuda a comprender cómo las variables originales contribuyen a la variabilidad capturada por los componentes principales y a identificar las variables más relevantes en cada componente.

Teorema de Dualidad:

El Teorema de Dualidad es un resultado matemático fundamental en el ACP que establece que los individuos y las variables se pueden analizar de manera equivalente en el mismo espacio de componentes principales. En otras palabras, cualquier análisis que se realice en los individuos (por ejemplo, un Plano de Similitud) también se puede realizar en las variables (por ejemplo, un Círculo de Correlaciones) y viceversa. Este teorema es importante porque muestra que la interpretación de los resultados es consistente, independientemente de si se enfoca en los individuos o en las variables.

Análisis

El programa realiza un análisis de componentes principales (ACP) en un conjunto de datos que contiene información sobre sitios del patrimonio mundial de la UNESCO que se encuentran en peligro de desaparecer. El análisis se lleva a cabo utilizando bibliotecas como requests, BeautifulSoup, pandas, sklearn, y matplotlib. A continuación, se ofrece un análisis detallado del programa:

- **Importación de Bibliotecas:**

El programa comienza importando las bibliotecas necesarias. requests se utiliza para realizar solicitudes web y obtener datos de una URL. BeautifulSoup es una biblioteca de análisis HTML que permite extraer información de páginas web. pandas se utiliza para la manipulación de datos tabulares, y PCA de sklearn se utiliza para realizar el Análisis de Componentes Principales. matplotlib es una biblioteca de visualización para crear gráficos.

Obtención de Datos desde Wikipedia:

El programa comienza accediendo a una página web específica de Wikipedia que contiene información sobre los sitios en peligro de la UNESCO.

- **Extracción y Limpieza de Datos:**

La biblioteca requests se utiliza para obtener el contenido HTML de la página web. Luego, se utiliza BeautifulSoup para buscar y extraer la tabla de datos relevante. Los datos se almacenan en un DataFrame de pandas, se eliminan las columnas innecesarias y se realizan transformaciones en las celdas para limpiar los datos.

- **Análisis de Componentes Principales (ACP):**

El ACP se aplica a los datos utilizando PCA de sklearn. El número de componentes principales se establece en 4 en este caso.

- **Visualización de Resultados:**

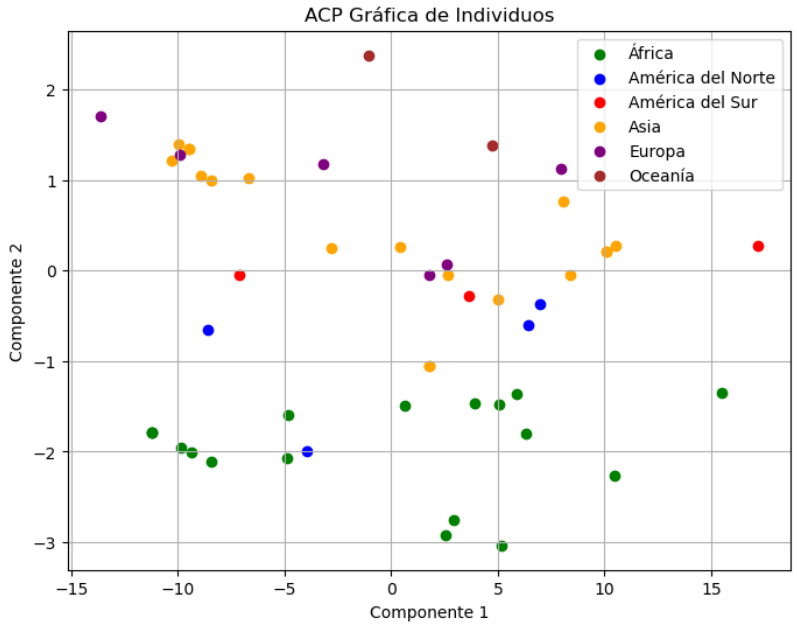
El programa genera dos gráficos. El primero muestra un gráfico de dispersión de individuos en el plano definido por los componentes principales 1 y 2. El segundo gráfico es un círculo de correlaciones que muestra la relación entre las variables originales y los componentes principales 3 y 4.

