VNIVERSIDAD NACIONAL AVFNOMA DE MEXICO

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

DESCUBRIMIENTO DE CONOCIMIENTO EN UNA BASE DE DATOS DE ASPIRANTES A EDUCACIÓN MEDIA SUPERIOR

T E S I S

QUE PARA OBTENER EL GRADO DE:

MAESTRA EN INGENIERÍA

(COMPUTACIÓN)

P R E S E N T A:

OLGA LIDIA ACOSTA LÓPEZ

DIRECTORA DE TESIS: DRA. AMPARO LÓPEZ GAONA

México, D.F. **2006.**

Agradecimientos

Deseo aprovechar este espacio para agradecer a las personas que han contribuido directa e indirectamente a la realización de este trabajo, así como también a aquellas que han apoyado de forma cordial y eficiente cada una de las gestiones que hemos realizado como estudiantes del posgrado en Ciencia e Ingeniería de la Computación.

A la Dra. Amparo López Gaona, por su ánimo, paciencia y apoyo que me brindó generosamente como Asesora de esta tesis.

A los lectores que revisaron y enriquecieron este trabajo: la Dra. Hanna Oktaba, el Mtro. Javier García García, la Mtra. Cecilia Pérez Colin y el Dr. Ragueb Chain Revuelta.

Al Dr. Ragueb Chain Revuelta, un especial agradecimiento por haber facilitado el acceso a los datos con los que se llevó a cabo este trabajo y por el tiempo invertido en la conducción y revisión del análisis.

A Lulú, Amalia y Diana, por la disposición y el apoyo incondicional brindado durante los casi tres años de estudio y trabajo en el posgrado.

A mis compañeros y amigos de generación con quienes compartí momentos muy gratos y que han contribuido también al buen logro de esta meta.

A mis padres (Valentín y Elena), hermanos (Valentín y Rafael) y sobrinitos (Mariela, Saúl, Iván y Ángel de Jesús), por ser la fuente de inspiración constante en mi vida.

A César: por representar también esa *condición de felicidad* con la que se emprende sin miedo cualquier reto en la vida.

Finalmente, a Dios, por crear siempre los escenarios perfectos para el *Aprendizaje Significativo*.

La presente investigación contó con el apoyo del Consejo Nacional de Ciencia y Tecnología (CONACYT), durante el periodo en que se realizaron los estudios de posgrado (septiembre 2003-julio 2005).

México, D.F. 23 de marzo de 2006.

Índice

Introducción

La observación y la búsqueda, son dos procesos claves que conducen al descubrimiento. Sin duda, desde el nacimiento, se tiene la fortuna de observar y buscar en el entorno que nos rodea para desarrollar estas habilidades, ya sea siguiendo una teoría formal o de manera casual. De forma paralela, se desarrolla la agudeza mental y visual necesaria, que a su vez permite mejorar los procesos de descubrimiento para comprender más y de mejor manera los fenómenos que tienen lugar tanto dentro de este mundo como fuera de él.

Actualmente, se dice que todo cambia rápidamente y que estos cambios se deben en parte a que el descubrimiento, de índole científico, se ha convertido en un dominio legítimo de información. La inspiración y la intuición que parecían guiarlo ya dejaron de ser misterios eternos y ahora se saben motivados por el reconocimiento de patrones en montañas de datos de diversos tipos, que ya rebasan las capacidades actuales y donde se parte generalmente de búsquedas en grandes laberintos de posibilidades cuando se emprende la ardua tarea de encontrar información relevante. Ahora, el reconocimiento y la búsqueda selectiva, son las herramientas utilizadas para el descubrimiento y se han adquirido a lo largo de muchos años de observar y aprender acerca de nuestro mundo.

Existe otra razón por la que se tiene este renovado interés por el descubrimiento. La tecnología de la información y de las bases de datos han permitido la acumulación de grandes cantidades de datos; así como también facilitado el acceso a los mismos, situación que en años anteriores no era posible. Esto ha generado una necesidad urgente de nuevas teorías y herramientas de análisis computacional que auxilien en el aprovechamiento eficiente de estos recursos. Estas teorías y herramientas son materia de estudio del campo *Descubrimiento de Conocimiento en Bases de Datos* (DCBD), dicha frase fue propuesta en el primer taller *Knowledge Discovery of Databases* (KDD) realizado en 1989 [9] para enfatizar que el conocimiento es el producto final del descubrimiento *guiado por los datos*.

Se ha mostrado que las computadoras, que han facilitado la acumulación de estas vastas cantidades de datos, y las herramientas de análisis actualmente disponibles, pueden colaborar en la tarea de extraer información útil a partir de estas montañas de datos. Si se analiza específicamente la tarea de la minería de datos, se observa que la necesidad más grande no es ampliar nuestras bases de conocimiento, sin embargo esto es deseable, sino ser capaces de acceder a los datos para extraer el conocimiento que se necesita, ya sea que se tenga idea de que éste exista o no. Debido a que la capacidad de atención de los seres humanos está estrictamente limitada, los motores de búsqueda deben ser más inteligentes, de modo que puedan seleccionar y filtrar, de todo el conjunto, aquellos elementos de interés que se requieren. Deben no sólo ser capaces de responder a solicitudes específicas, sino usar también un conocimiento más extenso de las necesidades para recuperar información importante que no se ha solicitado y, que de existir, podría resultar sorprendente [32].

Las facilidades de captación y acumulación de grandes conjuntos de datos se han extendido a diversos sectores. Es decir, ya no es sólo el sector comercial el que está en posibilidades de acumularlos y sacar provecho de los mismos, cada

día es más claro que existe una expansión gradual en este sentido hacia otros ámbitos, como es el caso de algunas dependencias gubernamentales, instituciones educativas o de evaluación externa independientes, por mencionar algunos. En el último caso, existe una necesidad creciente de procesar las vastas cantidades de datos almacenadas, esto para dar respuesta a las nuevas interrogantes que han surgido en torno a la educación y contribuir de esta manera a lograr cambios importantes en su beneficio.

Una de las líneas de investigación del ámbito educativo sobre las que se han realizado un sinnúmero de investigaciones a la fecha, emerge a partir de los años 60's, en Estados Unidos, donde se da un movimiento de rendición de cuentas, promotor de la elaboración de informes nacionales acerca del sistema educativo. Uno de los primeros trabajos de gran impacto fue el Informe Coleman [4], que investigó la eficiencia de las escuelas. En sus conclusiones se señala que los resultados escolares se debían básicamente al origen social del alumno, y que la escuela tenía escasa influencia. Con estos resultados, parecía que la educación no constituía un mecanismo de nivelación social, sino que, por el contrario, contribuía a conservar las diferencias.

Por otro lado, algunos analistas critican el modelo utilizado en los estudios referidos, por ser de "caja negra". Se advierte que, si bien la clase social puede constituir un predictor importante del desempeño de los alumnos, esto no anula la gran influencia que puede tener la escuela: si se ponderan las condiciones de partida de los alumnos y las situaciones del contexto, se podrán apreciar mejor las diferencias debidas a la escolaridad. Bien puede ocurrir que un resultado inferior llegue a ser, paradójicamente, mejor que otro superior, si hubo una mayor contribución a la formación de los estudiantes: en términos relativos, una escuela que recibe alumnos en condiciones precarias puede aportar más beneficios educativos (valor añadido) que otra escuela que recibe estudiantes que gozan de condiciones favorables y ya dominan mucho de lo evaluado desde un inicio. En esto radica el interés por evaluar la eficacia en el logro escolar

En México, investigaciones de esta índole son incipientes y se reconoce que con los datos generados, por ejemplo, de las pruebas estandarizadas que evalúan un currículo específico, se está en posibilidades de iniciar este tipo de estudios con la finalidad de que arrojen información relevante en torno a la situación del sistema educativo mexicano, lo que indudablemente podría conducir a la elaboración de políticas de mejoramiento para elevar su calidad.

