PUCE Virtual
EDUCACIÓN FLEXIBLE

Adquisición, gestión y gobernanza de datos

Preprocesamiento de datos

Clase 2

MAESTRÍA EN
SISTEMAS DE INFORMACIÓN
Mención Data Science

La excelencia no se improvisa

INTRODUCCIÓN DE LA CLASE

El preprocesamiento de datos es una etapa crucial en el análisis y la gestión de datos, ya que prepara la información cruda para su uso en modelos analíticos y de machine learning. Consiste en una serie de técnicas que transforman y limpian los datos, asegurando que sean consistentes, completos y adecuados para el análisis. Este proceso incluye la eliminación de valores atípicos, la imputación de datos faltantes y la normalización de características, lo que facilita la obtención de resultados más precisos y confiables.

Además, el preprocesamiento aborda la calidad de los datos, un aspecto fundamental para la toma de decisiones informadas. Las técnicas utilizadas no solo ayudan a mejorar la precisión de los modelos, sino que también optimizan el rendimiento computacional. Al aplicar métodos como la codificación de variables categóricas y la escalación de características, los analistas pueden asegurar que los datos se ajusten a los requisitos específicos de los algoritmos, lo que, a su vez, potencia la efectividad de los proyectos de análisis de datos y gobernanza.

RDA2:

Entender cada una de las etapas del proceso de Descubrimiento de Conocimiento en el contexto de Big Data.

2. Preprocesamiento de datos

El preprocesamiento de datos es una etapa crucial en el análisis de datos y la minería de datos, ya que prepara los datos brutos para su posterior análisis. Esta fase implica una serie de técnicas y transformaciones que aseguran que los datos sean precisos, consistentes y adecuados para los modelos analíticos que se aplicarán. Sin un adecuado preprocesamiento, los resultados de los análisis pueden ser engañosos o inexactos. A continuación, se presentan algunos ejemplos de las técnicas comunes de preprocesamiento y su relevancia en diferentes contextos.

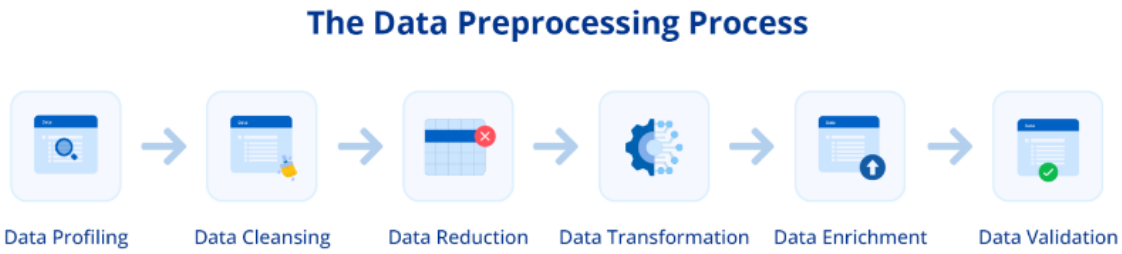


Figura 1. ***Data Preparation Process*****.**
Fuente: https://www.astera.com/es/type/blog/data-preprocessing/

Limpieza de Datos

Uno de los primeros pasos en el preprocesamiento es la limpieza de datos, que se refiere a la identificación y corrección de errores o inconsistencias en los datos. Esto puede incluir la eliminación de duplicados, el manejo de valores perdidos y la corrección de errores tipográficos.

Ejemplo: en un conjunto de datos de ventas, puede haber registros duplicados para una misma transacción. La limpieza de datos implica eliminar estos duplicados para asegurar que cada venta se cuente solo una vez. Además, si un registro tiene un valor perdido en el campo de la cantidad vendida, se puede optar por imputar un valor promedio o eliminar el registro, dependiendo de la cantidad de datos faltantes (Rahm & Do, 2000).

Reducción de Datos

La reducción de datos busca simplificar el conjunto de datos original mediante la eliminación de variables irrelevantes o redundantes, lo que puede mejorar la eficiencia del análisis. Esta etapa es crucial cuando se trabaja con conjuntos de datos grandes que contienen muchas variables.

