Finalmente propone una *definición integrada* como: “Concepto que ha surgido en la bibliografía por la “necesidad de explicar la aparición de productos que suponen una transformación radical de un estado anterior, lo cual lleva a inferir la existencia de una persona que posee una constelación de rasgos intelectuales, de personalidad y motivacionales que le capacitan para utilizar la metacognición de un modo óptimo. Dicha utilización óptima permite dar respuesta a un fallo o hueco en el conocimiento. Además, la persona ha de encontrarse inmersa en un contexto carente de obstáculos, que le presente modelos o parangones y le facilite los recursos necesarios”.

3.3 Característica medioambiental de la creatividad

C. González en [5] plantea un entramado ambiental en aspectos cognoscitivos, afectivos, comunicativos, axiológicos, formativos, productivos, lúdicos y políticos que favorecen la creatividad. En la *Ilustración 3*, se resume a través de un mapa conceptual, su planteamiento.

El desarrollo cognoscitivo constituye al fortalecimiento de habilidades de pensamiento, formas y estilos cognitivos, formulación y resolución de retos. El desarrollo comunicativo se caracteriza por la búsqueda de diálogo, argumentación y entendimiento. La dimensión axiológica se fija en los principios, la ética y el respeto a lo humano. El desarrollo productivo atiende a la elaboración, el logro y la innovación. El desarrollo formativo se orienta hacia el aprendizaje significativo, el desarrollo personal y prospectivo. El desarrollo lúdico contribuye a través de la posibilidad, el juego y el disfrute. El desarrollo político, contribuye a la búsqueda de clima, organización, convivencia y equilibrio.

Con base en lo anterior, un ambiente creativo se construye con la interacción de las siguientes dimensiones:

- Ambiente psicosocial
- Ambiente didáctico
- Ambiente físico

Un *ambiente psicosocial*, planteado como un ámbito de de relaciones e interacciones que brinden identidad, seguridad, confianza y autonomía, que favorezca la integración, la comunicación y la manifestación del saber, que genere la alegría y la emoción de crear a través diferentes formas. Un ambiente que adicionalmente que induzca la capacidad de asombro, el interés y la atención, como fuentes primarias del aprendizaje. Un ambiente en donde la reflexión y el conflicto, sean vistos como parte del proceso. El ambiente psicosocial, se concibe también como la afectividad de la interacción entre los actores sociales, y de éstos con su contexto; y como la relación sinérgica entre la organización y los intereses y expectativas de los actores sociales.

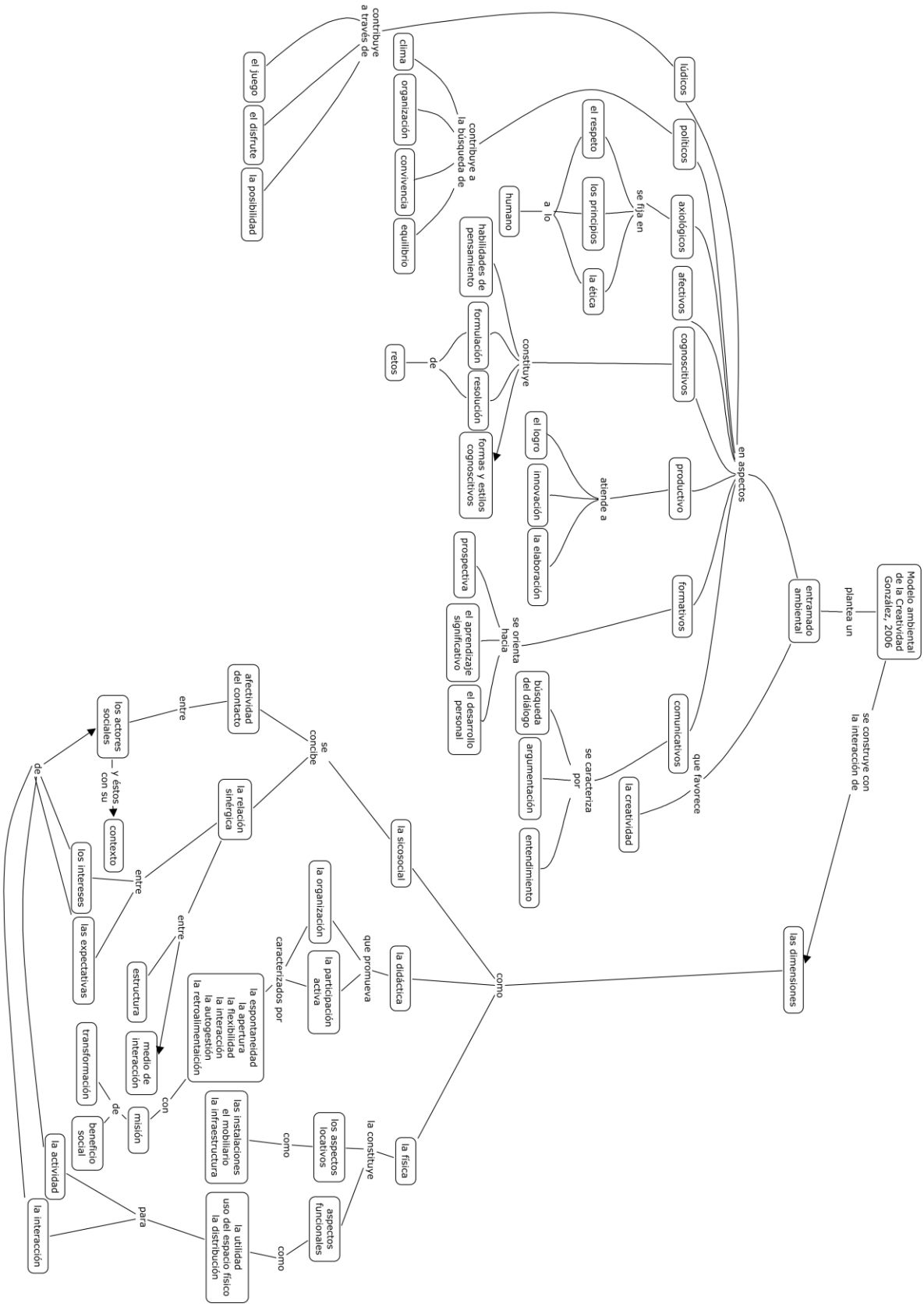

Ilustración 3: Mapa conceptual de modelo medioambiental de la creatividad según González, 2006