Considerando lo mencionado en párrafos anteriores y el desarrollo que se ha dado en el campo de Descubrimiento de Conocimiento en Bases de Datos, tanto a nivel teórico como práctico; en este trabajo se realiza un proceso de descubrimiento de conocimiento que, a partir de una base de datos relacionada con los resultados de una prueba estandarizada para complementar los criterios de ingreso a la Educación Media Superior (EMS) y del cuestionario de datos personales y socioeconómicos que deben requisitar los aspirantes, se enfoque en la obtención de un modelo que explique parte de las diferencias del desempeño en la prueba entre las escuelas secundarias de origen, incorporando predictores relevantes a nivel individual, esto con la finalidad de realizar comparaciones en igualdad de condiciones entre las mismas.

La producción de resultados útiles requiere de la definición clara del problema y de las metas del descubrimiento formuladas desde el inicio, junto con

planes de desarrollo. Para garantizar esto, se utiliza un modelo de procesos estándar que asegure la preparación correcta y proporcione un lenguaje común para la comunicación de métodos y resultados. Esta metodología ha sido denominada CRISP-DM y la selección de la misma se debe principalmente a que proporciona una visión muy clara del proceso de descubrimiento; facilitando con ello la planeación, documentación y comunicación para este tipo de proyectos. Esta metodología fue propuesta a finales del año 1996 por un consorcio de compañías Europeas para servir como modelo estándar de procesos, no propietario, para la minería de datos [34].

Con respecto a la organización de este trabajo, los contenidos se desarrollan básicamente en cuatro capítulos. En el capítulo 1, se presenta un recorrido sobre los temas que se consideran más importantes sobre el proceso de Descubrimiento de Conocimiento. En primera instancia, se describe a grandes rasgos cómo surge este campo, proporcionando una definición del proceso, que si bien no todos la aceptan como adecuada para describirlo, una abrumadora mayoría de estudiosos del tema la enuncia en sus trabajos para dar cuenta del mismo. Posteriormente, se describen cada una de las fases que comprende el proceso y se hacen algunas aclaraciones sobre las confusiones frecuentes que surgen al emplear los términos Minería de Datos y Descubrimiento de Conocimiento como sinónimos. Asimismo, se describe una propuesta de arquitectura para sistemas de Descubrimiento de Conocimiento en Bases de Datos y la función de cada uno de sus componentes, haciendo siempre énfasis en que la Minería de Datos es una etapa del proceso completo. De manera semejante, se considera importante presentar el análisis de la ruta evolutiva que ha seguido el proceso de Descubrimiento de Conocimiento y las dudas que surgen sobre si realmente se estará haciendo ciencia bajo este enfoque de análisis de datos, por ello se incluyen dos secciones que dan un panorama muy general sobre estas inquietudes. Por supuesto esta visión de contenidos importantes resultaría incompleta si no se incluyeran los retos que se vislumbran en este campo.

El capítulo 2, tiene como objetivo mostrar los esfuerzos que se han hecho en el desarrollo de modelos de procesos estandarizados para la tarea de descubrimiento de conocimiento y que podrían contribuir a lograr que el mismo deje de percibirse como una práctica demasiado especializada, por lo que se presenta la descripción de dos metodologías.

El capítulo 3, desarrolla la aplicación del proceso de descubrimiento de conocimiento, específico para la base de datos de aspirantes a EMS, en términos de la metodología CRISP-DM y considerando únicamente las tres primeras fases correspondientes al análisis previo de los datos. En primer lugar, se inicia con la comprensión del problema, lo que permite determinar los objetivos que se pretenden alcanzar en el proyecto, así como también evaluar la situación actual del problema y cerrar esta fase con un plan del proyecto. Posteriormente, con la fase de comprensión de los datos, se obtienen las primeras impresiones sobre los mismos y se vislumbra su importancia para las etapas siguientes. El siguiente apartado, correspondiente a la preparación o pre-procesamiento de los datos, es muy importante debido a que facilita la siguiente etapa, aquí se describe el tratamiento que se realiza para cada variable con miras a la realización del proceso de Minería de Datos.

En el capítulo 4, se presentan las tres últimas fases, incluyendo la fase de Minería de Datos, lo que incluye una descripción de la selección que se hizo de la técnica de modelación, del diseño de prueba de resultados adoptado, así como también de la construcción y la evaluación del modelo. Finalmente, se presenta una fase de evaluación y presentación, que da cuenta de los resultados obtenidos, en términos del contexto del problema, y una revisión del proceso de descubrimiento de conocimiento realizado.

Para concluir, se presentan dos glosarios, uno correspondiente a los términos utilizados en el proceso de descubrimiento y otro relacionado con los términos de la investigación educativa que se han utilizado en este trabajo, concretamente sobre los estudios de eficacia escolar. La finalidad de incluir los glosarios es facilitar la comunicación entre los expertos en el dominio y en la minería de datos. Aunado a lo anterior, se incluyen tres anexos: el anexo A contiene una porción relevante de la exploración, en términos de la distribución de frecuencias, de las variables más relevantes. El anexo B presenta los indicadores estadísticos más importantes calculados para cada una de las categorías de las variables de interés. Finalmente, el anexo C contiene una descripción de la estructura de la base de datos original.

Capítulo 1. Descubrimiento de Conocimiento en Bases de Datos

1.1Antecedentes

El descubrimiento de conocimiento en bases de datos es un intento por resolver un problema de la era de la información digital: la sobrecarga de datos [9]. Desde este punto de vista, los esfuerzos están encaminados a la automatización, por lo menos parcialmente, del trabajo de análisis de grandes conjuntos de datos, debido a que la tarea de procesarlos manualmente para descubrir patrones y estructuras significativas resulta lenta, cara y altamente subjetiva. En consecuencia, es justo recurrir al apoyo de aquello que hizo posible este crecimiento explosivo en la acumulación de datos: la tecnología computacional.

En un nivel abstracto, el proceso de Descubrimiento de Conocimiento en Bases de Datos (DCBD) se enfoca en el desarrollo de métodos y técnicas para extraer conocimiento de grandes volúmenes de datos. El problema básico del proceso DCBD es el de establecer una correspondencia entre los datos *crudos* en otras formas que podrían ser más compactas, como un reporte breve; más abstractas, como una aproximación descriptiva; o todavía más útil, un modelo predictivo.

1.2 Definición

Existen varias definiciones informales de DCBD que contrastan en diferentes campos. En estadística, DCBD se introdujo como análisis exploratorio automatizado por computadora de conjuntos grandes de datos complejos [14] o análisis secundario de grandes conjuntos de datos [24]. El término "secundario" significa que el propósito principal de la recolección de datos no es el descubrimiento de conocimiento. Desde la perspectiva de las bases de datos, DCBD se caracterizó como consulta de patrones, lo que enfatiza un acoplamiento cercano de búsqueda de patrones con las consultas ejecutadas por los Sistemas Administradores de Bases de Datos (SABD). En aprendizaje máquina, se hizo popular tratar DCBD como una versión del aprendizaje máquina que se aplica a conjuntos de datos grandes y a un espectro más amplio de tareas y métodos no supervisados; así como también, el típico concepto de aprendizaje supervisado a partir de ejemplos [32].