Ejemplo: en un análisis de satisfacción del cliente, se puede contar con cientos de preguntas en una encuesta. Algunas de estas preguntas pueden ser redundantes o tener correlaciones altas. Al aplicar técnicas de reducción de dimensionalidad, como el Análisis de Componentes Principales (PCA), se pueden combinar varias variables relacionadas en un número menor de componentes que retengan la variabilidad más significativa del conjunto de datos original (Jolliffe & Cadima, 2016).

Transformación de Datos

La transformación de datos es la etapa en la que los datos son convertidos a un formato adecuado para el análisis. Esto puede incluir normalización, estandarización o codificación de variables categóricas.

Ejemplo: En un conjunto de datos que contiene información de clientes, donde se incluyen ingresos en diferentes escalas (algunos en miles y otros en millones), la normalización es esencial. Se pueden transformar los ingresos a una escala común, por ejemplo, utilizando una escala de 0 a 1. Esto ayuda a asegurar que todas las variables se consideren de manera equitativa en el análisis (Khan et al., 2019).

Enriquecimiento de datos

El enriquecimiento de datos implica agregar información adicional que puede mejorar la calidad y el valor de los datos existentes. Esta etapa es particularmente útil para crear un contexto más completo que permita un análisis más profundo.

Ejemplo: una empresa que recopila datos de ventas puede enriquecer sus datos de clientes incorporando información demográfica, como edad, nivel educativo y ubicación geográfica. Al agregar estos atributos, la empresa puede segmentar mejor a sus clientes y diseñar estrategias de marketing más efectivas basadas en el comportamiento de compra (Bizer et al., 2009).

Validación de datos

La validación de datos es la etapa final del preprocesamiento y se refiere a la verificación de que los datos cumplen con criterios específicos de calidad y precisión. Esta etapa es esencial para garantizar que los datos sean confiables antes de ser utilizados en análisis posteriores.

Ejemplo: después de haber limpiado, reducido, transformado y enriquecido un conjunto de datos, es crucial realizar una validación. Esto puede incluir la verificación de que todos los valores están dentro de rangos aceptables, que no existen duplicados y que los formatos de datos son consistentes (Pipino, Lee, & Wang, 2002). Por ejemplo, si un campo de fecha debe seguir un formato específico (como DD/MM/YYYY), se debe comprobar que todos los registros cumplan con este formato.

Importancia del preprocesamiento de datos

El preprocesamiento de datos es un paso crítico en el análisis de datos y la minería de datos, ya que se encarga de transformar los datos brutos en un formato adecuado para el análisis posterior. La calidad de los datos es fundamental, ya que datos erróneos o mal estructurados pueden llevar a conclusiones inexactas y decisiones equivocadas. A continuación, se presentan ejemplos que destacan la importancia del preprocesamiento de datos en diversas aplicaciones y contextos.

Mejora de la Calidad de los Datos

Uno de los principales beneficios del preprocesamiento es la mejora de la calidad de los datos. Esto se logra mediante la limpieza de datos, que implica identificar y corregir errores, como duplicados o valores perdidos.

Ejemplo: en un estudio de mercado, una empresa puede recopilar datos de encuestas para entender las preferencias de los consumidores. Si el conjunto de datos incluye múltiples registros para la misma persona debido a errores de entrada, esto puede llevar a una interpretación errónea de las preferencias. Al limpiar estos datos, la empresa puede obtener una visión más precisa y representativa del comportamiento del consumidor (Rahm & Do, 2000).

Aumento de la Eficiencia Analítica

El preprocesamiento también contribuye a la eficiencia analítica. Al reducir el volumen de datos mediante la eliminación de atributos irrelevantes o redundantes, se facilita el análisis.

Ejemplo: en un análisis de rendimiento académico en instituciones educativas, puede existir un conjunto de datos con cientos de variables sobre estudiantes, cursos y calificaciones. Si se eliminan las variables que no aportan información significativa, como comentarios de texto no estructurados, los analistas pueden concentrarse en las variables que realmente afectan el rendimiento. Esto no solo acelera el proceso de análisis, sino que también mejora la capacidad de interpretación de los resultados (Jolliffe & Cadima, 2016).

