Universidad Nacional del Nordeste

Facultad de Ciencias Exactas y Naturales y Agrimensura

Licenciatura en Sistemas de Información

Cátedra: Base de Datos II

Grupo 4

Integrantes:

Chiari, Guillermo Daniel	DNI:38.236.129
Perez, Gonzalo Fabian	DNI: 41.612.228
Ramirez, Enzo Agustín	DNI: 42.602.916
Salguero, Benjamín Jesús	DNI:41.412.172
Vallejos, Gonzalo Emanuel	DNI: 43.268.163
Vargas, Carlos Esteban	DNI: 42.743.291

Monografía

Tema: Fase de Limpieza y Transformación de datos

Año: 2022

Índice

Índice de Figuras

Introducción

En la siguiente monografía se desarrollarán los temas correspondientes a la Fase de Limpieza y Transformación de datos, los cuales son de gran importancia a la hora del mantenimiento, el orden y el buen manejo de las bases de datos.

La limpieza es el acto de descubrimiento y corrección o eliminación de registros de datos erróneos de una tabla, base de datos o Dataframes. El proceso de limpieza de datos permite identificar datos incompletos, incorrectos, inexactos, no pertinentes, etc. y luego sustituir, modificar o eliminar estos datos sucios.

Por otro lado, la transformación de datos es el proceso de convertir datos o información de un formato a otro, usualmente desde el formato de un sistema fuente hasta el formato requerido de un nuevo sistema de destino.

A continuación, se expondrán con mayor profundidad y detalle los temas propuestos anteriormente.

Limpieza y transformación de datos en el proceso de extracción del conocimiento

El proceso de extracción del conocimiento contempla la fase de limpieza de datos (data cleaning). La información puede contener entre valores atípicos, valores faltantes y valores erróneos. Durante esta fase, se analiza la influencia de estos valores y qué se puede hacer con estos. La presencia de datos atípicos o faltantes pueden llevarnos a usar algoritmos robustos para evitarlos, en vez de filtrar la información, y reemplazar valores mediante técnicas de imputación y transformar los datos continuos en discretos mediante técnicas de discretización.

Valores Atípicos (Outliers)

Un valor atípico u outlier es una puntuación extrema dentro de una variable. Este tipo de valores afectan fuertemente a los análisis en que intervenga la variable citada, aumentando a medida que se trabaja con muestras pequeñas.
Se los puede definir más concretamente como observaciones aisladas cuyo comportamiento se diferencia claramente del comportamiento medio del resto de las observaciones.

Existe una primera categoría de casos atípicos formada por aquellas observaciones que provienen de un error de procedimiento, como errores de codificación o entradas de datos. Si dichos datos no son detectados durante el filtrado, deben ser eliminados o recodificados como datos faltantes.

Otra categoría adicional está compuesta por las observaciones extraordinarias para la cual el investigador no tiene explicación. Normalmente estos datos se eliminan del análisis

Una última categoría de datos atípicos formada por las observaciones que se sitúan fuera del rango ordinario de valores de la variable, los cuales son denominados valores extremos y se eliminan del análisis si no son elementos significativos para la población.

Dependiendo de las características de los datos atípicos, tanto como el objetivo del análisis que se realiza, van a determinar cuáles van a ser los casos a eliminar. Aunque puede ocurrir que el investigador decida no eliminar un valor extremo al considerar un número suficiente de otras variables en el análisis.

Herramientas de análisis exploratorio de datos en contexto invariante

Gráfico de caja y bigotes

El cual consiste en un rectángulo dividido por un segmento que posiciona la mediana de la variable, y su relación con los cuartiles. En esta caja, se posicionan segmentos llamados bigotes que tienen como extremos los valores mínimo y máximo de la variable.
Los valores atípicos, entonces, se presentan como puntos aislados fuera del rango marcado por los bigotes y los extremos tachados por una x.
En la figura se muestra un gráfico de caja y bigotes para una variable V1.

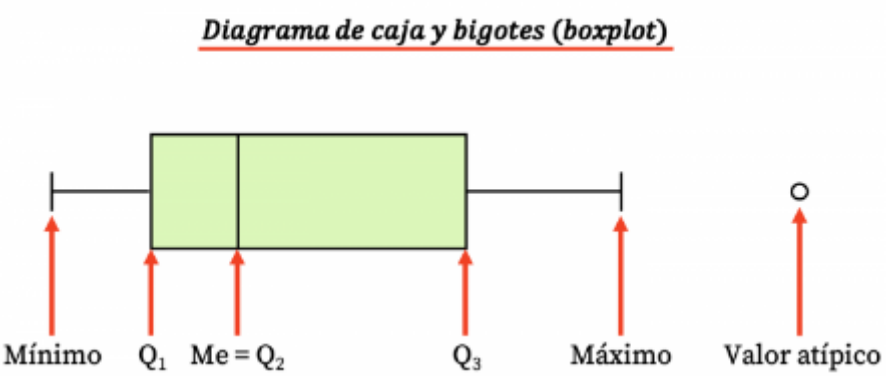

Figura 1. Diagrama de caja y bigotes (boxplot).

Diagrama de control

Consiste en la representación gráfica con una línea central que denota el valor medio de la variable y otras dos líneas horizontales, llamadas Límite superior del control(LSC) y Límite inferior del control (LIC). Estos límites son escogidos de manera que casi la totalidad de los puntos tomados por la variable se hallen dentro de ellos.

Los diferentes puntos son unidos por segmentos rectilíneos para visualizar mejor la secuencia de valores, y aquellos que se encuentren fuera de los límites establecidos se considerarán valores atípicos y se realizarán las acciones de investigación y corrección pertinentes.

Figura 2. Gráfico de control: VAR00001

Estadísticos robustos de la variable

Se pueden detectar valores atípicos mediante los estadísticos robustos de la variable y ver su diferencia con respecto de los estadísticos no robustos. Se suele considerar a los estadísticos robustos de centralización a la mediana, la media truncada y la media winsorizada. La media truncada prescinde del 15% de los valores de la variable por cada extremo y la media winsorizada sustituye ese 15 % por valores del centro de la distribución. Cuando no hay valores atípicos los estadísticos robustos y los estadísticos normales no difieren mucho.

Herramientas de análisis exploratorio de datos en contexto bivariante

Gráfico de cajas y bigotes

El cual puede utilizarse también para representar distintos gráficos de una variable para diferentes niveles de una segunda.