Un *ambiente didáctico*, dimensionado como escenario de formación y aprendizaje creativo, donde se promueva la indagación, la formulación y reformulación de problemas, centrados no tanto en la certeza del acierto como en la posibilidad de acierto, donde se generan más preguntas que respuestas. Un ambiente orientado al aprendizaje significativo creativo, a la educación en valores y a la posibilidad de juego, encaminados al logro de los objetivos y las metas en un proceso flexible, abierto y lúdico. Un ambiente didáctico, referido a métodos, técnicas y estrategias creativas, a formas eficientes de asimilación, comprensión y apropiación de los conocimientos y a elementos y materiales educativos significativos. Un ambiente didáctico que promueva la organización y la participación activa, determinadas por la espontaneidad, la apertura, la flexibilidad, la interacción, la autogestión y la retroalimentación permanente.

Un *ambiente físico*, que soporta y dinamiza el ambiente psicosocial y didáctico desde la dimensión funcional, técnica, semiótica y ecológica del espacio. El ambiente físico es un comunicador social a través de las diferentes representaciones y símbolos que materializa; reproduce y manifiesta mensajes traducidos en lo espacial, contiene pedagogías invisibles que por su carácter figurativo, holista y vivencial, los cuales tienen un alto poder interiorizante. Lo constituyen: los aspectos locativos en cuanto a las instalaciones, mobiliario e infraestructura, que conllevan en su diseño, procesos y condiciones simbólicas que expresan los paradigmas de la cultura. También lo constituyen los aspectos funcionales, en cuanto a la utilidad, uso de espacio físico y su distribución para la actividad e interacción de los actores. Finalmente el ambiente físico también está configurado por los aspectos técnicos, referidos a las condiciones ingenieriles y medioambientales en cuanto a recursos naturales, condiciones de iluminación, ventilación, acústica y radiación, para comodidad de las funciones y actividades realizadas al interior del mismo. También se orienta a dar cuenta de la tecnología y sus diversas aplicaciones en el espacio que facilitan los procesos de comunicación, aprendizaje y creación. El ambiente físico debe concebir al ser humano como principio y fin de su toda su meta productiva, propiciando condiciones medioambientales amables, placenteras y apropiadas a los sentidos, tal que favorezcan el aprendizaje y la creación.

3.4 Conclusión del capítulo

El concepto de creatividad ha tenido una evolución que partió de su negación en los griegos hasta una progresiva aceptación y uso hoy, no solamente en el ámbito del arte sino también en la ciencia y en las distintas áreas del quehacer humano.

Su definición también ha presentado perfeccionamiento y actualmente en ella, se incluyen aspectos referentes a la persona, al proceso, al producto y ambiente creativo.

En este último aspecto se debe considerar tres dimensiones, como son la didáctica, la física y la psicosocial que favorecen o inhiben la creatividad como tal.

4. Descubrimiento de conocimiento en bases de datos

En este capítulo se define los conceptos de Descubrimiento de Conocimiento en Bases de Datos y Minería de Datos. A continuación se describe una metodología que plantea el proceso del KDD. Por último, se presentan algunas técnicas de Minería de Datos que se usaron en el marco experimental de esta tesis.

4.1 Definición de KDD

Fayyad et al, citado por [7] define el "descubrimiento de conocimiento en bases de datos" (Knowledge Discovery in Databases, KDD) como *"el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos"*. En esta definición se resumen cuáles deben ser las propiedades deseables del conocimiento extraído:

- *Válido*: hace referencia a que los patrones deben seguir siendo precisos para datos nuevos (con un cierto grado de certidumbre), y no sólo para aquellos que han sido usados en su obtención.
- *Novedoso*: que aporte algo desconocido tanto para el sistema y preferiblemente para el usuario.
- *Potencialmente útil*: la información debe conducir a acciones que reporte algún tipo de beneficio para el usuario.
- *Comprensible*: la extracción de patrones no comprensibles dificulta o imposibilita su interpretación, revisión, validación y uso en la toma de decisiones. De hecho, una información incomprensible no proporciona conocimiento (al menos desde el punto de vista de su utilidad).

Referente al proceso, en [8] se plantea de manera general las etapas o fases a seguir en su desarrollo: integración de los datos, selección y preproceso, minería de datos, interpretación y asimilación (ver *Ilustración 4*).

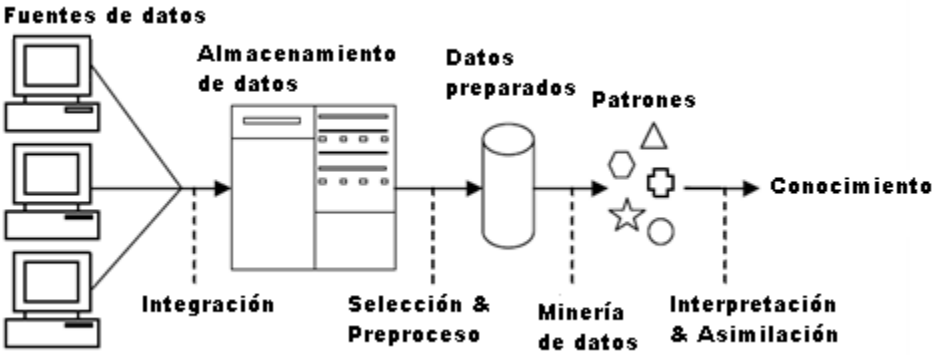

Ilustración 4: Proceso general de KDD (Bramer, 2007)

La Minería de Datos es el núcleo del proceso KDD. Pero entonces ¿qué es? A continuación se describe el concepto.

4.2 ¿Qué es la minería de datos?

El campo de la Minería de Datos es relativamente nuevo y está en estado de evolución. La primera conferencia internacional de KDD y DM, se llevó a cabo en 1995 y hay una variedad de definiciones para este concepto.