- **Interpretación y Conclusiones:**

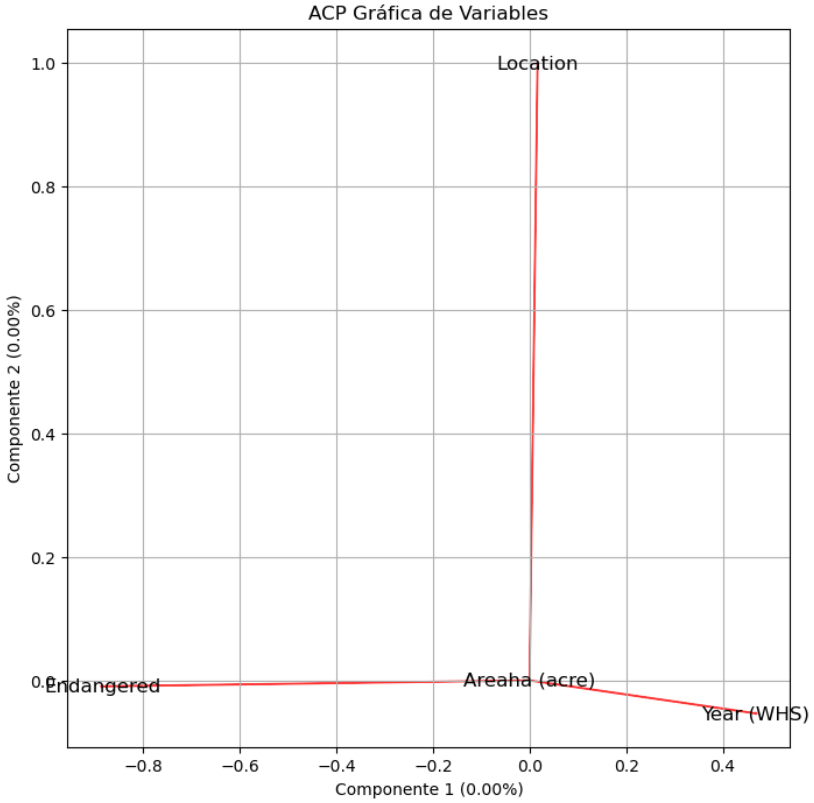
A partir de los gráficos y resultados obtenidos, se pueden derivar conclusiones sobre la distribución geográfica de los sitios, la influencia de las variables en los componentes principales y otras observaciones relevantes.

En resumen, el programa realiza un ACP en datos relacionados con sitios del patrimonio mundial en peligro y proporciona visualizaciones que ayudan a comprender la estructura de los datos. El análisis de los resultados y las conclusiones requiere una evaluación más profunda y contextualizada, pero el programa establece las bases para explorar y comprender mejor la situación de estos sitios en peligro.

Gráfica de Individuos:



Gráfica de Variables:



Implementación

La implementación se realizó mediante el uso de Jupyter Notebooks, usando lenguaje Python.

- ACP_WebMining.html

Conclusiones

Por medio del análisis de componentes principales (ACP) aplicado a la lista de sitios del patrimonio mundial de la UNESCO en peligro, se obtuvieron conclusiones significativas. En primer lugar, se realizó una adecuada preparación de los datos, extrayendo y limpiando la información de la lista de sitios en peligro disponible en Wikipedia.

La aplicación del ACP permitió reducir la dimensionalidad de los datos y revelar patrones latentes en la información. El análisis se centró en cuatro componentes principales que capturaron la variabilidad subyacente en los datos.

A través de la visualización de datos, se observó que los sitios tendían a agruparse por continente en el gráfico de dispersión de individuos, lo que sugiere una influencia geográfica en las amenazas a los sitios del patrimonio mundial. Esta observación inicial plantea preguntas sobre las razones detrás de estas tendencias geográficas y cómo pueden afectar a la preservación de estos sitios.

Adicionalmente, el círculo de correlaciones reveló las variables clave que más influyen en los componentes principales. Esta información es fundamental para identificar las amenazas principales a los sitios y comprender mejor los factores que contribuyen a su estado de peligro.

Es importante destacar que estas conclusiones iniciales brindan una visión general del análisis, pero para obtener resultados definitivos y acciones orientadas a la preservación de estos sitios, se requiere un análisis más profundo y contextualizado de los datos. El ACP es una herramienta valiosa para la exploración inicial de la estructura de los datos y la identificación de patrones, pero su interpretación completa requiere un análisis más detallado y una comprensión más profunda del contexto.

Referencias

No se consultaron fuentes.