La definición que se cita con frecuencia es: "DCBD es el proceso no trivial de identificación de patrones válidos, novedosos, potencialmente útiles y finalmente comprensibles a partir de los datos" [9]. Donde el término *proceso* implica que DCBD está conformado de varias etapas (como se verá más adelante) y lo *no trivial* significa que está involucrada alguna búsqueda o inferencia, es decir, que no se trata de un cálculo directo de cantidades predefinidas como lo sería, por ejemplo, la obtención de un promedio de algún conjunto de datos. Los *patrones* son expresiones, en algún lenguaje específico, que describen un subconjunto de los datos o un modelo. En este contexto, la extracción de un patrón también incluye ajustar un modelo o encontrar estructura en los datos; o, en general, hacer cualquier descripción de alto nivel de un conjunto de datos.

Los patrones descubiertos deben ser *válidos,* con algún grado de certidumbre, en nuevos datos. Es deseable que los patrones sean *novedosos*

(preferiblemente para el usuario) y *potencialmente útiles*, es decir, que lleven cierto beneficio para el usuario o tarea. Los patrones deben ser *comprensibles*, si no lo fuesen inmediatamente, entonces podrían serlo después de un paso adicional de procesamiento. Finalmente, y no por ello menos importante, ya que representan la materia prima para el proceso, los datos son un conjunto de hechos (casos en una base de datos).

La discusión anterior implica que se pueden definir medidas de evaluación para determinar la relevancia de los patrones extraídos. En muchos casos es posible definir medidas de certidumbre (por ejemplo, la exactitud de la predicción estimada en nuevos datos) o la utilidad (por ejemplo, las ganancias en términos monetarios, que se puedan obtener debido a la realización de mejores predicciones o a la velocidad en tiempo de respuesta de un sistema). Las nociones relacionadas con la novedad y el nivel de comprensión son mucho más subjetivas. Una noción importante, llamada *relevancia* (interestingness) con frecuencia se toma como una medida del valor global del patrón, que combina validez, novedad, utilidad y simplicidad. Las funciones de *relevancia* pueden definirse explícitamente o pueden enunciarse implícitamente a través de un orden impuesto por el sistema DCBD en los patrones o modelos descubiertos.

Con base en las nociones descritas anteriormente, es posible considerar que un patrón es conocimiento si excede algún umbral de relevancia, lo que no necesariamente implica un intento de definir conocimiento desde el punto de vista filosófico o incluso popular. En este caso, lo que se considera conocimiento está únicamente controlado por el usuario; así como también por el dominio específico en que se esté trabajando y se determina por las funciones y umbrales que el mismo usuario establece.

A pesar de que la definición anterior es citada con mucha frecuencia, en [32] los autores manifiestan que esta definición es útil para vender herramientas y servicios DCBD con la promesa que se hace generalmente a los clientes de que obtendrán resultados válidos, novedosos, útiles y comprensibles, pero no describe un dominio intelectual sistemático de actividades. Mientras ninguna definición es completamente adecuada para cualquier campo de actividad humana, argumentan una solución simple. El descubrimiento de conocimiento en bases de datos es exactamente lo que dice bajo el significado normal. No se requiere parafrasear. El conocimiento es cualquier verdad articulada y justificada con respecto a un dominio, representado en algún lenguaje formal.

El descubrimiento de conocimiento produce enunciados que describen objetos del mundo real, conceptos y regularidades. Estos enunciados se derivan de un proceso de generación autónoma y de la verificación de nuevas hipótesis. Las bases de datos son repositorios de datos bien estructurados y mantenidos, que están relacionados con dominios del mundo real. DCBD es más que análisis de datos y más que detección de patrones en los datos.

1.3 Descripción del proceso DCBD

La gran mayoría de los estudiosos del tema coinciden en que el proceso de descubrimiento de conocimiento es interactivo e iterativo. Este proceso incluye varias etapas donde el usuario debe tomar muchas decisiones. A continuación, se describen las diferentes etapas del proceso DCBD [9]:

1. La comprensión del dominio de aplicación y del conocimiento previo relevante; así como la identificación del objetivo del proceso DCBD desde el punto de vista del cliente.
2. La creación de un conjunto de datos, objeto de análisis: la selección de un conjunto de datos o de un subconjunto de variables o muestras de datos, sobre el que se desea realizar el descubrimiento.
3. La limpieza y el pre-procesamiento de los datos. Entre las operaciones básicas se incluyen: la eliminación del ruido, la colección de la información necesaria para modelar o explicar el ruido, la decisión sobre las estrategias que se utilizarán para la manipulación de los datos ausentes, etc.
4. La reducción y la proyección de los datos: consiste en encontrar las características útiles para representar los datos dependiendo de los objetivos de la tarea. Con los métodos de reducción de dimensionalidad o de transformación, se puede reducir el número de variables bajo consideración o pueden encontrarse representaciones invariantes de los datos.
5. La correspondencia de las metas del proceso DCBD con algún método de Minería de Datos (MD). Por ejemplo, clasificación, regresión, agrupación, etc.
6. El análisis exploratorio, el modelo y la selección de las hipótesis: la selección de algoritmo(s) de minería y del método que se utilizará para la búsqueda de patrones en los datos. Este proceso incluye la decisión de qué modelos y parámetros podrían ser apropiados (por ejemplo, los modelos de datos categóricos son diferentes de los modelos de vectores de reales) y la correspondencia de un método de MD específico con el criterio global del proceso DCBD (por ejemplo, el usuario final podría interesarse más en la comprensión del modelo que en sus capacidades de predicción).
7. La MD es la tarea de búsqueda de patrones de interés en una forma representacional o un conjunto de representaciones, que incluyen las reglas o árboles de clasificación, la regresión y la agrupación.
8. La interpretación de los patrones extraídos: en este paso es posible el regreso a alguno de los pasos anteriores (1 a 7) para realizar otras iteraciones. Este paso puede incluir también la visualización de los patrones extraídos y de los modelos; o la visualización de los datos dados los modelos extraídos.
9. La puesta en práctica del conocimiento descubierto: uso del conocimiento directamente, ya sea incorporándolo en otro sistema para realizar otras acciones o simplemente documentándolo y reportándolo a las personas interesadas. Este proceso también incluye la verificación y la resolución de conflictos potenciales con el conocimiento previo.

El proceso DCBD puede ser altamente iterativo y contener ciclos entre cualesquiera dos pasos de los mencionados anteriormente. El flujo básico de pasos (aunque no la potencial multiplicidad de iteraciones y ciclos) se ilustra en la figura 1.1, extraída de [26]. La mayoría del trabajo se ha enfocado en el paso 7, la minería de datos. Sin embargo, los pasos restantes son muy importantes para

la aplicación exitosa de DCBD en la práctica. Lo anterior, es por el hecho de que precisamente se hace mucho énfasis en la MD y poco en otras etapas a las que se destina generalmente la mayor cantidad de tiempo del proceso y que son indispensables para garantizar la calidad de los datos y facilitar el proceso de minería. Como se mencionó anteriormente, una de estas etapas es la de pre-procesamiento, que contempla: la limpieza, la integración, la transformación y la reducción de los datos; así como también una última etapa, que es la de evaluación y presentación de los resultados de la minería.