En [9] se plantean las siguientes: *extracción de información útil desde grandes conjuntos de datos*. Otra: *DM es el proceso de exploración y análisis, por medios automáticos o semiautomáticos, de grandes cantidades de datos con el fin de descubrir patrones significativos y reglas.*

Otra definición viene de Gartner Group, firma de investigación en tecnología de la información: *DM es el proceso de descubrir nuevas correlaciones significativas, patrones y tendencias por indagación a través de grandes cantidades de datos almacenados en repositorios, usando tecnologías así como técnicas matemáticas y estadísticas.*

En [7] se define la Minería de Datos como *el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos*.

4.3 El proceso de extracción de conocimiento

En la bibliografía se encuentra variadas formas de presentar el proceso de KDD. En general, son similares, cambia es la forma de agrupación y detalle de las fases o etapas planteadas. Se propone la siguiente [10]:

1. Desarrollar un entendimiento del dominio de la aplicación
2. Seleccionar y crear un conjunto de datos en el cual se ejecutará el descubrimiento.
3. Limpieza y preprocesamiento de los datos.
4. Transformación de los datos
5. Elección de la tarea de minería de datos apropiada
6. Elección de algoritmos de minería de datos
7. Empleo del algoritmo de minería de datos
8. Evaluación
9. Uso del conocimiento descubierto

En la *Ilustración 5* se presenta gráficamente.

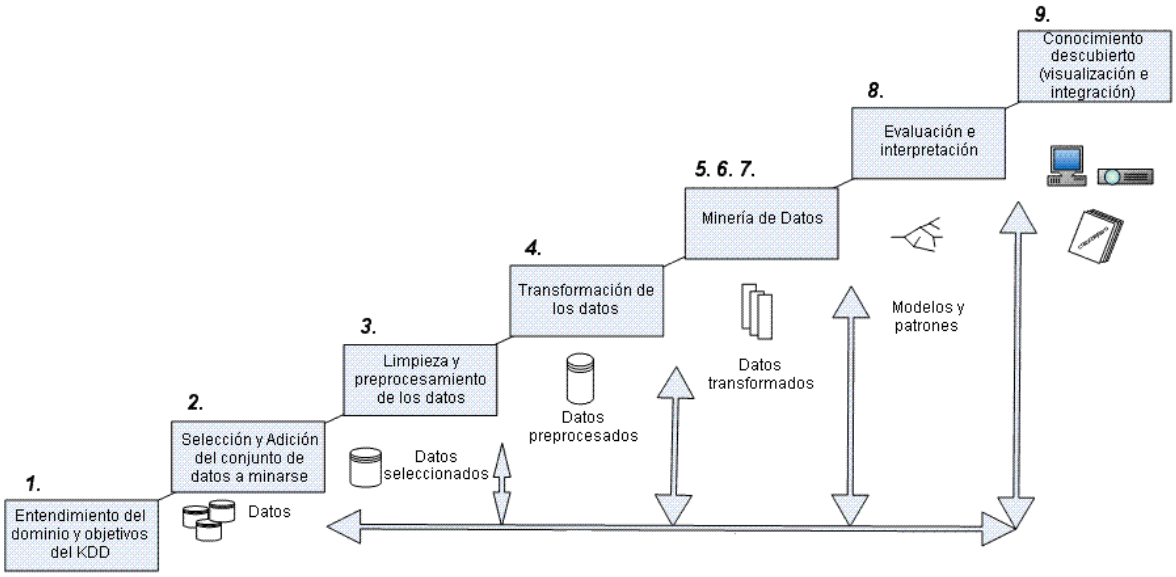

Ilustración 5: Proceso KDD

A continuación se describe brevemente cada paso, con base en [10] y [11].

1. **Desarrollar un entendimiento del dominio de la aplicación.**

En esta fase se prepara la escena para entender qué se debe hacer con varias decisiones (acerca de la transformación, algoritmos, representación, entre otros). El encargado del proyecto KDD necesita entender y definir los objetivos del usuario final y el ambiente en el cual tomará lugar el proceso de descubrimiento de conocimiento (incluyendo conocimiento previo relevante). A medida que avanza el proceso KDD, se puede hacer una revisión de este paso.

Teniendo entendido los objetivos del KDD, se inicia el preprocesamiento de los datos, definidos en los tres próximos pasos.

2. **Seleccionar y crear un conjunto de datos en el cual se ejecutará el descubrimiento.**

Teniendo definidos los objetivos, se deben determinar los datos que serán usados para el descubrimiento del conocimiento. Esto incluye averiguar qué datos están disponibles, la obtención de datos adicionales necesarios y a continuación la integración de todos los datos para el descubrimiento de conocimientos en un conjunto de datos, incluyendo los atributos que se considerarán para el proceso. Este proceso es muy importante porque la DM aprende y descubre conocimiento de los datos disponibles. Si faltan algunos datos importantes, el estudio entero puede fallar. Por otra parte, recoger, organizar y operar complejos repositorios de datos es costoso y hay una compensación con la posibilidad de comprender mejor los fenómenos.

3. Limpieza y preprocesamiento de los datos.

En esta etapa, la confiabilidad en los datos se eleva. Incluye la claridad en los datos, tal como el manejo de valores faltantes y la remoción de ruido o datos anómalos.

Puede convertirse en la mayor parte (en términos de tiempo invertido) de un proyecto de KDD. Puede involucrar métodos estadísticos complejos o el uso de algoritmos DM en este contexto. Por ejemplo, si uno sospecha que cierto atributo de confiabilidad insuficiente o tiene muchos datos faltantes, entonces este atributo puede convertirse en el objetivo de un algoritmo de minería de datos supervisado. Se desarrollará un modelo de predicción para este atributo y a continuación los datos faltantes se pueden predecir.

4. Transformación de los datos

En esta etapa, se preparan mejores datos para la DM. Incluye la reducción de dimensión (tales como selección de características y extracción y registro de muestras), y transformación de atributos (tales como discretización de atributos numéricos y transformación funcional). Este paso puede ser crucial para el éxito del proyecto KDD entero, y suele ser específico para cada proyecto.

Teniendo completos los cuatros pasos mencionados, los siguientes cuatro pasos están relacionados con la parte de minería de datos, en donde el foco está en los aspectos algorítmicos para cada proyecto.

5. Elección de la tarea de minería de datos apropiada

En este punto estamos listos para decidir cuál tipo de tarea de DM usar, por ejemplo, clasificación, regresión o agrupamiento. Esto depende sobre todo de los objetivos de KDD, y también en los pasos anteriores.