La figura siguiente representa una gráfica de cajas y bigotes donde las dos variables consisten de la potencia de los automóviles y los países donde se realiza el estudio.

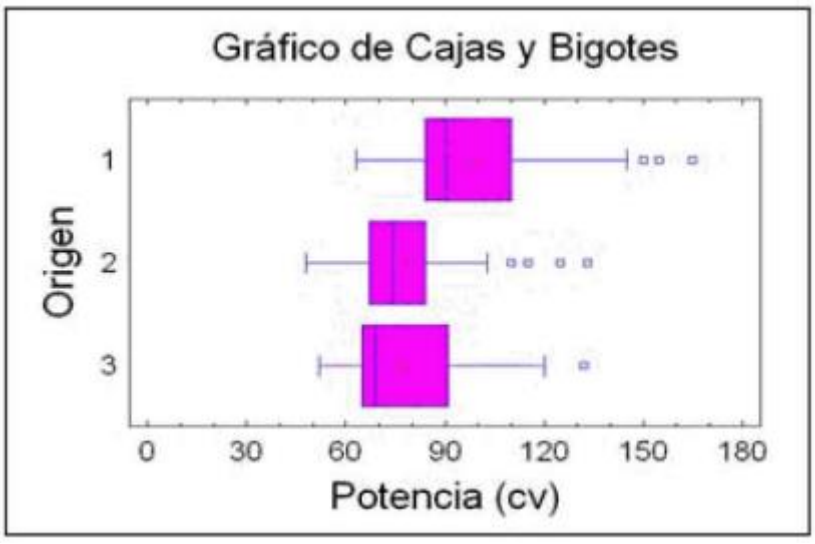

Figura 3. Gráfico de cajas y Bigotes.

Gráfico de dispersión

Los gráficos de dispersión se usan para averiguar la intensidad de la relación entre dos variables numéricas. El eje X representa la variable independiente, mientras que el eje Y representa la variable dependiente.

En la siguiente figura, se representa el consumo de combustible de los coches en función de su potencia, y aquellos puntos fuera del rango del resto de observaciones son identificados como valores atípicos.

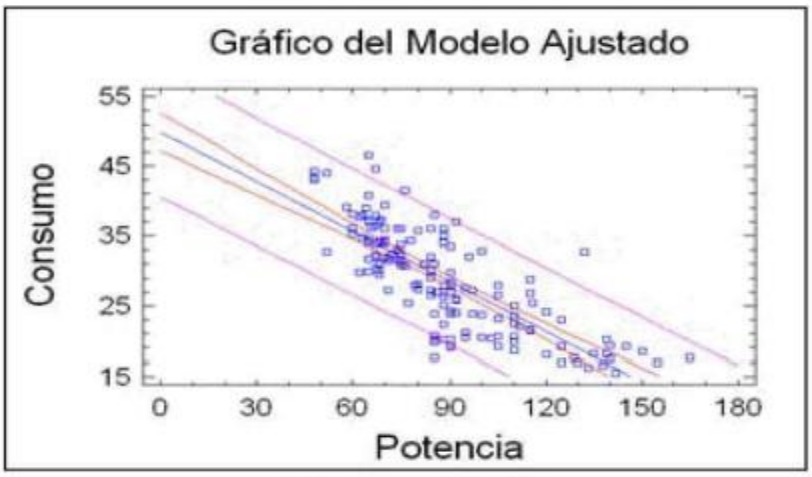

Figura 4. Gráfica del modelo ajustado.

Para detectar casos atípicos en un contexto multivariante, se pueden utilizar estadísticos basados en distancias, para detectar puntos influyentes. Tales como la distancia D2 de Mahalanobis, la cual es una medida de distancia de cada observación en un espacio multidimensional respecto al centro medio de las observaciones.

Información Faltante (Datos MISSING)

El tratamiento de la información faltante es una tarea previa a cualquier análisis. Estos valores los llamaremos valores ausentes o valores missing. La presencia de esta información faltante puede deberse a distintos motivos tales como pueden ser un registro defectuoso de la información, la ausencia natural de la información buscada o también a la falta de respuesta tanto total como parcial.

A la hora de realizar pruebas cuando existan datos missing debe de comprobarse si estos están distribuidos aleatoriamente sobre todo el conjunto de datos. Se considera primero si los datos ausentes para una única variable Y y separar los datos agrupando quienes tienen datos ausentes y cuáles no, luego se realizan test para determinar si existen diferencias significativas entre los grupos de valores determinados por la variable Y. Si se va considerando Y como cada una de las variables del análisis y se repite el test hasta encontrar todas las diferencias son no significativas, se puede concluir que los datos missing obedecen un proceso completamente aleatorio, por ende, se pueden realizar análisis estadísticos fiables. Si un porcentaje alto de las diferencias son no significativas, se puede concluir que es un proceso aleatorio, pero no completamente aleatorio, a lo que nuestros análisis estadísticos tendrán menos fiabilidad debido a la imputación de la información faltante.

Es habitual comprobar la distribución aleatoria de los datos missing mediante pruebas como la de correlaciones dicotomizadas o el test conjunto de aleatoriedad de Little.

Soluciones para los datos ausentes: Supresión de Datos o imputación de la información faltante.

Una vez que se ha contrastado la existencia de aleatoriedad en los datos ausentes ya se puede tomar una decisión para dichos datos antes de comenzar cualquier análisis estadístico con ellos. Podemos comenzar incluyendo solo en el análisis las observaciones (casos) con datos complejos (filas cuyos valores para todas las variables sean válidos), es decir cualquier fila que tenga algún dato desaparecido se elimina del conjunto de datos antes de realizar el análisis. Este método se denomina aproximación de casos complejos o supresión de casos según lista y suele ser el método por defecto en la mayoría del software estadístico. Este método es apropiado cuando no hay demasiados valores perdidos, porque su supresión provocaría una muestra representativa de la información total. En caso contrario se reduciría mucho el tamaño de la muestra a considerar para el análisis y no sería representativa de la información completa. [1]

Supresión de datos según pareja: Se trabaja con todos los casos (filas) posibles que tengan valores válidos para cada par de variables que se consideren en el análisis independiente de lo que ocurra para cada par de variables que se consideren en el análisis independiente de lo que se ocurra en el resto de las variables. Este método elimina menos información y se utiliza siempre en cualquier análisis bivariante o transformable en bivariante.