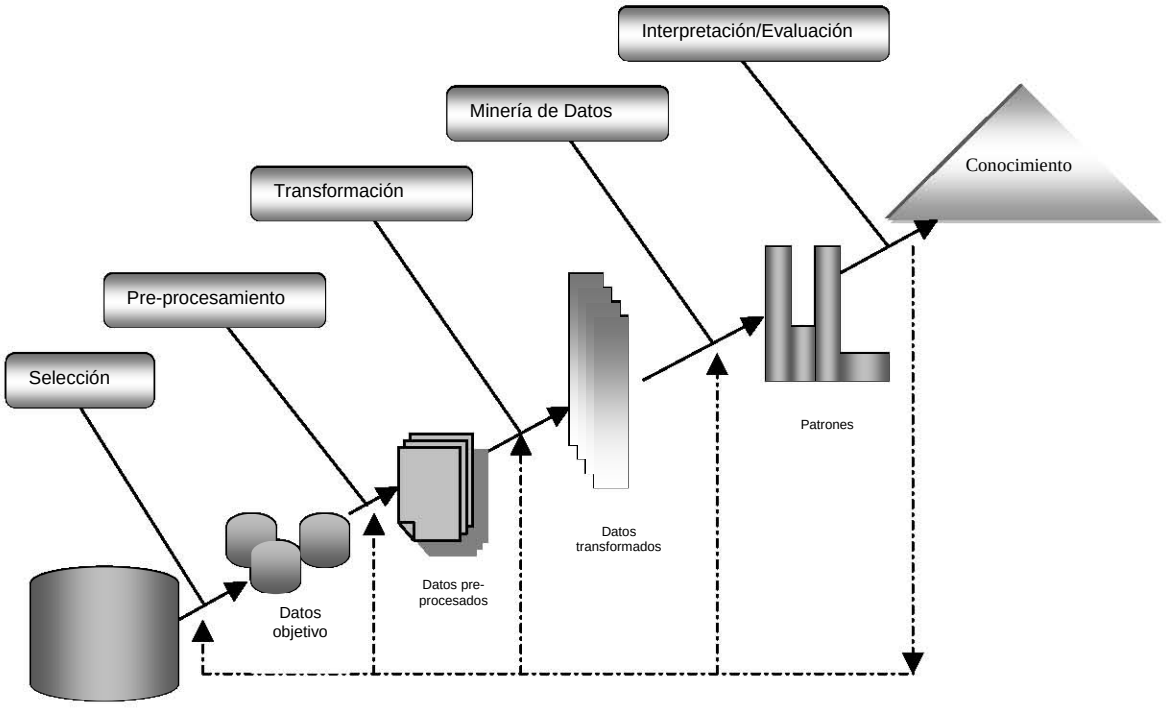

Figura 1.1. Etapas que conforman el proceso DCBD

1.4 Arquitectura de un Sistema de DCBD

La figura 1.2 muestra los pasos involucrados en el proceso de descubrimiento de conocimiento y cómo la MD podría interactuar con el usuario o una base de conocimiento. Los patrones interesantes se le presentan al usuario y podrían ser almacenados como nuevo conocimiento en la base de conocimientos. De acuerdo a esta visión, la MD es sólo un paso en el proceso completo, esencial por cierto, ya que apoyada por la gran variedad de técnicas de análisis, descubre patrones ocultos que estarán sujetos a la evaluación correspondiente para determinar su importancia práctica dentro del contexto del problema.

Entre la mayoría de los expertos del tema está claro que la MD es sólo un paso en el proceso de descubrimiento de conocimiento. Sin embargo, en la industria y en algunas comunidades de investigación como la estadística y la de bases de datos, por mencionar algunas, el término más popular es el de minería de datos. En consecuencia, se considera necesario aclarar que en este trabajo se utilizará el término MD para hacer referencia sólo a una de las etapas del proceso de descubrimiento de conocimiento.

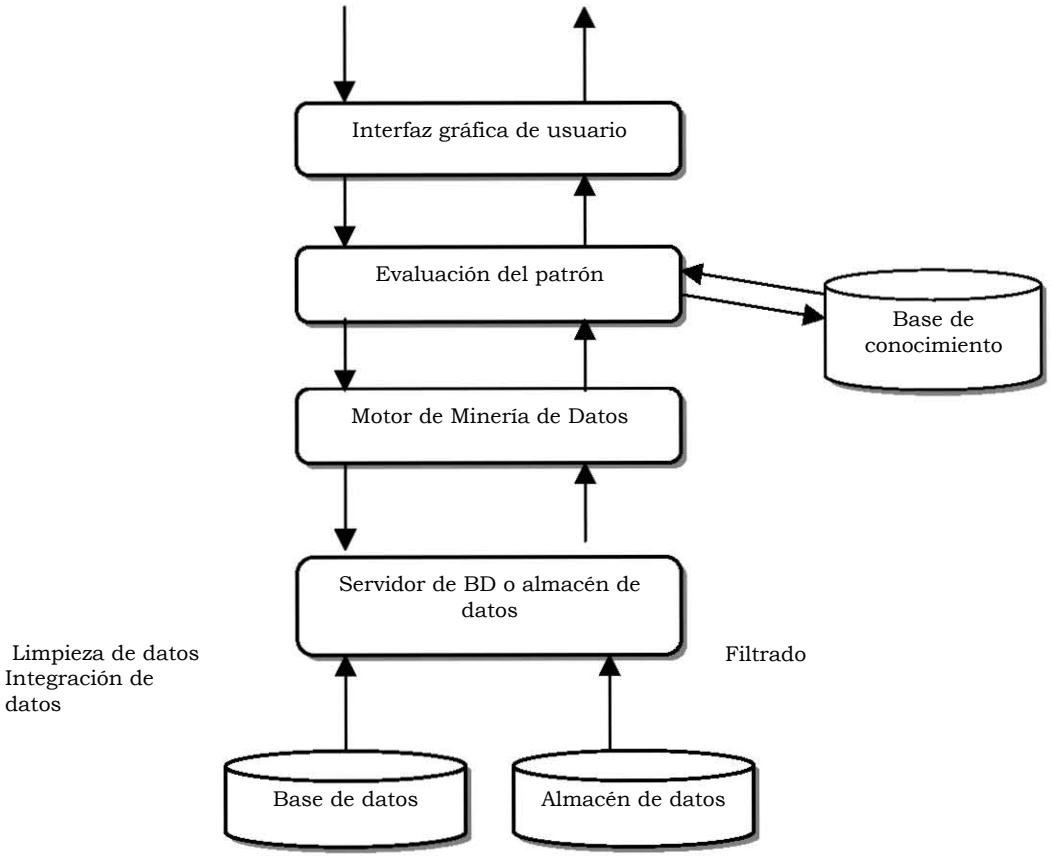

Figura 1.2. Arquitectura de un sistema típico de DCBD

Desde esta perspectiva, la arquitectura de un sistema común de descubrimiento de conocimiento podría tener los siguientes componentes principales [26]:

Base de datos, almacén de datos u otro repositorio de datos

Es un conjunto de bases de datos, almacenes de datos, hojas electrónicas u otros tipos de repositorios de datos. Podrían desarrollarse técnicas de limpieza e integración de los datos.