Hay dos objetivos principales en Minería de Datos: predicción y descripción. La predicción se refiere a menudo a minería de datos supervisada, mientras la descriptiva incluye aspectos no supervisados y visualización de la minería de datos.

Hablamos de supervisado cuando se parte de un conocimiento previo de los datos. En cambio, en el método no supervisado se buscan automáticamente grupos de valores para que después el usuario intente encontrar las correspondencias entre esos grupos seleccionados automáticamente y las categorías que le puedan ser de interés.

La mayoría de técnicas de DM están basadas en aprendizaje inductivo, en donde un modelo se construye explícita o implícitamente por la generalización desde un número suficiente de ejemplos de entrenamiento. El supuesto subyacente del enfoque inductivo es que el modelo entrenado es aplicable a casos futuros.

6. Elección de algoritmos de minería de datos

Teniendo la estrategia, ahora se decide la táctica. Esta etapa incluye la selección del método específico que se usará para buscar patrones. Por ejemplo, considerando la precisión versus el entendimiento, el primero es mejor con las redes neuronales, mientras que el segundo es mejor con árboles de decisión. Para cada estrategia de meta-aprendizaje hay varias posibilidades de cómo puede ser logrado. El meta-aprendizaje se centra en explicar lo que hace que un algoritmo de minería de datos sea exitoso o no en un problema particular. Así, este enfoque intenta entender las condiciones bajo las cuales un algoritmo de DM es más apropiado.

7. Empleo del algoritmo de minería de datos

Finalmente se alcanza la implementación del algoritmo de minería de datos. En este paso se puede necesitar emplear el algoritmo varias veces hasta que se obtienen resultados satisfactorios, por ejemplo por la puesta a punto de los parámetros de control del algoritmo, tal como el número mínimo de instancias en una sola hoja de un árbol de decisiones.

8. Evaluación

En esta etapa se evalúan e interpretan los patrones minados (reglas, confiabilidad, etc), con respecto a los objetivos definidos en el primer paso. Aquí se consideran los pasos de preprocesamiento con respecto a sus efectos en los resultados del algoritmo de DM (por ejemplo, adición de características en el paso 4, y repetir desde allí). Este paso se centra en la comprensibilidad y utilidad del modelo inducido. En este paso el conocimiento descubierto es también documentado para su uso más adelante.

El último paso es el uso y retroalimentación general sobre los patrones y resultados de descubrimiento, obtenidos por la Minería de Datos.

9. Uso del conocimiento descubierto

Ahora se está listo para incorporar el conocimiento en otros sistemas para acciones más adelante. El conocimiento llega a ser activo en el sentido que se pueden hacer cambios al sistema y medir los efectos. En realidad el éxito de este paso determina la efectividad del proceso entero de KDD. Hay muchos retos en este paso, tales como pérdida de las "condiciones de laboratorio" bajo las cuales han operado. Por ejemplo, el conocimiento fue descubierto desde una cierta instantánea estática (usualmente una muestra) de los datos, pero ahora los datos son dinámicos. La estructura de los datos pueden cambiar (ciertos atributos llegan a ser no disponibles), y se puede modificar el dominio de los datos (tal como, un atributo puede tener un valor que no fue asumido antes).

A continuación se amplían conceptos relacionados con lo antedicho.

4.4 Tareas de minería de datos

Uno de los aspectos que se debe tener claro en el proceso KDD es distinguir entre una tarea y un método de minería de datos. Las tareas de DM pueden ser predictivas y las descriptivas [7].

Las *tareas predictivas* tratan de problemas y tareas en los que hay que predecir uno o más valores para uno o más ejemplos. Dependiendo de cómo sea la correspondencia entre los ejemplos y los valores de salida y la presentación de los ejemplos podemos definir varias tareas predictivas:

- *Clasificación (o discriminación)*
- *Clasificación suave*
- *Estimación de probabilidad de clasificación*
- *Categorización*
- *Preferencias o priorización*
- *Regresión*

Las *tareas descriptivas* buscan describir los datos existentes. A continuación las tareas descriptivas más delimitadas:

- *Agrupamiento (clustering)*
- *Correlaciones y factorizaciones*
- *Reglas de asociación*
- *Dependencias funcionales*
- *Detección de valores e instancias anómalas*

4.5 Técnicas de minería de datos

Cada una de las tareas anteriores, requieren métodos, técnicas o algoritmos para resolverlas.

Veamos algunas técnicas:

- *Técnicas algebraicas y estadísticas*
- *Técnicas bayesianas*
- *Técnicas basadas en conteos de frecuencias y tablas de contingencia*
- *Técnicas basadas en árboles de decisión y sistemas de aprendizaje de reglas*
- *Técnicas relacionales, declarativas y estructurales*
- *Técnicas basadas en redes neuronales artificiales*
- *Técnicas basadas en núcleo y máquinas de soporte vectorial*
- *Técnicas estocásticas y difusas*
- *Técnicas basadas en casos, en densidad o distancia*

La *Tabla 1: Algunas técnicas de Minería de DatosTabla 1*, tomada de [7], muestra algunas técnicas o algoritmos asociadas al tipo de tarea que realizan.

Nombre	PREDICTIVO		DESCRIPTIVO		
	Clasificación	Regresión	Agrupamiento	Reglas de asociación	Correlaciones/ Factorizaciones
Redes neuronales	✓	✓	✓		
Árboles de decisión ID3, C5.0	✓				
Árboles de decisiones CART	✓	✓			
Otros árboles de decisión	✓	✓	✓	✓	
Redes de Kohonen			✓		
Regresión lineal y logarítmica		✓			
Regresión logística	✓			✓	
Kmeans			✓		
Apriori				✓	
Naive Bayes	✓				
Vecinos más próximos	✓	✓	✓		
Análisis factorial y de componentes principales					✓
Twostep, Cobweb			✓		
Algoritmos genéticos y evolutivos	✓	✓	✓	✓	✓
Máquinas de vectores de soporte	✓	✓	✓		
CN2 rules (cobertura)	✓			✓	
Análisis discriminante multivariante	✓				

Tabla 1: Algunas técnicas de Minería de Datos

A continuación se presentan y describen sucintamente, las técnicas e instrumentos de Minería de Datos usados en este trabajo.