Facilita el Aprendizaje Automático

El preprocesamiento de datos es esencial en el ámbito del aprendizaje automático. Muchos algoritmos requieren que los datos estén en un formato específico y que sean normalizados o estandarizados.

Ejemplo: en un modelo predictivo para la detección de fraudes en transacciones financieras, es crucial que los datos de las transacciones sean normalizados. Si los datos de las transacciones incluyen montos que varían en rangos muy diferentes, esto puede afectar el rendimiento del modelo. La normalización asegura que todos los atributos tengan un impacto equitativo en el resultado final del modelo (Khan et al., 2019)

Mejora de la Interpretabilidad

El preprocesamiento también mejora la interpretabilidad de los resultados. Al transformar los datos en un formato más comprensible, los analistas pueden comunicar sus hallazgos de manera más efectiva.

Ejemplo: en un análisis de satisfacción del cliente, convertir las respuestas de encuestas en datos categóricos (como "satisfecho", "neutral" o "insatisfecho") permite que los

resultados sean más fáciles de entender para los tomadores de decisiones. Esto ayuda a las empresas a identificar áreas de mejora de manera clara y concisa (Feldman & Sanger, 2007).

Validación y Confiabilidad

El preprocesamiento de datos asegura que los datos sean válidos y confiables, lo que es esencial para la toma de decisiones informadas.

Ejemplo: en estudios científicos, los investigadores deben asegurarse de que los datos recopilados sean precisos y válidos antes de realizar cualquier análisis. El preprocesamiento, que incluye la validación de datos, permite a los investigadores verificar que los datos cumplan con criterios específicos de calidad. Esto aumenta la confianza en los resultados y las conclusiones del estudio (Pipino, Lee, & Wang, 2002).

Detección de Outliers

La detección de outliers, o valores atípicos, es un componente crítico en el análisis de datos, ya que estos puntos pueden influir de manera significativa en la validez y precisión de los resultados obtenidos a partir de un conjunto de datos. Un outlier se define como un dato que se encuentra considerablemente alejado de otros valores en un conjunto. Esta anomalía puede deberse a variaciones naturales en el fenómeno que se está estudiando, errores en la recolección de datos o condiciones extraordinarias. Identificar y tratar adecuadamente los outliers es fundamental para garantizar la integridad de los análisis posteriores.

Importancia de la Detección de Outliers

Los outliers pueden distorsionar las medidas estadísticas, como la media y la desviación estándar, lo que puede llevar a conclusiones erróneas. Por ejemplo, en un análisis de ingresos, un único ingreso extremadamente alto puede elevar la media, dando la impresión de que el nivel de ingreso de la población es más alto de lo que realmente es. Por lo tanto, la detección y el manejo adecuado de los outliers son esenciales para proporcionar un análisis preciso y representativo (Iglewicz & Hoaglin, 1993).

Métodos de Detección

Existen varios métodos para detectar outliers, que se pueden agrupar en enfoques estadísticos y de aprendizaje automático.

- **Métodos Estadísticos:** Un método común es el uso de la desviación estándar. Un valor se considera un outlier si se encuentra a más de tres desviaciones estándar de la media. Este método es efectivo para conjuntos de datos que siguen una distribución normal, pero puede no ser adecuado para distribuciones asimétricas (Grubbs, 1969).

- **Métodos de Aprendizaje Automático:** Los algoritmos de clustering, como DBSCAN, son útiles para identificar outliers al agrupar datos y detectar puntos que no pertenecen a ningún grupo. Esta técnica es particularmente valiosa en conjuntos de datos de alta dimensión donde la visualización es complicada (Ester et al., 1996).