Servidor de bases de datos o almacén de datos

El servidor de bases de datos o almacén de datos es responsable de cargar (o disponer) los datos importantes, basados en la solicitud del usuario.

Base de conocimiento

Es el conocimiento del dominio que se utiliza para guiar la búsqueda o evaluar qué tan interesantes son los patrones que se obtengan. Este conocimiento puede incluir jerarquías de conceptos, usadas para organizar atributos o valores de atributos en diferentes niveles de abstracción. Además, podría incluirse el conocimiento representado por la creencia del usuario, que también puede usarse para evaluar la importancia de un patrón inesperado. Otros ejemplos del

conocimiento del dominio son restricciones de importancia adicionales o umbrales y meta-datos (que describen datos de fuentes heterogéneas múltiples).

Motor de Minería de datos

Es esencial para el sistema de minería de datos e idealmente consiste de un conjunto de módulos funcionales para tareas relacionadas con la caracterización, asociación, clasificación, análisis cluster, análisis de evolución y desviación.

Módulo de evaluación de patrones

Este componente emplea comúnmente medidas de importancia e interactúa con los módulos de MD y enfoca la búsqueda hacia patrones interesantes. Podría usar umbrales de importancia para filtrar los patrones descubiertos. De manera alternativa, el módulo de evaluación de patrones se puede integrar con el módulo de minería, dependiendo de la implementación del método de minería de datos usado. Para la minería eficiente, se recomienda introducir la evaluación de la importancia del patrón en el proceso de minería de modo que se enfoque en la búsqueda de patrones interesantes.

Interfaz gráfica de usuario

Este módulo comunica al usuario con el sistema de MD, al permitir que el usuario interactúe con el sistema y especifique una consulta o tarea proporcionando información para enfocar la búsqueda y que realice minería exploratoria basada en los resultados intermedios. Además, este componente permite al usuario mostrar esquemas de bases de datos, almacenes de datos o estructuras de datos, evaluar patrones extraídos y visualizarlos en diferentes formas.

1.5 Minería de Datos: una etapa del proceso DCBD

La minería de datos, como una etapa del proceso de descubrimiento de conocimiento global, contempla la aplicación iterativa y repetida de métodos de minería específicos. El objetivo de esta sección es presentar un resumen de las metas principales de la MD; una descripción breve de los algoritmos que incorporan los métodos de la MD y mencionar algunos métodos o técnicas de MD, descritos en términos de sus componentes algorítmicos.

Las metas principales de la MD

Se pueden distinguir dos tipos de metas: (1) la *confirmación* y (2) el *descubrimiento*. En la *confirmación*, el sistema se limita a verificar las hipótesis del usuario. Con el *descubrimiento*, el sistema de manera autónoma encuentra nuevos patrones. Además, podemos subdividir la meta de descubrimiento en *predicción*, donde el sistema localiza patrones para predecir el comportamiento futuro de algunas entidades y la *descripción*, donde el sistema obtiene patrones que presenta al usuario en una forma comprensible.

La MD incluye también el ajuste de modelos o la determinación de patrones a partir de los datos observados. Estos modelos representan entonces el conocimiento inferido: si los modelos reflejan o no conocimiento útil o interesante, es parte del todo, por esta razón se requiere que el proceso DCBD sea interactivo y que el juicio humano esté siempre presente para determinar la importancia práctica de los resultados.

Para el caso del ajuste de modelos, se utilizan dos formalismos matemáticos primarios: (1) el estadístico y (2) el lógico. El enfoque estadístico permite efectos no determinísticos en el modelo, mientras que un modelo lógico es puramente determinístico. El enfoque estadístico se usa más ampliamente para las aplicaciones prácticas de MD dada la presencia típica de incertidumbre en los procesos del mundo real que generan los datos.

Los componentes de los algoritmos de MD

De acuerdo con [9] se pueden identificar tres componentes básicos en cualquier algoritmo de MD: 1) La *representación del modelo*, 2) La *evaluación del modelo*, y 3) la *búsqueda del modelo*.

Esta visión reduccionista no necesariamente es completa; en lugar de ello, más bien es una forma conveniente de expresar los conceptos clave de los algoritmos de MD en una forma unificada y compacta.

La *representación del modelo* es el lenguaje usado para describir los patrones descubiertos. Si la representación es demasiado limitada, entonces ninguna cantidad de tiempo de entrenamiento o ejemplos podrán producir un modelo preciso para los datos. Es importante que un analista de datos comprenda completamente los supuestos de representación que podrían ser inherentes a un método específico. Es igualmente importante que un diseñador de algoritmos enuncie claramente qué supuestos de representación se hacen en un algoritmo particular.

Los criterios de *evaluación del modelo* son declaraciones de qué tan bien cumple un patrón particular las metas del proceso de descubrimiento de conocimiento. Por ejemplo, los modelos predictivos son juzgados por la exactitud de la predicción empírica sobre algún conjunto de prueba. Los modelos descriptivos pueden evaluarse a lo largo de las dimensiones de la exactitud predictiva, novedad, utilidad y comprensión del modelo ajustado.

El *método de búsqueda* consiste de dos componentes: 1) búsqueda de parámetros y 2) búsqueda del modelo. Una vez que se especifica la representación del modelo (o familia de representaciones) y el criterio de evaluación, entonces el problema de MD se reduce a una tarea de optimización, es decir, a encontrar los parámetros y modelos, a partir de la familia seleccionada, que optimicen el criterio de evaluación. En la búsqueda de parámetros, el algoritmo debe buscar los parámetros que optimicen el criterio de evaluación dado los datos observados y el modelo de representación establecido. La búsqueda del modelo se presenta entonces como un ciclo del método de búsqueda de parámetros.

Algunos métodos de MD descritos en términos de sus componentes algorítmicos

Existe una amplia variedad de métodos de minería de datos. Aquí sólo se presentará un subconjunto de técnicas populares y se discutirán en términos de los tres elementos mencionados anteriormente: representación, evaluación y búsqueda [9].

Métodos de Regresión no lineal y de Clasificación

Estos métodos consisten de una familia de técnicas para la predicción que ajustan combinaciones lineales y no lineales de funciones base (sigmoides, splines, polinomiales) para las combinaciones de las variables de entrada. Ejemplos incluyen las redes neuronales y métodos spline, por mencionar algunos. En términos de la evaluación del modelo, aunque las redes del tamaño apropiado se pueden aproximar universalmente a cualquier función suavizada para cualquier grado de exactitud deseada, relativamente se conoce poco con respecto a las propiedades de representación de las redes de tamaño fijo estimadas a partir de conjuntos de datos finitos. Además, el error cuadrático estándar y de funciones de perdida cross-entropy usadas para entrenar las redes neuronales pueden verse como funciones log-verosimilitud para la regresión y la clasificación, respectivamente. La retropropagación es un método de búsqueda de parámetros que realiza el gradiente descendente en el espacio de parámetros (pesos) para encontrar un máximo local de la función de verosimilitud a partir de condiciones iniciales aleatorias. Los métodos de regresión no lineal, aunque poderosos en el poder de representación, pueden ser difíciles de interpretar.

Métodos basados en el ejemplo

La representación es simple: utiliza ejemplos representativos de la base de datos para aproximar un modelo; es decir, las predicciones sobre nuevos ejemplos se derivan de las propiedades de casos similares en el modelo cuya predicción se conoce. Entre estas técnicas se incluyen los algoritmos de clasificación del vecino más cercano y de regresión y sistemas de razonamiento basado en casos.