4.5.1 Fiabilidad de los datos: Alfa de Cronbach

Se usa en la etapa de limpieza y preprocesamiento de los datos. Es el método más usado para determinar si el instrumento evaluado (como en este caso, la encuesta) es estable y consistente o por el contrario recopila información defectuosa y por tanto, no podrí usarse porque nos llevaría a conclusiones equivocadas.

Se trata de un índice de consistencia interna que es el camino más habitual para estimar la fiabilidad de pruebas, escalas o test, cuando se utilizan conjuntos de ítems o reactivos que se espera midan el mismo atributo o campo de contenido. El Alfa de Cronbach es, sin duda, el más ampliamente utilizado por los investigadores. Alfa estima el límite inferior del coeficiente de fiabilidad y se expresa como [12]:

$$\alpha = \left(\frac{k}{k-1}\right).\left(1 - \frac{\sum S_i^2}{S_{sum}^2}\right)$$

Donde k es el número de ítems de la prueba, S_i^2 es la varianza de los ítems (desde 1...i) y S_{sum}^2 es la varianza de la prueba total. El coeficiente mide la fiabilidad del test en función de dos términos: el número de ítems (o longitud de la prueba) y la proporción de varianza

total de la prueba debida a la covarianza entre sus partes (ítems). Ello significa que la fiabilidad depende de la longitud de la prueba y de la covarianza entre sus ítems.

Para poder calcular la fiabilidad de una prueba, este debe cumplir con dos requisitos previos:

1. Estar formado por un conjunto de ítems que se combinan aditivamente para hallar una puntuación global (esto es, la puntuaciones se suman y dan un total que es el que se interpreta).
2. Todos los ítems miden la característica deseada en la misma dirección. Es decir, los ítems de cada una de las escalas tienen el mismo sentido de respuesta.

Su valor va de *0* a *1*, considerándose *1* la mejor fiabilidad y adecuada, desde *0,8*.

4.5.2 Representación gráfica: diagramas de dispersión

Usado en la etapa de limpieza y preprocesamiento de los datos. También ayuda a determinar las tareas de minería a seguir. Muy útil cuando los datos son de más de tres dimensiones y se requiere un primer acercamiento a la información contenida en ellos.

Los diagramas de dispersión son gráficos en los que se representan las observaciones por puntos asociados a cada par de coordenadas. En la *Ilustración 6* se puede observar un ejemplo de diagrama de dispersión sobre tres variables.

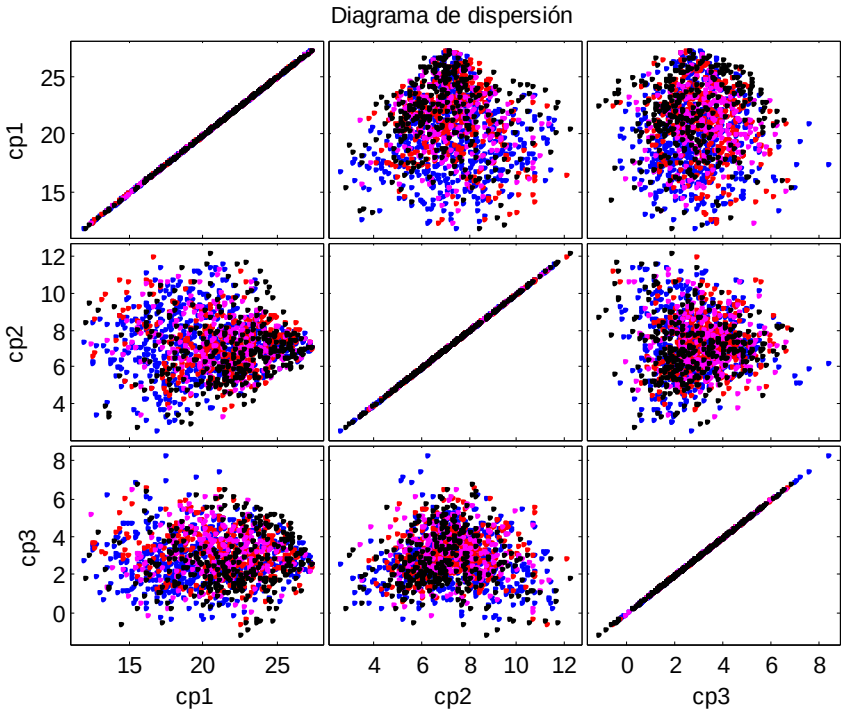

Ilustración 6: Diagrama de dispersión

4.5.3 Detección de datos atípicos

Los datos atípicos son aquellas observaciones extremas que se apartan de los demás datos [13]. Su presencia puede ser grave: distorsionan las medias y desviaciones típicas de las variables y destruyen las relaciones existentes entre ellas [14].

Se deben considerar dos aspectos con respecto a los datos atípicos: uno, su detección o identificación, para el cual existen herramientas gráficas y de cálculo; otro, es el tratamiento de las observaciones detectadas como atípicas, lo que puede implicar la modificación de los datos, cuando se descubre que son causados por errores de medición, registro o de concepto.

La detección de datos atípicos hace parte de la etapa de limpieza y preprocesamiento de los datos.

De todos modos, se debe tener cuidado con su manejo, ya que su misma aparición puede llevar consigo circunstancias que pueden aportar información interesante sobre el problema analizado.

Un algoritmo para detectar las observaciones atípicas, es la usada en la función *findoutd* del toolbox de Reconocimiento de Patrones de Matlab [15], que se basa en considerar dato atípico aquel que esté alejado *n* desviaciones de la media con el resto de datos en un porcentaje dado del total. Normalmente se toman observaciones cuyas distancias superan la media en *3* desviaciones estándar con el *10%* o más de los datos.

Se usará este algoritmo por la simpleza de concepto y su potencia al realizar la labor de detección.

4.5.4 Reducción de dimensión: Análisis de Componentes Principales

Hace parte de la etapa de transformación de los datos.

El Análisis de Componentes Principales es el método no supervisado más popular en la reducción de dimensiones de un conjunto de datos, además de ser ampliamente validado.