Boxplot

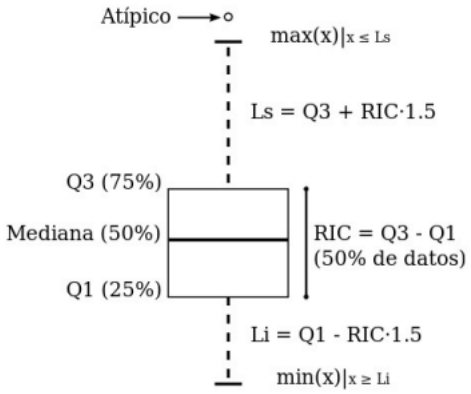


Figura 2. Características del Boxplot. Elaboración Propia

El boxplot, también conocido como diagrama de caja, es una herramienta gráfica fundamental en la estadística que se utiliza para visualizar la distribución de un conjunto de datos. Su diseño compacto permite representar de manera clara y concisa la mediana, los cuartiles y los valores atípicos, facilitando así la comparación de diferentes conjuntos de datos. Esta visualización es especialmente valiosa en el análisis exploratorio de datos, donde se busca entender la variabilidad y las características de los datos de manera rápida y efectiva.

En el siguiente enlace puedes encontrar la descripción de un boxplot: https://www.youtube.com/watch?v=Bh6EZnbfXzA

Componentes del Boxplot

Un boxplot se compone de varios elementos clave:

- **Caja:** Representa el rango intercuartílico (IQR), que es la distancia entre el primer cuartil (Q1) y el tercer cuartil (Q3). Esta caja contiene el 50% central de los datos y su longitud indica la variabilidad de la mitad central del conjunto de datos.
- **Línea Mediana:** Dentro de la caja, se traza una línea que indica la mediana del conjunto de datos. Este valor central proporciona una medida robusta de tendencia central que no se ve afectada por outliers.

- **"Bigotes":** Las líneas que se extienden desde la caja, conocidas como "bigotes", muestran el rango de los datos. Generalmente, se extienden hasta 1.5 veces el IQR desde los cuartiles, lo que ayuda a identificar valores atípicos.
- **Valores Atípicos:** Los puntos que caen fuera del rango de los bigotes se consideran outliers y se representan como puntos individuales. Esto permite a los analistas identificar rápidamente datos que pueden requerir atención especial.

Ejemplo de Uso del Boxplot

El uso de boxplots es común en diversas disciplinas, desde la biología hasta la economía. Por ejemplo, en un estudio sobre el rendimiento académico de estudiantes en diferentes grupos de edad, un boxplot puede ayudar a comparar la mediana y la variabilidad de las calificaciones entre los distintos grupos. Esto permite identificar no solo qué grupo tiene el rendimiento más alto, sino también cuál presenta más variabilidad en sus calificaciones (Tukey, 1977).

Ventajas del Boxplot

Una de las principales ventajas del boxplot es su capacidad para resumir un gran volumen de datos en una representación gráfica simple. A diferencia de otros gráficos, como los histogramas, los boxplots no solo muestran la forma de la distribución, sino que también destacan valores atípicos y la variabilidad dentro de los datos. Además, su diseño permite realizar comparaciones rápidas entre múltiples conjuntos de datos, lo que es especialmente útil en investigaciones científicas y estudios de mercado (McGill et al., 1978).

Limitaciones del Boxplot

A pesar de sus ventajas, los boxplots tienen algunas limitaciones. No muestran la distribución completa de los datos, como la forma exacta de la distribución o la presencia de agrupamientos. Además, pueden ser menos informativos en conjuntos de datos muy pequeños o en distribuciones con sesgos extremos, donde otras visualizaciones, como los histogramas, pueden ser más adecuadas (Cleveland, 1994).

Análisis de Componentes Principales

El Análisis de Componentes Principales (PCA, por sus siglas en inglés) es una técnica estadística multivariada ampliamente utilizada en el campo de la exploración y el análisis de datos. Su principal objetivo es reducir la dimensionalidad de un conjunto de datos mientras se conserva la mayor cantidad de variabilidad posible. Esta técnica es especialmente útil en contextos donde los conjuntos de datos son de alta dimensión, lo que puede dificultar la visualización y el análisis efectivo.

