ACTIVIDAD - SEMANA 4
REDUCCIÓN DE DATOS
ANÁLISIS DE COMPONENTES PRINCIPALES (PCA)

EQUIPO 68

DENISSE MARÍA RAMÍREZ COLMENERO A01561497
EMMANUEL GONZÁLEZ CALITL A01320739

MATERIA: CIENCIA Y ANALÍTICA DE DATOS

PROFESOR TITULAR: JOBISH VALLIKAVUNGAL DEVASSIA
PROFESOR TUTOR: ORLANDO FIGÓN CRUZ

FECHA: 11 DE OCTUBRE DE 2022

—----------

Transformación de datos:
Una transformación es el proceso de cambiar el formato y/o estructura de un dato, para que así estos tengan una misma escala y puedan ser comparados, además de que ayuda a una mejor organización de la base de datos y aumenta la calidad de los datos, pues rectifica los valores nulos, las entradas duplicadas, los defectos y los formatos incorrectos.
Un ejemplo de transformación de datos es cuando se convierten datos de caracteres en su formato equivalente ASCII.

Reducción de datos:
Por otro lado, la reducción de datos es el proceso de obtener una representación reducida de un conjunto de datos con el fin de reducir su volumen y simplificar el proceso de análisis del conjunto de datos. Lo más importante de una reducción de datos es que mantiene toda información importante, pues solo elimina información repetida o sin valor.
Para reducir datos de un conjunto se pueden utilizar diferentes estrategias como: muestreo, selección de características o reducción de dimensionalidad.

PCA:
El análisis de componentes principales es un tipo de reducción de dimensionalidad donde se transforman las columnas de un conjunto de datos (variables) en un nuevo conjunto de características llamados componentes principales. Estos componentes principales están formados de información proveniente de las variables originales del dataset. Los componentes principales de un dataset son menos que las variables originales, por lo que la información se comprime en menos columnas de características, lo que hace que las dimensiones de un dataset se reduzcan.

Los componentes principales de un conjunto de datos se enlistan de forma que el primer componente es el que mayor cantidad de varianza explicada tiene, es decir, más información aporta.

Valores atípicos:
Un valor atípico es una observación del conjunto de datos que se encuentra muy lejos del resto de las observaciones, es decir, que es mucho más grande o mucho más pequeño. Los valores atípicos ocasionan una disminución en la precisión del análisis estadístico. Para tratar los valores atípicos existen varias opciones: eliminación de los valores atípicos, imputación de media/mediana, revestimientos, entre otras.

Para identificar los valores atípicos en un conjunto de datos con gran cantidad de registros, se utilizan técnicas de visualización y matemáticas.

PARTE 2:

1. **¿Cuál es el número de componentes mínimo y por qué?**

   No existe un método único que permita identificar cual es el número óptimo de componentes principales a utilizar. La forma más confiable es seleccionar los componentes principales a partir de observar la varianza explicada acumulada y seleccionar los componentes en el momento en que se tiene un porcentaje alto de esta varianza.

   La cantidad de componentes que se forman es igual al menor número de variables originales o al número de observaciones menos uno, y a partir de obtener todos los componentes se seleccionan el menor número de componentes que acumulen un porcentaje alto de varianza y estos se convierten en los componentes más importantes.

   En nuestro caso se formaron 15 componentes pero solo se seleccionaron 9 pues entre estos acumulaban un 92% de la varianza explicada.

2. **¿Cuál es la variación de los datos que representan esos componentes?**

   Los componentes están enlistados en orden descendente de su varianza explicada, es decir, al porcentaje que tiene cada componente de la información total del dataset. Esto quiere decir que el primer componente siempre contribuye o da información en mayor medida. La variación explicada acumulada nos arroja la suma de los porcentajes que cada componente contribuye, es por eso que con la contribución de todos los componentes se tendría el 100% de la varianza.

   En nuestro caso, el primer componente contribuye en un 39.47%, mientras que el componente número nueve contribuye en un 4.84%. Juntando los porcentajes de los primeros nueve componentes, que son nuestros componentes principales, juntamos un 92.64% de la varianza total.

3. **¿Cuál es la pérdida de información después de realizar PCA?**

   Sabemos que al proyectar las dimensiones en una menor dimensión, la información total de nuestro conjunto de datos se ve comprometida. En este caso, ya que mantuvimos un 92.64% de la variación, un 7.36% aproximadamente se perderá después de realizar el PCA. Entre menos componentes principales se quieran utilizar, menos va a ser la información mantenida. Sin embargo, si hablamos de un conjunto de datos con una gran cantidad de registros, la información disponible seguiría siendo

mucha con la ventaja de que el proceso sería más rápido y sencillo pues la dimensión del conjunto ahora es mucho menor.

4. **De las variables originales, ¿Cuál tiene mayor y cuál tiene menor importancia en los componentes principales?**

   La variable Credible? tiene un peso en total de 3.024 puntos (sumando los valores absolutos de cada término), y esta variable junto a Pp-Jun son los que mayor peso tienen en dos de los componentes principales. La variable con menor importancia sería Bs-May pues es el que tiene menor peso en total en los componentes con solo 0.522 puntos.

5. **¿Cuándo se recomienda realizar un PCA y qué beneficios ofrece para Machine Learning?**

   Se recomienda utilizar el método PCA (Principal Component Analysis) las dimensiones de las características de entrada son muy altas y se requiere reducir la dimensionalidad o variables dentro de nuestro set de datos, dado que estas variables están correlacionadas. Al lograr esto se reducen tiempos de procesamiento y de complejidad matemática para los cálculos y visualización.
   También se puede utilizar para reducir el ruido y comprimir los datos.

REFERENCIAS:

Transformación de datos - BizTalk Server. (2022, 24 septiembre). *Microsoft Learn*. Recuperado 10 de octubre de 2022, de https://learn.microsoft.com/es-es/biztalk/core/data-transformation

Prabhakaran, S. (2022, 20 abril). *Principal Component Analysis (PCA) - Better Explained | ML+*. Machine Learning Plus. Recuperado 11 de octubre de 2022, de https://www.machinelearningplus.com/machine-learning/principal-components-analysis-pca-better-explained/

Recursos: Big Data. (2021, 31 agosto). Datapeaker. Recuperado 11 de octubre de 2022, de https://datapeaker.com/big-data/recursos-big-data/

Reducción de la dimensionalidad: Análisis de Componentes Principales (PCA). (2022, 21 enero). profesorDATA.com. Recuperado 11 de octubre de 2022, de https://profesordata.com/2020/09/01/reduccion-de-la-dimensionalidad-analisis-de-componentes-principales-pca/

Liu, Y. (2020). *Python Machine Learning By Example: Build intelligent systems using Python, TensorFlow 2, PyTorch, and scikit-learn, 3rd Edition*. Van Haren Publishing.