Los parámetros del modelo pueden ser calculados directamente de la matriz de datos centralizada X bien sea por descomposición en valores singulares o la diagonalización de la matriz de covarianza [16]. Sea x_i el i-ésimo vector de observación de longitud p, siendo $x = (x_1, x_2, \ldots, x_n)^T$, la matriz de rotación U con la que se calculan las p' componentes principales denotadas por z que resumen x, como sigue.

$$z = U^T x$$

U puede ser calculada a partir de los primeros p' valores propios de X^TX, esto es,

$$X^T X U = U \Lambda$$

Se debe establecer un criterio para determinar cuántas componentes se usarán para la representación de los datos. Veamos algunas [14]:

1. Realizar un gráfico (de sedimentación) de los valores propios ordenados contra su respectiva posición ordinal y seleccionar los primeros componentes hasta que el resto de componentes cambie poco de valor, es decir, cuando hayan generado un codo en el gráfico.
2. Seleccionar los primeros componentes hasta que cubran un valor de varianza acumulada, como por ejemplo el *80%* o el *90%*.
3. No tener en cuenta las componentes asociadas a valores propios inferiores a una cota, que normalmente se fija como la varianza media.

4.5.5 Reducción de dimensión: Escalamiento multidimensional

Es una técnica que se puede enmarcar en la etapa de transformación de los datos. Favorece el análisis de los datos en dos o tres dimensiones.

El Análisis de Componentes Principales se limita a la búsqueda de combinaciones lineales de las características de los datos originales a mapear en menos dimensiones. Esto es suficiente para muchas aplicaciones.

Sin embargo, si el objetivo de un mapeo es inspeccionar los datos en proyecciones de dos o tres dimensiones, PCA puede descartar muchísima información. Por ejemplo, se puede proyectar dos grupos distintos en los datos uno encima del otro, o no pueden mostrar si los datos se distribuyen en forma no lineal en el espacio de medida de alta dimensión original.

Para conservar esa información, se necesita un método de asignación no lineal. Como hay muchos criterios posibles para adaptarse al mapeo, hay muchos métodos diferentes disponibles. Uno de ellos es el escalamiento multidimensional (MDS).

MDS se basa en la idea de que la proyección debe mantener las distancias entre objetos, tanto como sea posible. Dado un conjunto de datos que contiene los N vectores de medición (o vectores de características) z_i $i = 1,. . . N_S$, tratamos de encontrar D nuevos y_i $i = 1,. . . N_S$ de acuerdo con este criterio. Por lo general, $D << N$; si el objetivo es visualizar los datos, elegimos $D = 2$ o $D = 3$. Si δ_{ij} denota la distancia conocida entre los objetos z_i y z_j y d_{ij} denota la distancia entre los objetos proyectados y y_i y y_j, entonces las distancias se pueden conservar lo mejor posible, colocando y_i tal que la medida de estrés [17]

$$E_S = \frac{1}{\sum_{i=1}^{N_S} \sum_{j=i+1}^{N_S} \delta_{ij}^2} \sum_{i=1}^{N_S} \sum_{j=i+1}^{N_S} \left(\delta_{ij} - d_{ij}\right)^2$$

se minimiza. Para ello, tomamos la derivada de la ecuación con respecto a los objetos *y*. Se deduce que cuando las distancias euclidianas se usan, entonces

$$\frac{\partial E_S}{\partial \mathbf{y}_i} = -\frac{2}{\sum_{j=1}^{N_S}\sum_{k=j+1}^{N_S}\delta_{jk}^2}\sum_{\substack{j=1\\ j\neq i}}^{N_S}\frac{\delta_{ij}-d_{ij}}{d_{ij}}(\mathbf{y}_i-\mathbf{y}_j)$$

4.5.6 Análisis de Conglomerados

Partiendo de la definición de Everit (citado por [13]), los conglomerados son *"regiones continuas de un espacio que contiene una densidad relativamente alta de puntos, los cuales están separados por regiones que contienen una densidad relativamente baja de puntos".*

Esta tarea está enmarcada en la etapa de empleo del algoritmo de minería de datos, dentro del proceso KDD.

Agrupamiento por K medias

El agrupamiento por K medias es un método de partición de datos que busca dividirlos en un número de grupos definido previamente, de forma que cada observación pertenezca a uno y solamente a un grupo; toda observación quede clasificada y cada grupo sea interiormente homogéneo [14].

El algoritmo sigue los siguientes pasos [17]:

1. Asocia aleatoriamente cada observación a uno de los grupos *k=1,2,…,K.*
2. Calcula la media de cada uno de los grupos:

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{\mathbf{z}_i \in C_k}\mathbf{z}_i$$

3. Reasigna cada observación z_i al grupo con la media más cercana μ_k.
4. Retorna al paso 2 hasta que las medias de los grupos no cambien más

Es un método ampliamente validado de partición de datos.

Agrupamiento Jerárquico

Se puede incluir en las etapas de elección de tareas y algoritmos de minería de datos. Ayuda a mostrar la estructura general de los datos y los posibles números de grupos existentes.

La idea básica de agrupamiento jerárquico es recoger los objetos en grupos mediante la combinación de los objetos más cercanos y los grupos más grandes grupos hasta que todos los objetos están en un grupo. Una ventaja importante es que los objetos no sólo son colocados en K grupos distintos, sino que se colocan en una jerarquía de grupos. Esto da más información sobre la estructura del conjunto de datos, y muestra que las agrupaciones son similares o diferentes. Esto hace que sea posible detectar subgrupos en grupos grandes, o para detectar grupos de valores atípicos [17].

El algoritmo es el siguiente:

1. Asignar a cada objeto a su propio grupo, dando lugar a N_S grupos, cada uno con un solo objeto. La distancia inicial entre todos los grupos, por lo tanto sólo las distancias entre todos los objetos.
2. Encontrar el más cercano par de grupos y fusionarlos en un único grupo, de modo que el número de grupos se reduce a uno.
3. Calcular la distancia entre el nuevo grupo y cada uno de los antiguos grupos, donde la distancia entre los dos grupos se puede definir de varias maneras (véase más abajo).
4. Repita los pasos 2 y 3 hasta que todos los objetos se agrupan en un único grupo de tamaño N_S, o hasta que se logre un número predefinido de K grupos.