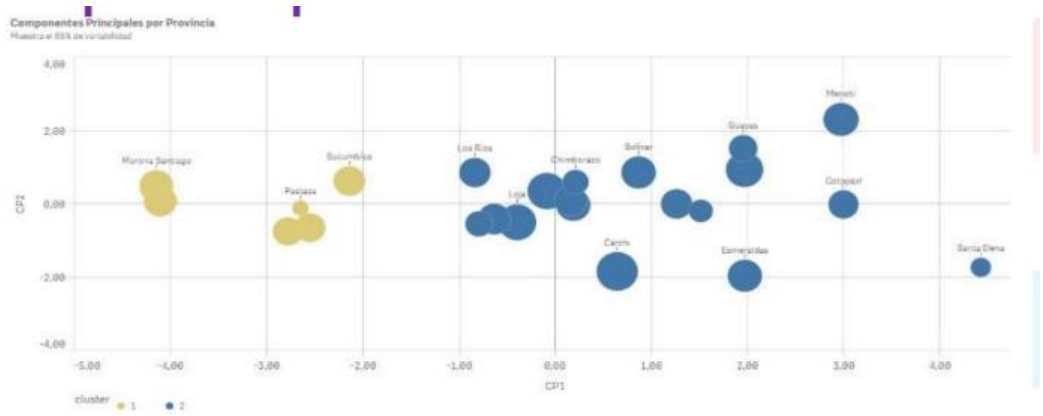


Figura 3. Análisis de Componentes Principales Datos de Salud y Nutrición ENSANUT. Elaboración Propia

Objetivos del PCA

El PCA busca transformar un conjunto de variables correlacionadas en un nuevo conjunto de variables no correlacionadas, llamadas componentes principales. Los primeros componentes principales retienen la mayor parte de la variabilidad del conjunto de datos original, lo que permite una representación más compacta y manejable. Esta reducción dimensional facilita la visualización de datos y la identificación de patrones subyacentes (Jolliffe, 2002).

Proceso del PCA

El proceso del PCA implica varios pasos clave:

- **Normalización de los Datos:** Dado que el PCA es sensible a las escalas de las variables, es crucial normalizar los datos. Esto se suele hacer mediante la estandarización, que implica restar la media y dividir por la desviación estándar de cada variable.
- **Cálculo de la Matriz de Covarianza:** Se calcula la matriz de covarianza para identificar cómo varían conjuntamente las variables del conjunto de datos. Esta matriz permite entender la relación entre las variables.
- **Cálculo de los Valores y Vectores Propios:** Se obtienen los valores y vectores propios de la matriz de covarianza. Los valores propios indican la cantidad de varianza que se captura por cada componente principal, mientras que los vectores propios representan la dirección de esos componentes.
- **Selección de Componentes Principales:** Se seleccionan los componentes principales más significativos, generalmente aquellos con los valores propios más

altos, que son los que retienen la mayor parte de la variabilidad del conjunto de datos.

- **Transformación de los Datos:** Finalmente, se proyectan los datos originales en el espacio de los componentes principales seleccionados, lo que resulta en un nuevo conjunto de datos con menos dimensiones.

Ejemplos de Aplicación

El PCA se utiliza en diversas disciplinas. En biología, por ejemplo, se emplea para analizar datos genómicos y para identificar patrones de expresión génica. En marketing, el PCA puede ayudar a segmentar clientes al identificar factores subyacentes que influyen en las preferencias de compra (Kahn et al., 2005). En ambos casos, el PCA permite a los investigadores y analistas extraer información valiosa de grandes volúmenes de datos complejos.

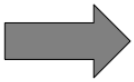
Ventajas y Limitaciones

Entre las ventajas del PCA se encuentran la simplificación del análisis de datos, la reducción del ruido y la mejora en la visualización de datos de alta dimensión. Sin embargo, el PCA también tiene limitaciones. Por ejemplo, puede ser difícil interpretar los componentes principales, ya que son combinaciones lineales de las variables originales. Además, el PCA asume que las relaciones entre las variables son lineales, lo que puede no ser cierto en todos los casos (Iglewicz & Hoaglin, 1993).