Una desventaja potencial de los métodos basados en el ejemplo (comparados con los métodos de árbol) es que se requiere de una métrica de distancia bien definida para la evaluación de la distancia entre los puntos. El modelo de evaluación está basado comúnmente en estimadores de validación cruzada de un error de predicción: los parámetros del modelo que se estima pueden incluir el número de vecinos a usar para la predicción y la métrica de distancia misma. Semejante a los métodos de regresión no lineal, los métodos basados en el ejemplo son con frecuencia asintóticamente poderosos en términos de las propiedades de aproximación pero pueden ser difíciles de interpretar porque el modelo está implícito en los datos y no explícitamente formulado. Entre las técnicas relacionadas se incluye la estimación densidad kernel y la modelación mezclada.

Modelos de dependencia gráfica probabilística

Los modelos gráficos especifican dependencias probabilísticas usando una estructura gráfica. En su forma más simple, el modelo especifica qué variables son directamente dependientes de las otras. Comúnmente, estos modelos se utilizan con variables categóricas o discretas, pero las extensiones a casos especiales, como densidades Gausianas, para variables reales son posibles también. Dentro de las comunidades de Inteligencia Artificial y estadísticas, estos modelos fueron inicialmente desarrollados dentro del marco de trabajo de sistemas expertos probabilísticos; la estructura del modelo y los parámetros (las probabilidades condicionales atadas a las ligas de la gráfica) fueron obtenidas de expertos. Existen avances significativos en ambas comunidades sobre métodos en los que la estructura y los parámetros de los modelos gráficos pueden ser aprendidos directamente de las bases de datos. El criterio de evaluación del modelo es comúnmente Bayesiano en la forma y la estimación de parámetros puede ser una mezcla de estimadores de forma cerrada y métodos iterativos dependiendo de si una variable es directamente observada u oculta. La búsqueda del modelo puede consistir de métodos Greedy hill-climbing sobre varias estructuras gráficas. El conocimiento previo, así como el orden parcial de las variables basado en relaciones causales, puede ser de utilidad en términos de la reducción del espacio de búsqueda del modelo. Aunque todavía, principalmente en la fase de investigación, los métodos de inducción del modelo gráfico son de interés particular para DCBD porque la forma gráfica del modelo conduce por sí misma fácilmente a la interpretación humana.

Modelos de aprendizaje relacional

Aunque los árboles de decisión y las reglas tienen una representación restringida para la lógica proposicional, el aprendizaje relacional (conocido también como inductive logic programming) utiliza el lenguaje de patrón más flexible de lógica de primer orden. Un aprendiz relacional puede fácilmente encontrar fórmulas como X=Y. A la fecha, la mayoría de la investigación sobre los métodos de evaluación de modelos para aprendizaje relacional es lógico en naturaleza. El poder extra representacional de los modelos relacionales viene del costo de las demandas computacionales significativas en términos de la búsqueda.

1.6 Cronología del DCBD

Históricamente, a la noción de encontrar patrones útiles en los datos se le ha dado una gran variedad de nombres, que incluyen: minería de datos, extracción de conocimiento, cosecha de información, arqueología de los datos y procesamiento de patrones de datos. El término minería de datos se ha usado principalmente por las comunidades estadísticas, analistas de datos y sistemas de administración de información. También ganó popularidad en el campo de las bases de datos. El término DCBD se ha popularizado más en los campos de la Inteligencia Artificial y de Aprendizaje Máquina [9].

En algunos textos sobre el tema, se plantea al DCBD como la evolución natural de la tecnología de las bases de datos [26]. En años recientes, la razón principal de la gran cantidad de atención que se le ha dado se debe a la amplia

disponibilidad de grandes cantidades de datos y a la inminente necesidad de convertirlos en información útil o en conocimiento.

Este campo se puede ver también como resultado de la evolución natural de la tecnología de la información. A partir de la figura 1.3, se puede observar que la industria de las bases de datos ha seguido una ruta evolutiva en el desarrollo de las siguientes funcionalidades: colección y creación de bases de datos, administración de datos (incluyendo almacenamiento y recuperación, procesamiento de transacciones) y el análisis y la comprensión de los datos (incluye los almacenes de datos y la Minería de Datos).

Existe consenso en que el próximo avance vendrá en soluciones integradas que permitan a los usuarios finales explorar sus datos usando también metáforas gráficas –la meta es unificar los algoritmos de MD e interfaces humanas visuales. Esto logrará que los usuarios entren en contacto con sus datos y hará la MD y el DCBD disponible a todo el mundo [21].

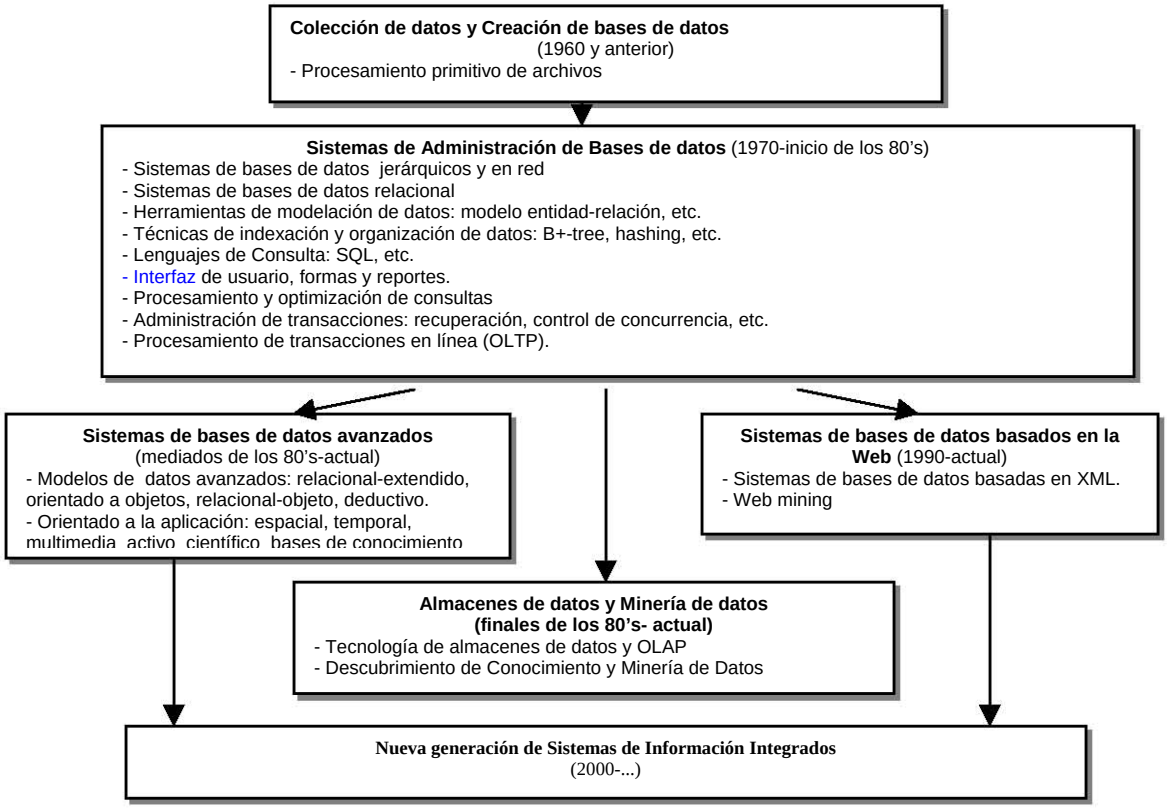

Figura 1.3. Cronología del desarrollo del DCBD y la MD

Herramientas de software para DCBD

Debido a que las aplicaciones DCBD se incrementan y difunden rápidamente, el mercado de herramientas avanzadas DCBD es dinámico y aparecen regularmente nuevos sistemas y versiones. Es difícil para cualquier usuario interesado conservar un registro actualizado del estado del arte en el desarrollo de herramientas. Una de las fuentes de información relacionada con herramientas