Suprimir los casos (filas) o variables(columnas): Nuevamente es necesario sopesar la cantidad de datos a eliminar. Debe siempre considerarse lo que se gana al eliminar una fuente de datos ausentes y lo que se pierde al no contar con una determinada variable o conjunto de casos en el análisis estadístico

Imputación de la información faltante: La imputación es el proceso de estimación de valores ausentes en valores válidos de otras variables o casos de muestra. Un primer método de imputación no reemplaza los datos ausentes, sino que imputa las características de la distribución (por ejemplo, la desviación típica) o las relaciones de todos los valores válidos posibles, como representantes para toda la muestra entera. Este método se denomina enfoque de disponibilidad completa.

Método de imputación por sustitución del caso: Las observaciones (casos) con datos ausentes se sustituyen con otras observaciones no maestras. Por ejemplo, en una encuesta sobre hogares a veces se sustituye un hogar de la muestra que no contesta por otro hogar que no está en la muestra y que probablemente lo contestará.

Método de imputación de sustitución por la media: Los datos ausentes se sustituyen por la media de todos los valores válidos de su variable correspondiente. Este método tiene la ventaja de que se implementa fácilmente y proporciona la información completa para todos los casos, pero tiene la desventaja de que modifica las correlaciones e invalida las estimaciones de la varianza derivadas de las fórmulas estándar de la varianza para conocer la verdadera varianza de datos.

Método de imputación de sustitución por la mediana: Cuando hay valores extremos en las variables, se sustituyen los valores ausentes por la mediana (en vez de por la media), ya que la mediana es un resumen estadístico de los datos más robustos. De esta forma se tiene el método de imputación de sustitución por la mediana.

Método de imputación por interpolación: A veces, cuando hay demasiada variabilidad de datos, suele sustituirse cada valor ausente por la media o mediana de un cierto número de observaciones adyacentes a él. En la cual se sustituye cada valor ausente de una variable por el valor resultante de realizar una interpolación con los valores adyacentes.

Método de imputación de sustitución por valor constante: Los datos ausentes se sustituyen por un valor constante apropiado derivado de fuentes externas o de una investigación previa. En este caso el investigador debe asegurarse de que la sustitución de los valores ausentes por el valor constante proveniente de una fuente externa es más válida que la sustitución de la media (valor generado internamente).

Método de imputación por regresión: Se utiliza el análisis de la regresión para predecir los valores ausentes de una variable basándose en su relación con otras variables del conjunto de datos a partir de la ecuación de regresión que las liga.

Método de imputación múltiple: Es una combinación de varios métodos entre los ya citados.

Transformación de datos

Cuando el análisis exploratorio lo indique, los datos originales pueden necesitar ser reemplazados. Suelen considerarse cuatros tipos de transformaciones:

- **Transformaciones Lógicas**: Se unen características del campo de definición de variable para reducir así su amplitud. De esa forma pueden eliminarse categorías sin respuestas. También pueden convertirse variables de intervalo en ordinales o nominales y crear variables factibles.
- **Transformaciones Lineales**: Se obtiene de sumar, restar, multiplicar o dividir las observaciones originales por una constante para mejorar su interpretación. Estas transformaciones no cambian la forma de la distribución ni la distancia entre los valores ni el orden, y por tanto no provocan cambios considerables entre las variables.
- **Transformaciones Algebraicas**: Se obtiene de aplicar transformaciones no lineales monotónicas a las observaciones originales por una constante para mejorar su interpretación. Estas transformaciones cambian la forma de distribución al cambiar las distancias entre los valores, pero mantienen el orden.
- **Transformaciones no Lineales no Monotónicas**: Cambian las distancias y el orden entre los valores. Pueden cambiar demasiado la información original.

Con estas transformaciones se arreglan problemas en los datos. Por ejemplo: Una asimetría negativa puede minorarse con una transformación parabólica o cubica, una asimetría positiva fuerte puede suavizarse mediante una transformación hiperbólica o hiperbólica cuadrática y una asimetría positiva débil puede suavizarse mediante una transformación de raíz cuadrada, logarítmica o recíproca de la raíz cuadrada. La transformación logarítmica puede conseguir estacionalidad en medio y en varianza para los datos. [1]

Suele elegirse como transformaciones aquella que arregla mejor el problema, una vez realizada. Si ninguna arregla el problema, realizamos el análisis sobre los datos originales sin transformaciones. Combinando transformaciones lineales y algebraicas pueden modificarse los valores extremos de la distribución.

Transponer, fusionar, agregar, segmentar y ordenar archivos

La organización de los archivos de datos no siempre se encuentra de la manera en que lo necesitamos, por ello en ocasiones podremos realizar cambios en el orden de los casos,

transponer tantos las filas como las columnas y también por otro lado barajar varios archivos distintos o tomar solo una muestra de casos, que sean necesario para la ocasión.

Transponer crear un archivo de datos nuevo en el que se transponen las filas y las columnas del archivo de datos original de manera que los casos (las filas) se convierten en variables, y las variables (las columnas) se convierten en casos. Normalmente, si el archivo de datos de trabajo contiene una variable de identificación o de nombre con valores únicos, podrá utilizarla como variable de nombre sus valores se emplearán como nombre de variable en el archivo de datos transpuesto.

La fusión de archivo consiste en la formación de un nuevo archivo con las mismas variables y casos diferentes. Se trata de Añadir casos fusionando el archivo de datos de trabajos con otro archivo de datos que contiene las mismas variables, pero diferentes casos. También es posible fundir archivos con los mismos casos, pero variables diferentes. En este caso es necesario que existan variables claves tanto en el archivo de trabajo como en el archivo externo que se funde con él. Ambos archivos deben estar ordenados según el orden ascendente de las variables clave. [1]

Agregar datos combina grupos de casos de resumen únicos y crea un nuevo archivo de datos agregado. Los casos se agregan en función del valor de una o más variables de agrupación. El nuevo archivo de datos contiene un caso para cada grupo. Por ejemplo, se puede agregar datos de regiones por estado y crear un nuevo archivo en el que el estado sea la unidad de análisis.

Segmentar un archivo es dividir el archivo de datos en distintos grupos para el análisis basándose en los valores de una o más variables de agrupación. Si se seleccionan varias variables de agrupación, los casos se agrupan por variable dentro de las categorías de la variable anterior de la lista.