Las componentes principales son variables que no están relacionadas entre sí y que explican el comportamiento de los datos acumulando cierto porcentaje de variabilidad. Adicionalmente, los loadings o cargas de las componentes principales ayudan a la interpretación de las mismas.

Si la componente principal tiene todas sus cargas positivas, esto indica que es una componente de tamaño; es decir, mientras más alta sea esta componente, más altos serán los valores de las variables. Por otro lado, si las cargas son positivas y negativas, esto convierte a la componente en una componente de contraste; es decir, mientras más alta sea esta componente, mayores serán los contrastes entre las variables con cargas positivas y negativas. El siguiente cuadro muestra un ejemplo:

id	chol	presa	ldlc	presb	insu	gluc	hdlc	trig
1	100	120	100	48	20	142	50	150
2	220	125	103	45	23	140	51	154
3	130	112	103	42	25	139	55	142
4	150	122	108	40	21	138	52	148



	CP1	CP2	CP3
chol	0,42	-0,11	0,15
presa	0,4	0,12	0,08
ldlc	0,4	-0,16	0,04
presb	0,36	0,07	0,35
insu	0,35	0,03	-0,46
gluc	0,32	0,12	-0,67
hdlc	0,29	-0,64	0,22
trig	0,23	0,7	0,32

Figura 4. Variables de salud y nutrición y tabla con *loadings* de las variables aplicación componentes principales. Elaboración Propia

Análisis de Correspondencias

El análisis de correspondencias es una técnica estadística utilizada principalmente para explorar y visualizar relaciones entre dos o más variables categóricas. Esta metodología permite representar gráficamente las asociaciones entre categorías, facilitando la identificación de patrones y estructuras subyacentes en los datos. Es particularmente útil en investigaciones en ciencias sociales, marketing, biología y en cualquier campo donde las relaciones entre variables cualitativas sean de interés.

Objetivos del Análisis de Correspondencias

El principal objetivo del análisis de correspondencias es proporcionar una representación gráfica que permita observar la relación entre filas y columnas de una tabla de contingencia. Esto se logra mediante la construcción de un espacio de representación en dos dimensiones, donde cada punto representa una categoría de las variables analizadas. Las distancias entre los puntos en este espacio reflejan las similitudes o diferencias entre las categorías. Así, el análisis de correspondencias ofrece una forma intuitiva de interpretar datos complejos (Greenacre, 1984).

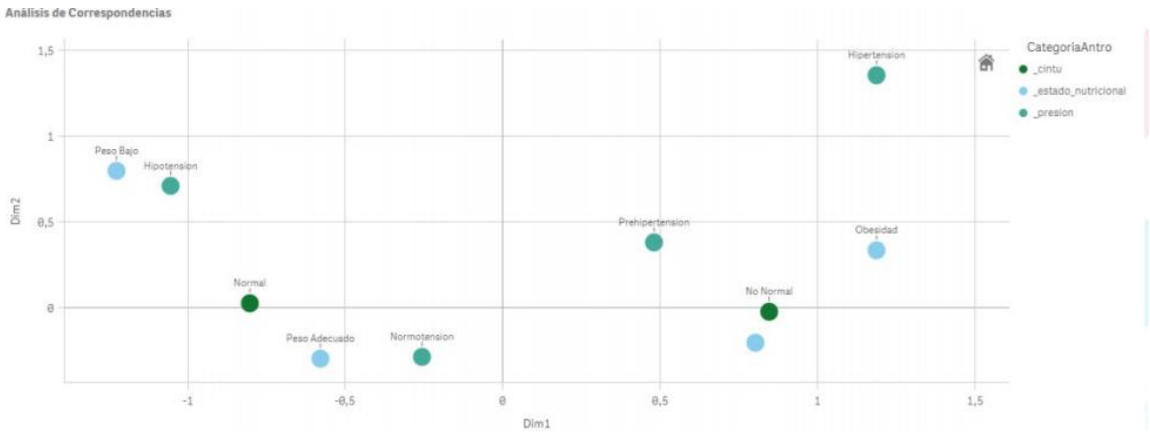


Figura 5. Análisis de Correspondencias Datos de Salud y Nutrición ENSANUT. Elaboración Propia

Fuente: Análisis de Componentes Principales en Python https://www.youtube.com/watch?v=FIc14qgsFJw

Proceso del Análisis de Correspondencias

El proceso de análisis de correspondencias incluye varios pasos fundamentales:

1. **Construcción de la Tabla de Contingencia:** Se inicia con la creación de una tabla que resume las frecuencias de las combinaciones de las categorías de las variables en estudio.