En el paso 3 es posible calcular las distancias entre las agrupaciones de varias maneras diferentes. Podemos distinguir de *enlace simple*, de *enlace promedio* y de *enlace completo*. En una sola agrupación de enlace, la distancia entre dos grupos de C_i y C_j se define como la distancia más corta de cualquier objeto en un grupo a cualquier objeto en el otro grupo:

$$d_{sl}(C_i, C_j) = \min_{\mathbf{x} \in C_i, \mathbf{y} \in C_j} \|\mathbf{x} - \mathbf{y}\|^2$$

Para el agrupamiento de enlace promedio, el operador mínimo se sustituye por la distancia media, y para el de enlace completo se reemplaza por el operador máximo.

Otra manera de construir el agrupamiento jerárquico fue propuesto por Ward y Wishart. Se diferencia de los anteriores, en que este parte de los elementos directamente, en lugar la matriz de distancias y se define una medida global de heterogeneidad de una agrupación de observaciones en grupos [14].

El dendograma

El dendograma o diagrama de árbol jerárquico ilustra gráficamente el resultado de agrupamiento jerárquico en forma de árbol (*Ilustración 7*). Por lo tanto, también enmarcado en las etapas de elección de tareas y algoritmos de minería de datos.

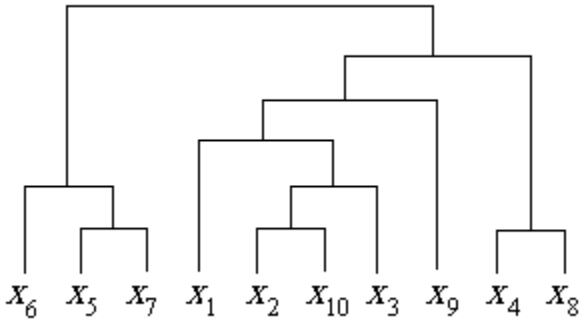

Ilustración 7: Dendograma

Determinación del número de grupos: Criterio de Hartigan

Hace parte de las etapas de elección de tareas y algoritmos de minería de datos.

No existe un criterio universal para determinar el número de grupos. El criterio de Hartigan es regla empírica que ha dado resultados razonables ([18] y [14]) fue dada por el autor del mismo apellido en 1975.

Consiste en ir calculando el valor de *F* para dos números de grupos consecutivos e introducir otro si este valor es mayor que 10. El valor de *F* se calcula con base en:

$$\boldsymbol{F = \frac{SCDG_G - SCDG_{G+1}}{\frac{SCDG_{G+1}}{n - G - 1}}}$$

Donde *G* es el número de grupos, *n* es el número de datos y *SCDG* es la suma de cuadrados dentro de los grupos dado por:

$$\boldsymbol{\sum_{g=1}^{G} \sum_{j=1}^{p} \sum_{i=1}^{n_g} (x_{ijg} - \overline{x}_{jg})^2}$$

4.6 Conclusión del capítulo

Existen variadas técnicas que permiten analizar los datos multivariados y así lograr un mayor conocimiento de ellos. Se debe tener la claridad sobre cuál es posible usar para un análisis específico.

Por lo anterior, es importante ayudarse de una metodología adecuada para el Descubrimiento de Conocimiento en Bases de Datos, que al aplicarla, favorece un trabajo más organizado y sistematizado durante el análisis y al validar los resultados.

5. Estado del arte

A continuación se presentan, de manera cronológica, estudios de los últimos cinco años sobre creatividad, basados en técnicas de minería de datos.

Las áreas involucradas van desde la psiquiatría, neurología, educación, mercadeo, análisis organizacional, entre otras.

5.1 Revisión del estado del arte

Cartwright et al. [19] presentan un estudio cuyo objetivo fue probar la hipótesis de que los procesos neurales que subyacen a la epilepsia del lóbulo temporal (ELT) también facilitan la creatividad. Para medir el potencial de creatividad, se usó la Prueba de Preferencia de Figuras de Polígono (rango de figuras de diferente complejidad) en 150 participantes conformados en 3 grupos. Su usó Análisis de Regresión Jerárquica para el análisis de los datos, entre otros estadísticos.

En [20] se describe los fundamentos teóricos y el desarrollo de un instrumento de medición (lista de chequeo) que provee a los maestros una herramienta para observar las características de la creatividad personal de cada alumno. Fue administrado en 24 aulas a 520 estudiantes de grado 6 y 7 de elemental. Se utilizó análisis factorial usando extracción de máxima verosimilitud con una rotación Oblimin para evaluar la estructura factorial de la encuesta. La correlación de Pearson fue usada para establecer las relaciones que existían entre los elementos individuales de la escala.

Los investigadores Wu et al. [21] dan a conocer un estudio que exploró actuaciones en 3 tipos de tareas de los Test de Pensamiento Creativo de Torrance (un problema del mundo real, completar figuras, y la verbal, con cajas de cartón) en 22 estudiantes de 6º grado y 22 estudiantes universitarios de Hong Kong. Fueron evaluados con MANOVA y la prueba *F* univariada y multivariada.

En el trabajo de Burch et al. [22] se plantea un estudio en el que se propuso investigar la relación entre la creatividad, la esquizotipia multidimensional y de la personalidad en general. Esto se logró mediante el análisis de las puntuaciones en un rango de escalas de personalidad y de las medidas de la creatividad sobre 7 pruebas de pensamiento divergente y medidas sicométricas. Se hizo sobre cien estudiantes de pregrado voluntarios (67 hombres y 33 mujeres). Para su evaluación se usó ANOVA, Correlación de Pearson y Análisis Factorial.

Por otro lado, en [23] se presenta un estudio en el que se examina la composición de la creatividad de productos mediante el examen de los elementos de los productos de consumo desde la perspectiva del consumidor. Se usó una encuesta de 52 puntos, basada en la Matriz de Análisis del Producto Creativo, completada por 205 estudiantes universitarios. Para su análisis utilizaron Análisis de Factores exploratorio y confirmatorio, y regresión paso a paso.

En [24] se examinó la posibilidad de una relación curvilínea entre la presión de tiempo creativo experimentado por los empleados y la creatividad. Los datos fueron obtenidos a partir de una encuesta que incluía ítems de evaluación de la educación, presión de tiempo creativo experimentado, apertura a la experiencia y apoyo a la creatividad, la cual se realizó a 170 empleados y 10 supervisores de una organización de la producción. Para su evaluación se usó el índice Alfa de Cronbach para verificar la consistencia de las encuestas. Análisis de Regresión Jerárquico se usó para la probar de hipótesis de los autores.