Ponderar casos y categorizar y numerizar variables

Ponderar casos proporciona a los casos diferentes ponderaciones (mediante una réplica simulada) para el análisis estadístico. Los valores de la variable de ponderación deben indicar el número de observaciones representadas por casos únicos en el archivo de datos. Los casos con valores perdidos, negativos o cero para la variable de ponderación se excluyeron del análisis. Los valores fraccionarios son válidos y se usan exactamente donde adquieren sentido y, con mayor probabilidad, donde se tabulan los casos.

Categorizar Variables consiste en crear una variable categórica a partir de una variable de escala, es decir, se trata de convertir datos numéricos continuos en un número discreto de categorías. Este procedimiento crea nuevas variables que contienen los datos categóricos.

También es posible crear una variable numérica a partir de una categórica asignando valores numéricos a las categorías.

Pareamiento o matching

Las técnicas de pareamiento o matching persiguen la comparabilidad de grupos utilizando características comunes de todos ellos. Aunque los grupos difieran respecto a algunas de sus variables, es posible compararlos mediante un procedimiento de ajuste o estandarización. Este procedimiento consiste en igualar ambos grupos con relación a algunas características, haciéndola homogénea en ambos grupos. Un efecto importante de matching es el aumento en la eficiencia del estudio, ya que permite circunscribir la población a estudiar a aquella en la cual la exposición es más representativa. [1]

Conceptualmente el matching corresponde a un procedimiento empleado a priori, en la fase de diseño del estudio. Ocasionalmente se puede efectuar pareamiento a posteriori, cuando el investigador decide parear observaciones una vez recogido los datos, a partir de un conjunto de individuos controles que previamente no fueron sometidos a matching. Sin embargo, se prefiere reservar el término matching para aquellos casos en que el procedimiento se emplea a priori.

El matching se usa también cuando se trabaja con variables confusas de difícil definición o medición, como, por ejemplo, las de tipo genérico, psicosocial o relacionadas a comportamientos humanos. En estos casos, los investigadores suelen utilizar “pares” de sujetos, con la finalidad de poder estudiar aisladamente el efecto de la variable de interés habiendo controlado la influencia de las variables sometidas a pareamiento, las que se asumen comunes. Los tipos de variables sometidas a matching pueden ser variados y dependerán lógicamente del problema a investigar.

Existen varias modalidades de pareamiento o matching. Dos de las más utilizadas, dependiendo si este procedimiento se aplica colectivamente o a observaciones específicas, son el matching de grupos o de frecuencia y el matching individual (Figura 5).

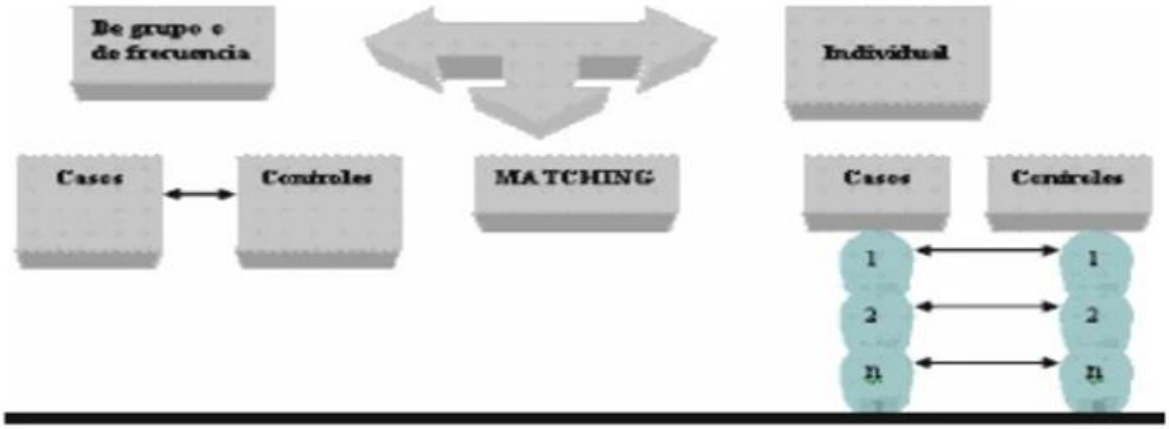

Figura 5. matching de grupos o de frecuencia y matching individual.

En la modalidad de matching de grupo o de frecuencia se restringe a priori el ingreso de sujetos en ambos grupos buscando estudiar a sujetos que representen adecuadamente los criterios de inclusión. Así, el ingreso al estudio puede estar regulado por características tales como sexo, edad, ocupación, lugar de resistencia o modalidad de cuidados médicos. La contribución de los grupos en cuanto a eventuales factores confusos tiende a ser homogénea en casos y controles, lo que incrementa la potencia del estudio.

En el matching individual, las características a parear se definen específicamente para cada caso y cada control simultáneamente. Se podrá apreciar que el efecto de este procedimiento tiene implicancias directas en la modalidad de análisis de la información: en este caso el análisis se efectúa por “parte” o “tríos” de observaciones, a diferencia de la modalidad de matching por grupos o de frecuencia, en la que se comparan grupos. También tiene implicaciones en la factibilidad de encontrar adecuados sujetos controles que ajusten a los requerimientos exigidos en el matching. A mayor cantidad de variable a “parear”, mayor dificultad de encontrar controles adecuados. En ambos casos, el matching puede considerar más de un control por cada caso.

El matching o pareamiento también presenta desventajas. Este procedimiento involucra dificultades técnicas y teorías en el desarrollo del estudio. El investigador se expone a encontrar dificultades para encontrar controles adecuados y en muchos casos debe descartar controles con el consiguiente riesgo de sesgar las mediciones en el caso de que la(s) variable(s) a parear no sean de valor epidemiológico. El investigador puede verse enfrentado a la realidad de encontrar en su base de una alta frecuencia de valores missing, debiendo descartar dichas observaciones o aplicar procedimientos de estimación de ellos usando procedimientos de poca aceptación epidemiológica. El estudio se hace también más largo y, por ende, de mayor costo. [1]

Transformación de datos mediante técnicas de reducción de la dimensión

En el mundo de la información de hoy, es habitual disponer de gran cantidad de variables medidas u observadas en una colección de individuos y pretender estudiarlas conjuntamente. Al observar muchas variables sobre una muestra es presumible que una parte de la información recogida pueda ser redundante o excesiva, en cuyo caso los métodos multivariantes de reducción de la dimensión tratan de eliminarla. Estos métodos combinan muchas variables observadas para obtener pocas variables ficticias que las representen con la mínima pérdida de información.

Estos métodos de reducción de la dimensión son métodos multivalentes de la interdependencia en el sentido de que todas sus variantes tienen una importancia equivalente.