2. **Cálculo de Proporciones y Distancias:** A partir de la tabla de contingencia, se calculan las proporciones y se determinan las distancias entre las categorías utilizando una medida adecuada, como la distancia euclidiana.

3. **Descomposición en Valores Singulares (SVD):** Se aplica la descomposición en valores singulares a la matriz de correspondencias para reducir la dimensionalidad y facilitar la visualización.

4. **Visualización:** Los resultados se representan gráficamente en un plano cartesiano, donde las coordenadas de cada categoría se derivan de los valores singulares. Esto permite observar claramente las relaciones entre las diferentes categorías.

Ejemplo de Aplicación

El análisis de correspondencias es ampliamente utilizado en encuestas de opinión. Por ejemplo, en un estudio de mercado, una empresa puede querer entender la relación entre diferentes características de un producto y la satisfacción del cliente. Al aplicar esta técnica, la empresa puede identificar cómo las diversas características del producto se asocian con

niveles específicos de satisfacción, permitiendo así un mejor enfoque en el desarrollo de productos futuros (Bourguignon & Pichot, 2005).

Ventajas y Limitaciones

Entre las ventajas del análisis de correspondencias se encuentran su capacidad para simplificar y resumir grandes conjuntos de datos categóricos, así como su fácil interpretación visual. Sin embargo, es importante tener en cuenta que esta técnica también tiene limitaciones. Por ejemplo, el análisis de correspondencias puede ser sensible al tamaño de la muestra y puede no ser tan efectivo cuando las frecuencias en la tabla de contingencia son muy bajas. Además, es fundamental que las categorías sean mutuamente excluyentes y exhaustivas para obtener resultados significativos (Greenacre & Hastie, 1987).

Referencias

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(4), 433-459. https://doi.org/10.1002/wics.101

Bourguignon, A., & Pichot, J. (2005). Analyse des correspondances et analyse factorielle. *Revue de Statistique Appliquée, 53*(2), 57-74.

Cleveland, W. S. (1994). *The elements of graphing data*. Wadsworth.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge University Press.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. Academic Press.

Greenacre, M. J., & Hastie, T. (1987). The geometry of duality in correspondence analysis. *Journal of the American Statistical Association, 82*(398), 277-283. https://doi.org/10.2307/2289381

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques*. Elsevier.

Iglewicz, B., & Hoaglin, D. C. (1993). *How to detect and handle outliers*. SAGE Publications.

Jolliffe, I. T. (2002). *Principal component analysis*. Springer.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Kahn, K. B., Mentzer, J. T., & Eroglu, S. (2005). The effects of the buyer-supplier relationship on a firm's product innovation success. *Journal of Business Research, 58*(7), 916-927. https://doi.org/10.1016/j.jbusres.2003.08.006

Khan, M. A., Shah, A. A., & Khan, I. (2019). Data preprocessing techniques in data mining: A comprehensive review. *International Journal of Computer Applications, 975*, 8887. https://doi.org/10.5120/ijca2019919674

McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of box plots. *The American Statistician, 32*(1), 12-16. https://doi.org/10.2307/2683468

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM, 45*(4), 211-218. https://doi.org/10.1145/505248.505275

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin, 23*(4), 3-13.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

VanderPlas, J. (2016). *Python data science handbook*. O'Reilly Media.

Weka. (n.d.). Retrieved from https://www.cs.waikato.ac.nz/ml/weka/

PUCE Virtual
EDUCACIÓN FLEXIBLE

La excelencia no se improvisa

síguenos

MDASC - 03 - 251