DCBD y que se actualiza regularmente es la sección de software de DCBD nuggets (http://www.kdnuggets.com/software/index.htm). Este sitio cuenta con descripciones de herramientas para las tareas principales de minería, como las redes Bayesianas y de Dependencia, Clasificación, Agrupación y Detección de desviación, pero también incluye pre-procesamiento, visualización y categorías de aplicación. Adicionalmente, hay una opción para bajar software que permite el acceso a sistemas comerciales o libres de costo; así como también versiones de prueba para su uso por tiempo limitado.

En el cuadro 1.1 se presenta un ejemplo de la selección de sistemas realizada por Gartner Group y que se estimaba probablemente conducirían el mercado de la minería de datos [32]. Estos sistemas se seleccionaron también incluyendo el criterio de viabilidad de los vendedores y los productos, servicios, soporte y consultoría que proporcionan. Gartner Group esperaba que IBM, Oracle y SAS tuvieran una ventaja a largo plazo debido a su sociedad, tamaño y experiencia.

Sistemas de Minería de Datos
Angoss Knowledge Suite
Intelligent MIner
Darwin
Enterprise Miner
Mine Set
Clementine

Cuadro 1.1. Sistemas para la Minería de Datos

Los reportes especiales y principalmente comerciales proporcionan un resumen extenso y de evaluación de herramientas DCBD (Two Crows, Aberdeen Group, Gartner Group). Sin embargo, estos reportes también corren el riesgo de resultar obsoletos por lo mencionado en líneas anteriores.

Dimensiones para la comparación de herramientas DCBD

El énfasis de este apartado es introducir algunas dimensiones que se consideran básicas e importantes para comparar, resumir y evaluar las herramientas DCBD. Se pueden definir diferentes tipos de sistemas DCBD de acuerdo a dimensiones simples, como las presentadas en la cuadro 1.2: sistemas de soporte único versus sistemas de tareas de minería múltiples, minería de datos versus sistemas DCBD y sistemas de dominio específico versus sistemas genéricos.

Ejemplos de sistemas de tarea única son los sistemas de clasificación (C4.5), sistemas de agrupación (AutoClass) y sistemas de visualización de datos (Netmap). Los dos primeros sistemas ofrecen sólo un método de minería para la clasificación o la agrupación. Algunos sistemas como Clementine, Darwin y Recon, ofrecen una combinación típica de métodos para la tarea de clasificación, que incluyen los árboles de regresión, reglas, redes neuronales y del vecino más cercano. La siguiente dimensión diferencia entre sistemas de minería de datos y DCBD. Los sistemas DCBD son sistemas completamente integrados (o sistemas con interfaces a bases de datos bien establecidas, sistemas estadísticos y de visualización) que soportan el proceso DCBD completo, es decir, pre-procesamiento, minería de datos y post-procesamiento.

Los sistemas de dominio específico soportan un dominio especial. Por ejemplo, los sistemas CoverStory, Spotlight y Opportunity Explorer son herramientas especializadas para aplicaciones de investigación de mercados.

Aunado a las dimensiones de clasificación simples, resultan importantes otras, como la dimensión de entrada de datos, que incluye los tipos de datos que pueden procesarse en el sistema, la interfaz e integración con el Sistema de Administración de Bases de Datos (SABD) y las operaciones de pre-procesamiento de datos. Si una herramienta DCBD puede procesar sólo tablas rectangulares simples –o multirelacionales, espaciales, texto, audio o datos multimedia- puede ser decisivo para una aplicación.

Las herramientas DCBD pueden explotar también conocimiento del dominio. El uso de reglas descubiertas previamente como conocimiento del dominio pueden ser relevantes como entrada para procesos DCBD, por ejemplo para pre-procesar los datos.

La dimensión algorítmica de un sistema DCBD incluye los métodos ofrecidos por el sistema del vasto conjunto de métodos de minería (y pre o post-procesamiento). El aspecto algorítmico también se refiere a propiedades como la exactitud, escalabilidad, interpretabilidad, robustez, rapidez y la sofisticación estadística de algoritmos, incluyendo técnicas avanzadas como muestreo adaptativo, bagging o boosting.

La siguiente dimensión se refiere a la salida, es decir, los tipos de conocimiento que son generados por un sistema. Esta es función de los algoritmos y métodos para las tareas de minería (clasificación, agrupación, etc.) que estén disponibles. Una aspecto adicional de la salida es la presentación del conocimiento generado y el papel de la visualización para los resultados de la minería. Reporte de resultados, especialmente basados en la Web, es un rasgo crecientemente importante de un sistema de descubrimiento.

La dimensión del usuario incluye aspectos como el usuario al que está enfocado (sofisticación requerida del usuario, analista contra usuario del dominio), los roles del usuario, aconsejar al usuario sobre los siguientes pasos razonables, documentación de los pasos usados en un proceso DCBD especial, replicabilidad y automatización (programación visual, macros, asistentes, bitácoras, agendas, administración de experimentos), interfaz de usuario gráfica o visual intuitiva, posición del usuario en el ciclo de descubrimiento y el grado de autonomía del sistema. Estos son importantes para la usabilidad y el grado de automatización del sistema.

Como se enfatizó por Gartner Group, el soporte y la consultoría es una dimensión prácticamente importante para un sistema DCBD. Mientras se demanda una práctica de consultoría grande como beneficio para los vendedores destacados (clasificados de acuerdo a su penetración en el mercado), algunos desarrolladores más pequeños (ejemplos, DataSurveyor y Kepler) están proporcionando también servicios y soportes apropiados. Con frecuencia los ingresos principales de estas compañías especializadas no es por la vía de su software sino por los honorarios de sus consultorías. Las dimensiones tecnológicas se refieren a las plataformas, arquitecturas, integración con otros sistemas y la Web.

Tareas soportadas	Tareas de minería únicas (ejemplo, clasificación), únicamente un método (ejemplo, árbol de clasificación)	Tarea única, métodos múltiples (ejemplo, árboles, redes neuronales, vecino más cercano, reglas)	Tareas de minería múltiples (ejemplo, clasificación, agrupación)	
Pasos del proceso soportados	Solo Minería de datos	También pre y post-procesamiento.		
Dominios soportados	Sistema genérico	Dominio específico (ejemplo, marketing).		
Enfoque de integración de herramienta	Macro integración (plug-in)	Micro integración		
Capas arquitecturales	Administración de datos Refinamiento de búsqueda	Agregación de datos (estadística, consultas) Interacción con el usuario	Evaluación de patrón y modelo	Búsqueda

Cuadro 1.2. Tipos de sistemas DCBD

Arquitectura de las herramientas DCBD

Como se mencionó en la sección 1.4, en los sistemas DCBD se pueden identificar capas arquitecturales: una capa de administración de datos básica es responsable de almacenar y actualizar los datos proporcionados. Una capa de agregación de datos controla las consultas estadísticas sobre la capa de administración de datos básica. Esta capa puede estar incrustada en la capa de administración de datos, por ejemplo, usando SQL para las agregaciones de datos, o puede depender de operaciones especializadas de acceso a los datos. La tercera capa es la de evaluación de hipótesis o patrones, que incluye tipos y métodos de problemas estadísticos para su verificación y cálculos de calidad. La capa del patrón usa los servicios del nivel de agregación de datos y es usado por la capa de búsqueda. La capa de búsqueda incluye las estrategias de búsqueda y las operaciones para la generación de hipótesis. Finalmente, una capa de refinamiento opera en los resultados de búsqueda, incluyendo los métodos de eliminación de redundancia.