Los métodos de la interdependencia se contraponen con métodos multivariantes de la dependencia en donde no es aceptable una importancia equivalente en variables, porque alguna se destaca como principal. En estos casos, la utilización de técnicas multivariantes analíticas o inferenciales considerando la variable dependiente como explicada por las demás variables y tratando de relacionar todas las variables por una posible ecuación o modelo que las ligue. El método elegido podría ser entonces la Regresión lineal, generalmente con las variables cuantitativas. Una vez configurado el modelo matemático se llegará a predecir el valor de la variable dependiente conociendo el perfil de todas las demás. Si fuera una variable cualitativa dicotómica podría usarse como clasificadora, estudiando la relación con el resto de variables similares mediante Regresión logística. Si la variable constatara la asignación de individuos a grupos definidos (dos o más de dos), se utilizaría para clasificar nuevos casos en que se desconozca el grupo al que pertenecen, ahí estamos ante el Análisis discriminante, que resuelve el problema de asignación en función de un perfil cuantitativo de variables clasificativas. Si la dependiente es cuantitativa y las explicativas cualitativas, estamos ante modelos de análisis de varianza, que extienden a modelos loglineales para analizar tablas de contingencia de dimensión elevada. Si puede ser cualitativa o cuantitativa y las independientes cualitativas, es segmentación.

Por otro lado, las técnicas de reducción de la dimensión juegan un papel muy importante dentro de las técnicas emergentes de análisis multivariante de datos.

Componentes principales

El análisis en componentes principales es una técnica de análisis estadístico multivariante. Se trata de un método multivariante de simplificación de la dimensión, que se aplica cuando se dispone de un conjunto elevado de variables con datos cuantitativos correlacionados entre sí persiguiendo obtener un menor número de variables, que resuman lo mejor posible a las variables iniciales con la mínima pérdida de información.

Esta reducción de variables puede simplificar la aplicación de otras técnicas multivariantes. El número elevado de variantes x1, x2,…, xn, se resumen en pocas variables C1, C2,…Cn (Componentes principales) perfectamente calculables y que sintetizan la información contenida en datos. Inicialmente se tienen tantas componentes como variables:

$$C_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p$$
$$\vdots$$
$$C_p = a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{pp}x_p$$

Pero, solo retienen las k componentes principales (k ≤ p) que explican un porcentaje alto, de la variabilidad de las variables iniciales (C1, C2, … …., Cn).

Como medida de la información incorporada en una componente, se utiliza su varianza. Es decir, a mayor valor de varianza, mayor la información incorporada en el componente.

Por lo general, la extracción de componentes principales se efectúa sobre variables tipificadas, para evitar problemas derivados de escala, aunque también se puede aplicar sobre variables expresadas en desviaciones respecto de la media.

Cuando las variables originales están muy correlacionadas entre sí, la mayor parte de su variabilidad se explica con muy pocas componentes. Si las variables originales fueran completamente incorrelacionadas entre sí, el análisis de componentes principales carecería de interés.

Cálculo de las componentes principales

En el análisis en componentes principales se dispone de una muestra de tamaño *n* acerca de *p* variables X1, X2,…, Xp (tipificadas o como desviaciones respecto de la media) inicialmente correlacionadas, así luego obtenemos gracias a ellas un número k ≤ p de variables incorrelacionadas C1, C2,……, Ck que son una combinación lineal de variables iniciales y expresan la mayor parte de la variabilidad. El primer componente, como las restantes, se expresa como una combinación de las variables originales como sigue:

$$C_{1i} = u_{11}X_{1i} + u_{12}X_{2i} + \cdots + u_{1p}X_{pi} \quad i=1,..,n$$

Para el conjunto de las n observaciones muestrales y para todas las componentes tenemos:

$$\begin{bmatrix} C_{11} \\ C_{12} \\ \vdots \\ C_{1n} \end{bmatrix} = \begin{bmatrix} X_{11} & X_{21} \cdots & X_{p1} \\ X_{12} & X_{22} \cdots & X_{p2} \\ & \vdots & \\ X_{1n} & X_{2n} \cdots & X_{pn} \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{1p} \end{bmatrix}$$

En notación abreviada tenemos: $C_1 = X\, u_1\, y$:

$$V(C_1) = \frac{\sum_{i=1}^{n} C_{1i}^2}{n} = \frac{1}{n} C_1' C_1 = \frac{1}{n} u_1' X' X u_1 = u_1' \left[\frac{1}{n} X'X \right] u_1 = u_1' V u_1$$

El primer componente C1 se obtiene de forma que su varianza sea máxima sujeta a la restricción de que la suma de los pesos u_{1j} al cuadrado iguale la unidad, la variable de pesos o ponderaciones (u11, u12,……, u1p) se considera normalizada. Entonces buscamos el C_1 maximizado $V(C_1) = u_1'Vu_1$, sujeta a la restricción:

$$\sum_{j=1}^{p} u_{1i}^2 = u_1' u_1 = 1$$

Se demuestra que para maximizar V(C_1) se toma el mayor valor propio de λ de la matriz V. Sea λ1 el mayor valor de V y tomando u_1 como su vector propio normalizado ($u_1'\, u_1$= 1), tenemos definido el vector de ponderaciones que se aplica a variables iniciales, para obtener el primer componente principal, que está definido como:

$$C_1 = u_1 X = u_{11}X_1 + u_{12}X_2 + \cdots + u_{1p}X_p$$

Para maximizar V(C_1) se toma el segundo mayor valor de λ de la matriz V. Tomando λ2 como el segundo valor de V y tomando u_2 como su vector asociado normalizado, tenemos definido el vector de ponderaciones que se aplica en variables iniciales para obtener la segunda componente principal, que vendrá definida como:

$$C_2 = u_2 X = u_{21} X_1 + u_{22} X_2 + \cdots + u_{2p} X_p$$

De forma similar la componente h-ésima se define como Ch = Xuh donde uh es el vector propio de V asociado a su h-ésimo mayor valor propio. Se le llama también a uh el eje factorial h-ésimo.