En el trabajo de Kim [25] da cuenta de una investigación en la que se usó Análisis Factorial confirmatorio con datos de 500 estudiantes de grado-6 (242 niños y 258 niñas) y se probaron varios modelos de factor sobre los Test de Pensamiento Creativo de Torrance. Los resultados de este estudio muestran que estas pruebas constan de 2 factores en lugar de uno solo, a diferencia de la mayoría de las investigaciones sobre este tema. Adicionalmente se usó curtosis multivariada, mínimos cuadrados de aproximación (RMSEA) y mínimos cuadrados estandarizado residual (SRMR).

En [26] se analiza la asociación de la edad, años de escolaridad y conocimiento de dominio específico en el desarrollo de la creatividad de los niños en matemáticas. Se usó la evaluación matemática DISCOVER para medir el conocimiento de matemáticas y los índices de creatividad de originalidad, flexibilidad, elaboración y fluidez en 841 estudiantes de primero a quinto grado de 4 escuelas. Para su evaluación se usó análisis de regresión jerárquica, análisis de regresión múltiple y MANOVA entre grupos.

Charyton y Snelbecker [27] hacen referencia a una investigación realizada con estudiantes de ingeniería, hombres (n = 85) y mujeres (n = 18) y estudiantes de música, hombres (n = 49) y mujeres (n = 50) en la que se plantean algunas preguntas sobre el uso de adjetivos como una metodología para detectar las similitudes y diferencias en la creatividad. Para el análisis de resultados se usó análisis de varianza de dos vías y análisis Tukey post hoc (comparación múltiple de medias).

Asimismo en [28] se presentan tres estudios que examinaron la relación entre la inteligencia emocional (IE) y la creatividad emocional (CE) y si cada constructo fue predictivo de comportamiento creativo. Se planteó la hipótesis de que la relación entre la IE y la CE, corresponde a la relación entre la inteligencia cognitiva y la capacidad creativa. Por lo tanto, se esperaba que la IE y la CE fueran dos grupos distintos de habilidades. Se usaron 12 pruebas distintas de medición sicométrica. Con el uso de análisis de correlación y análisis factorial confirmatorio, apoyaron la hipótesis.

Igualmente en [29] se revisaron los datos de una investigación de Wallach y Kogan (1965) en 151 niños con cinco pruebas sobre la relación entre creatividad e inteligencia y se volvieron a analizar utilizando el análisis de variables latentes (que no estaba disponible para los investigadores de la época), pudiendo evaluar las relaciones entre las variables latentes de creatividad e inteligencia. Afirman los autores que los modernos métodos estadísticos pueden ofrecer nuevas ideas sobre las investigaciones anteriores e inspirar nuevas investigaciones y resultados.

Wu y Chiou, en [30] presentan un estudio en el que se investigó la relación entre los niveles de desarrollo cognitivo y el rendimiento creativo entre jóvenes al final de su adolescencia desde una perspectiva post-piagetiana. Los participantes fueron 386 estudiantes universitarios, de edades comprendidas entre 19 a 26 años. Se utilizó el instrumento Social Paradigm Belief Scale para medir los tres estilos cognitivos del final de la adolescencia: el pensamiento formal, relativista y dialéctico. La Prueba de Pensamiento Divergente (DTT) se utilizó para medir el desempeño creativo. Para el análisis de resultados se utilizaron Análisis de Correlación en donde el pensamiento dialéctico y relativista se correlacionó positivamente con el desempeño creativo, mientras que el pensamiento formal se correlacionó negativamente. Con MANOVA se reveló que los pensadores posformales tuvieron puntajes más altos que los pensadores formales en todas las dimensiones de la creatividad. Análisis Discriminante Múltiple mostró que la combinación lineal de las seis dimensiones de la creatividad reconocido en la DTT discrimina entre pensadores formales y postformales que apoya la conclusiones de los autores acerca de la relación entre el pensamiento postformal y la creatividad posformal. Se discuten futuras líneas de investigación y las implicaciones para la pedagogía de la creatividad. También se utilizaron las puebas Z y T.

Freund y Holling, en [31] utilizan Análisis Multinivel (modelación jerárquica) para evaluar 1.133 estudiantes agrupados en 60 aulas, partiendo del test *Berlin Structure of Intelligence Test for Youth: Assessment of Talent and Giftedness*. Según lo planteado por los autores, su uso permitió crear una imagen más minuciosa de la relación entre la creatividad, la capacidad de razonamiento y el rendimiento escolar. Se usó también mínimos cuadrados generalizados iterativo para estimación de parámetros.

Por último, en [32] se menciona la creación de un instrumento de Evaluación del Diseño Creativo en Ingeniería (CEDA, por sus siglas en inglés), aplicado a 63 estudiantes de ingeniería (57 hombres y seis mujeres) y a 21 estudiantes de otras carreras distintas a la ingeniería (seis hombres y 15 mujeres), el cuál con cuatro cuestionarios distintos pretende evaluar el diseño creativo a nivel de persona, proceso, producto y ambiente. Para el análisis de resultados se utilizaron el análisis de correlación, el índice de consistencia interna Alfa de Cronbach, Análisis Multivariado de Varianza (MANOVA) de una vía y de dos vías para examinar el efecto de clase y género en las medidas de la creatividad general antes y después de sus clases, el análisis de Kruskal Wallis para analizar el efecto de la clase y género en el diseño creativo en ingeniería, la prueba de Mann-Whitney y el análisis Welch y Brown-Forsythe para comparar diseño creativo en ingeniería, entre otros estadísticos típicos.

5.2 Conclusión del capítulo

En la revisión del estado del arte, esencialmente se ha percibido que todavía es incipiente o por lo menos no tan generalizado, el uso de estas técnicas de minería de datos para la evaluación de la creatividad.

También nos brindó elementos para reafirmar una suposición que teníamos desde el inicio de este trabajo de investigación: el uso de estas herramientas permiten extraer nuevo conocimiento de investigaciones anteriores, en las cuales por distintas razones, no se hizo uso de ellas.