Se demuestra que la proporción de variabilidad total recogida por la componente principal h-ésima (porcentaje de inercia explicada) está dada por:

$$\frac{\lambda_h}{\sum_{h=1}^{p} \lambda_h} = \frac{\lambda_h}{traza(V)}$$

Si las variables están tipificadas, *traza (V) = p,* la proporción de la componente h-ésima en la variabilidad será λh/p. También se define porcentaje inercia explicada por las "k" primeras componentes principales como:

$$\frac{\sum_{h=1}^{k} \lambda_h}{\sum_{h=1}^{p} \lambda_h} = \frac{\sum_{h=1}^{k} \lambda_h}{traza(V)}$$

Puntuaciones o medición de las componentes

El análisis en componentes principales es habitualmente un paso previo a otros análisis, por lo que es necesario conocer los valores que toman las componentes en cada observación.

Una vez calculados los coeficientes u_{hj} se pueden obtener las puntuaciones Z_{hj}, es decir, los valores de las componentes correspondientes a cada observación, a partir de la siguiente relación:

$$Z_{hi} = u_{h1} X_{1i} + u_{h2} X_{2i} + \cdots + u_{hp} X_{pi} \quad h = 1,\ldots,p \quad i = 1,\ldots,n$$

Número de componentes principales a retener

Uno de los problemas que posee este tipo de análisis es determinar k, es decir, ¿Qué número componentes se deben retener? Para responder esta pregunta existen una serie de criterios

Criterio de la media aritmética

Según este criterio se van a seleccionar aquellas componentes cuya raíz característica (varianza) exceda la media de las raíces características. Analíticamente serán aquellas componentes que cumplan:

$$\lambda_h > \overline{\lambda} = \frac{\sum_{j=1}^{p} \lambda_h}{p}$$

Criterio del gráfico de sedimentación

Este se obtiene al representar en el eje de ordenadas las raíces características y en el eje de las abscisas los números de las componentes principales correspondientes a cada raíz característica en orden decreciente, al unir los puntos se obtiene una figura similar a una montaña con una pendiente marcada hasta la base, que es una meseta con una ligera inclinación, es aquí donde se sedimentan los guijarros que caen de la montaña y se sedimentan, de ahí el nombre del gráfico. Según este criterio las componentes se van a conservar son aquellas que se encuentren en las zonas previas a la de sedimentación.

Matriz de cargas factoriales, comunalidad y círculos de correlación

La principal dificultad que existe a la hora de interpretar los componentes radica en la necesidad de que estos tengan sentido y midan algo útil, por ello es indispensable considerar el peso de cada variable original dentro del componente elegido.

Un componente es una función lineal de todas las variables, pero puede estar muy bien correlacionados con algunas de ellas y menos con otras. El coeficiente de correlación entre componente y variable se calcula de la siguiente manera:

$$r_{jh} = u_{hj}\sqrt{\lambda_h}$$

Gracias a esto cada variable puede ser representada como una función lineal de los k componentes retenidos, donde los pesos de cada componente en la variable coinciden con los coeficientes de correlación.

El cálculo matricial permite obtener inmediatamente la tabla de coeficientes de correlación variables – componentes (pxk), denominada como matriz de cargas factoriales. Se expresa de la siguiente manera:

$$\begin{aligned} Z_1 &= r_{11}X_1 + \cdots + r_{1p}X_p \\ Z_2 &= r_{21}X_1 + \cdots + r_{2p}X_p \\ &\vdots \\ Z_k &= r_{k1}X_1 + \cdots + r_{kp}X_p \end{aligned} \quad \Rightarrow \quad \begin{aligned} X_1 &= r_{11}Z_1 + \cdots + r_{k1}Z_k \\ X_2 &= r_{12}Z_1 + \cdots + r_{k2}Z_k \\ &\vdots \\ X_p &= r_{1p}Z_1 + \cdots + r_{kp}Z_k \end{aligned}$$

La suma de las comunalidades de todas las variables representa la parte inercia global de la nube original explicada por los k factores retenidos y coincide con la suma de los valores propios de estas componentes.

También demuestra que la suma en vertical de los cuadrados de las cargas factoriales de todas las variables en un componente es su propio valor.

Su representación gráfica puede orientar al investigador en una primera aproximación a la interpretación de los componentes. Este gráfico se denomina círculo de correlación y están formados por puntos que representan cada variable por medio de dos coordenadas que miden coeficientes de correlación de dicha variable con los dos factores o componentes considerados.

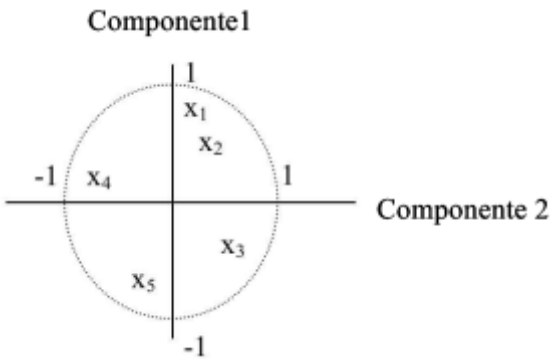

Figura 6. Círculo de correlación

Rotación de las componentes

Frecuentemente no se encuentran interpretaciones a los factores obtenidos. Sería ideal, para una interpretación más fácil, que cada componente estuviera relacionado muy bien con pocas variables y mal con las demás, esto se puede lograr mediante la rotación de los ejes.

Al hacer esto no cambia la proporción de inercia total, ni las comunalidades de cada variable, sin embargo, sí afectan los coeficientes. Las rotaciones más utilizadas son VARIMAX, la QUARTIMAX y la PROMAX.

El Análisis Factorial

El análisis factorial es una técnica de reducción de datos que sirve para encontrar grupos homogéneos de variables a partir de un conjunto numeroso de variables.

Los grupos homogéneos se forman con las variables que correlacionan mucho entre sí y procurando, inicialmente, que unos grupos sean independientes de otros.

Cuando se recogen un gran número de variables de forma simultánea (por ejemplo, en un cuestionario de satisfacción laboral) se puede estar interesado en averiguar si las preguntas del cuestionario se agrupan de alguna forma característica. Aplicando un análisis factorial a las respuestas de los sujetos se pueden encontrar grupos de variables con significado común y conseguir de este modo reducir el número de dimensiones necesarias para explicar las respuestas de los sujetos. [2]

El Análisis Factorial es, por tanto, una técnica de reducción de la dimensionalidad de los datos. Su propósito último consiste en buscar el número mínimo de dimensiones capaces de explicar el máximo de información contenida en los datos.

A diferencia de lo que ocurre en otras técnicas como el análisis de varianza o el de regresión, en el análisis factorial todas las variables del análisis cumplen el mismo papel: todas ellas son independientes en el sentido de que no existe a priori una dependencia conceptual de unas variables sobre otras.

Fundamentalmente lo que se pretende con el Análisis Factorial (Análisis de Componentes Principales o de Factores Comunes) es simplificar la información que nos da una matriz de correlaciones para hacerla más fácilmente interpretable. Se pretende encontrar una respuesta al preguntarnos ¿Por qué unas variables se relacionan más entre sí y menos con otras? Hipotéticamente es porque existen otras variables, otras dimensiones o factores que

explican por qué unos ítems se relacionan más con unos que con otros. En definitiva, se trata de un análisis de la estructura subyacente a una serie de variables.

Entre los métodos para obtener los factores se pueden mencionar:

• **Método de las componentes principales**: Método de extracción de factores utilizando para formar combinaciones lineales no correlacionadas de las variables observadas. El primer componente tiene la varianza máxima.

• **Método de Mínimos cuadrados no ponderados**: Método de extracción factorial que minimiza la suma de los cuadrados de las diferencias entre las matrices de correlaciones observada y reproducida, ignorando las diagonales.

• **Método de Mínimos cuadrados generalizados**: Minimiza el mismo criterio. La suma de las diferencias al cuadrado entre las matrices de correlación observada y reproducida ponderando las correlaciones inversamente por la varianza del factor específico. Este método permite, además, aplicar contraste de hipótesis para determinar el número de factores. [2]

• **Método de máxima verosimilitud**: Método de extracción factorial que proporciona las estimaciones de los parámetros que con mayor probabilidad han producido la matriz de correlaciones observada, si la muestra procede de una distribución normal multivariada.

• **Factorización de ejes principales**: La saturación de los factores resultantes se utilizan para estimar de nuevo las comunalidades y reemplazan a las estimaciones previas en la diagonal de la matriz. Se repite iterativamente hasta que satisfagan el criterio de convergencia para la extracción.

• **Alfa**: Maximiza el alfa de Cronbach para los factores. [2]

• **Factorización imagen**: Consiste en aplicar el método de componentes principales a la matriz de correlaciones obtenida a partir de las partes predichas de las diversas regresiones lineales de cada una de las variables sobre las demás (dicha parte recibe el nombre de imagen de la variable). [2]

Contrastes en el método factorial

El Análisis Factorial puede ser Exploratorio o Confirmatorio:

En el modelo factorial pueden realizarse varios tipos de contrastes los cuales suelen agruparse en estos dos bloques según se apliquen previamente a la extracción de los factores o que se apliquen después. Con los contrastes aplicados previamente a la extracción de los factores trata de analizar la pertinencia de la aplicación del análisis

factorial a un conjunto de variables observables. Con los contrastes aplicados después de la obtención de los factores se pretende evaluar el modelo factorial una vez estimado.

Por lo tanto, antes de realizar un análisis factorial debe cumplirse que las p variables originales están altamente inter correlacionadas, puesto que, si no lo están o, si sus correlaciones son muy bajas, no existirían factores comunes y el Análisis Factorial no tendría sentido su aplicación.

Contenido en el grupo de contrastes que se aplican previamente a la extracción de los factores se conocen el contraste de esfericidad de Bartlett y la medida de adecuación muestral de Kaiser, Meyer y Olkin.

Prueba de esfericidad de Bartlett contrasta la hipótesis de que la matriz de correlaciones es una matriz de identidad, lo que indicaría que las variables no están relacionadas y, por lo tanto, no son adecuadas para la detección de estructuras. Los valores pequeños (menores que 0.05) del nivel de significación indican que un análisis factorial puede ser útil con los datos. [3]

La **Medida Kaiser-Meyer-Olkin de adecuación de muestreo** es un estadístico que indica la proporción de varianza en las variables que pueden ser causadas por factores subyacentes. Los valores altos (cercanos a 1.0) generalmente indican que un análisis factorial puede ser útil con los datos. Si el valor es menor que 0.50, los resultados del análisis factorial probablemente no serán muy útiles. [3]

Rotación de los factores

Transforma la matriz factorial inicial en otra denominada matriz factorial rotada, más fácil de interpretar, que consiste en una combinación lineal de la primera y que explica la misma cantidad de varianza inicial. Los factores rotados tratan de que cada una de las variables originales tenga una correlación lo más próxima a uno que sea posible con uno de los factores, y correlaciones próximas a cero con los restantes, consiguiendo así correlaciones altas con un grupo de variables y baja con el resto.

Sin embargo, su interpretación, a veces, puede llegar a ser muy compleja, por lo que se puede recurrir a la rotación de los componentes (ejes).

Existen varias formas de rotar los ejes:

- Método Varimax. Método de rotación ortogonal que minimiza el número de variables que tienen cargas altas en cada factor. Simplifica la interpretación de los factores.

- Método Quartimax. Método de rotación que minimiza el número de factores necesarios para explicar cada variable. Simplifica la interpretación de las variables observadas.
- Método Equamax. Método de rotación que es combinación del método varimax, que simplifica los factores, y el método quartimax, que simplifica las variables. Se minimiza tanto el número de variables que saturan alto en un factor como el número de factores necesarios para explicar una variable. [3]

Características de los métodos de rotación oblicua más importantes:

Esta rotación se puede calcular más rápidamente que una rotación oblimin directa, por lo que es útil para conjuntos de datos grandes.

- Criterio Oblimin *directo*. Método para la rotación oblicua (no ortogonal). Si el delta es igual a cero (el valor predeterminado) las soluciones son las más obligatorias. A medida que delta se va haciendo más negativo, los factores son menos oblicuos. Para anular el valor predeterminado 0 para delta, introduzca un número menor o igual que 0,8.
- Rotación Promax. Rotación oblicua que permite que los factores estén correlacionados. Esta rotación se puede calcular más rápidamente que una rotación oblimin directa, por lo que es útil para conjuntos de datos grandes. [3]

Interpretación gráfica de los factores

A continuación, se presenta un gráfico relativo a cuatro variables X1, X2, X3 y X4 representadas por dos factores F1 y F2.

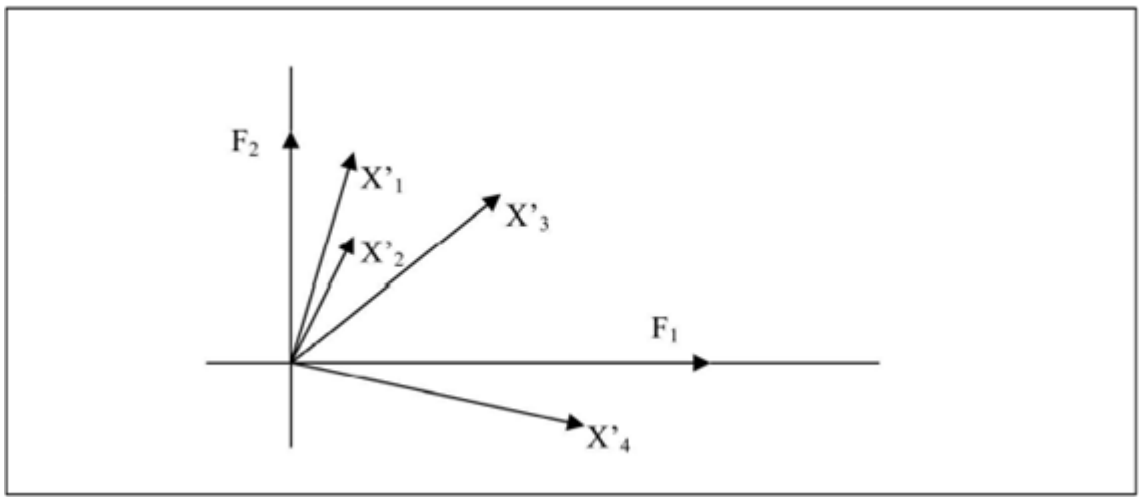

Figura 7. Gráfico relativo a cuatro variables X1, X2, X3 y X4 representadas por dos factores F1 y F2.

Puntuación o Medición de los factores

Una vez identificados y nombrados los factores o componentes latentes de un conjunto de variables, puede sernos de gran utilidad conocer qué puntuaciones obtienen los sujetos o unidades de análisis; esto es: las variables son sustituidas por las unidades de análisis, lo que nos permitirá analizar las similitudes que se den entre individuos, casos o unidades respecto a sus puntuaciones factoriales.

Métodos para estimar coeficientes de puntuación de factor:

- Método *de* regresión. Método para estimar los coeficientes de las puntuaciones factoriales. Las puntuaciones que se producen tienen una media de 0 y una varianza igual al cuadrado de la correlación múltiple entre las puntuaciones factoriales estimadas y los valores factoriales verdaderos. La puntuación puede correlacionarse incluso si los factores son ortogonales.
- Puntuaciones *de* Bartlett. Método para estimar los coeficientes de las puntuaciones factoriales. Las puntuaciones resultantes tienen una media de 0. Se minimiza la suma de cuadrados de los factores exclusivos sobre el rango de las variables.
- Método *de* Anderson-Rubin. Método para calcular los coeficientes para las puntuaciones factoriales; es una modificación del método de Bartlett, que asegura la ortogonalidad de los factores estimados. Las puntuaciones resultantes tienen una media 0, una desviación estándar de 1 y no correlacionan entre sí. [3]

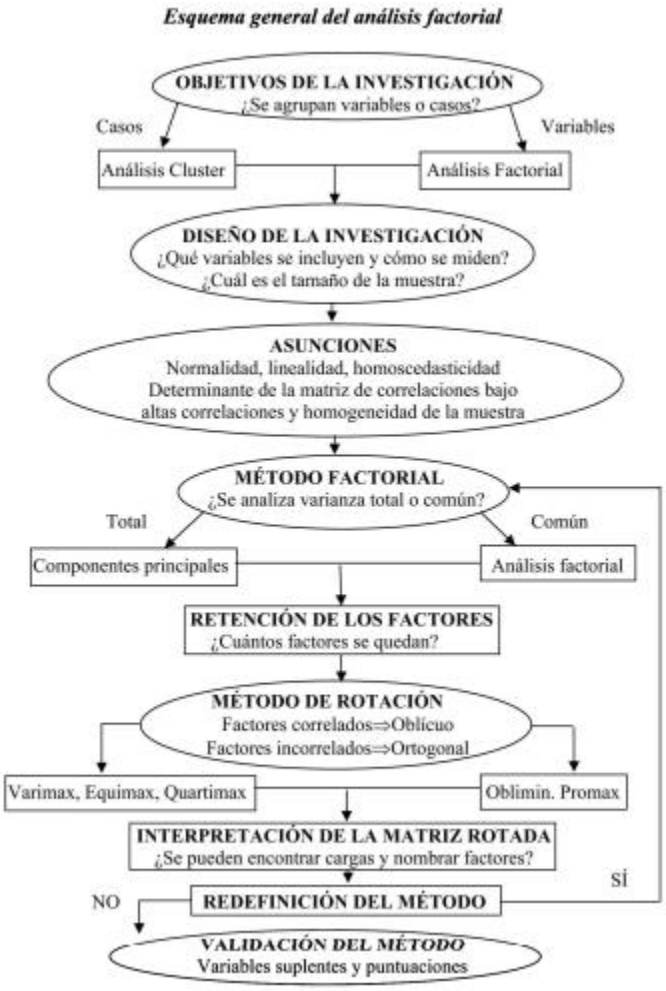

Figura 8. Esquema general del análisis factorial.

Conclusión

La limpieza de datos nos permite descubrir y corregir o eliminar registros de datos erróneos en la tabla de la base de datos. Esto ayuda en proyectos de big data, para la toma de decisiones, teniendo en cuenta que fuentes de información sean confiables, ya que si la información o los datos son erróneos pueden acarrear conclusiones erróneas.

Hay algoritmos y plataformas de aprendizaje automático que proporcionan estos datos y ayudan a analizar de forma descriptiva la situación para cualquier tamaño de empresas. Esto nos demuestra que no basta con tener información disponible, sino que también se debe tener una buena recopilación y comprensión de datos anteriores, una correcta gestión, limpieza, validación de ellos, creación de reglas de negocios en base a los objetivos de las empresas involucradas.

Con respecto a la transformación de datos, se pueden utilizar distintas formas o métodos de transformar los datos de acuerdo a los diferentes tipos de necesidades o requerimientos que se puedan llegar a tener.

Referencias Bibliográfica

[1] Perez Lopez y Santin Gonzalez, Minería de Datos - Técnicas y Herramientas.

[2] Santiago de la Fuente Fernandez, Análisis Factorial, Facultad de Ciencias Económicas y Empresariales, UAM, 2011.

[3] Documentación IBM SPSS Statistics 29.0.0, Septiembre 2022. [Online]. Disponible: :https://www.ibm.com/docs/es/spss-statistics/29.0.0