Se pueden distinguir dos enfoques de integración para la inclusión modular de métodos de minería en un sistema: macro y micro integración. La macro integración se basa en un marco de sistema que ofrece la administración general de datos, visualización, funciones de prueba estadística, como en el sistema Kepler, que está basado en una arquitectura plug-in. Esta integración plug-in permite que un algoritmo de minería sea incluido en el sistema sin tener que modificar la implementación existente del algoritmo. La integración está basada principalmente en una descripción de la interfaz entre el sistema y el algoritmo.

La micro integración es un enfoque altamente modular que se realiza, por ejemplo, en el sistema Data Surveyor. La arquitectura Data Surveyor (figura 1.4) permite una implementación modular eficiente de las estrategias de búsqueda. Cuando se incluye un nuevo algoritmo en el sistema, la estructura modular requerida por el sistema debe ser observada y el algoritmo tiene que volver a implementarse con frecuencia. Data Surveyor está basado en una representación innovadora orientada a objetos de un proceso de descubrimiento.

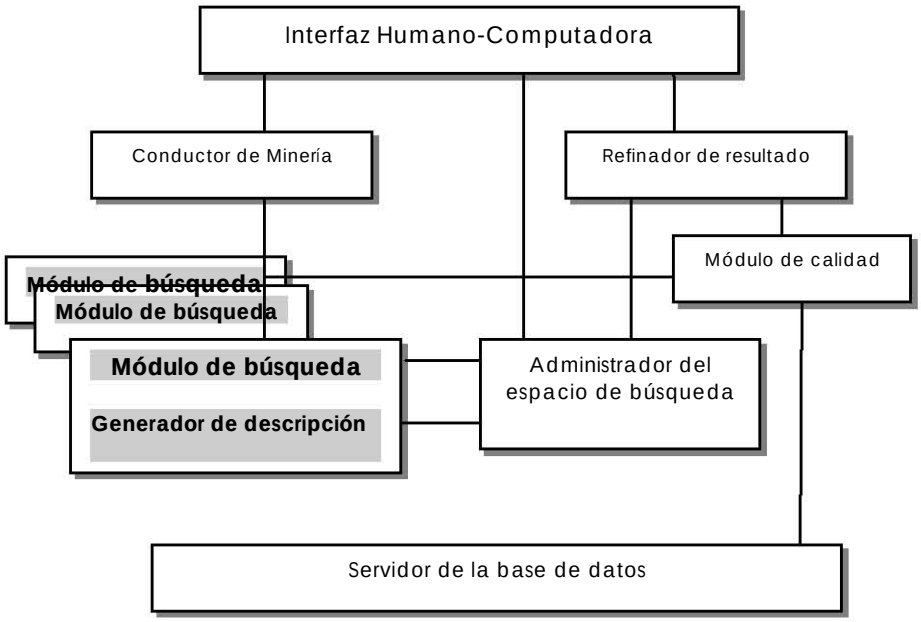

Figura 1.4. Estructura modular de los algoritmos de MD

A partir de la figura 1.4, se puede observar que el sistema Data Surveyor se adapta a la arquitectura general mencionada en la sección 1.4. Donde, en el núcleo de esta arquitectura existe un administrador del espacio de búsqueda realizada en una base de datos orientada a objetos que administra todas las hipótesis generadas. Los procesos operan de manera separada en este administrador del espacio de búsqueda: una interfaz de usuario permite dinámicamente la presentación del estado real de la búsqueda y la que va a ser re-direccionada. Los módulos de búsqueda llaman a un generador de hipótesis o descripción para expandir las hipótesis vía operadores genéticos o vecinos e insertar las nuevas hipótesis en el administrador del espacio de búsqueda. Un módulo de calidad prueba las hipótesis y calcula sus calidades usando una función de evaluación. Esta arquitectura depende en un modelo orientado a objetos para DCBD (descripción de proyectos, tareas, hipótesis, calidades, etc.) y un conductor procedimental de minería soportando un enfoque general de generar y probar, que iterativamente llama a los procesos anteriores.

Un tema arquitectural central es la integración de algoritmos de minería en sistemas de administración de bases de datos. Especialmente para enfoques de subgrupos de minería que requieren muchos procesos de consultas estadísticas, podría tener ventajas una implementación basada en SQL por la facilidad de desarrollo, portabilidad y posibilidades de paralelización automática.

La mayoría de los criterios tecnológicos básicos se refieren a plataformas en las que los sistemas están disponibles, que son principalmente las plataformas Unix o NT, pero también mainframes (Intelligent Miner from IBM). Las versiones de usuario único PC o Unix se están reemplazando cada vez más por arquitecturas cliente-servidor. El cuadro 1.3 resume los criterios.

Criterio	Características
Entrada Datos	Qué tipos de datos son soportados Qué tipo de opciones de transformación/pre-procesamiento
Conocimiento del dominio	Qué tipo de conocimiento del dominio son explotados
Salida Algoritmos	Qué tareas de minería son soportadas Qué métodos están disponibles para una tarea Propiedades algorítmicas (exactitud, robustez, escalabilidad, ...) Sofisticación algorítmica (bagging, boosting,...)
Presentación	Visualización Reportes
Usuario	Tipos y roles de usuario Guía, documentación de pasos, repetibilidad, automatización Lo intuitiva que resulte la interfaz
Tecnología	Plataformas Lenguaje de programación Integración con otros sistemas (especialmente sistemas de bases de datos) Capas arqueticturales y modularidad
Soporte	Documentación Servicio y consultoría Viabilidad del vendedor y producto.

Cuadro 1.3. Criterios para la comparación de herramientas DM

1.7 ¿El proceso DCBD es ciencia?

Para comprender la filosofía del enfoque metodológico utilizado en el DCBD se requiere proporcionar algún antecedente y el contexto. Aristóteles (384 a.C) y Bacon (1561-1626) defendieron un enfoque para la metodología científica que se utilizó aproximadamente por 2000 años. Sugirieron analizar grandes cantidades de datos, buscando patrones y después proponer hipótesis con respecto a estos patrones. Galileo (1564-1642) continuó la defensa de este enfoque pero sugirió además que los científicos debían también hacer experimentos para verificar sus hipótesis. El enfoque Galileano para la metodología científica se aceptó dentro de la comunidad científica aproximadamente por 300 años; incluso, era todavía de uso común durante todo el siglo XIX.

Sin embargo, en el siglo XX, se dio un cambio importante en la forma en que se practicaba el método científico. Se modificó algo del método científico Galileano; defendiéndose que la teoría debía ser postulada primero y que los datos experimentales debían ser colectados en un esfuerzo por apoyar esa teoría. Este enfoque ha sido denominado *ciencia confirmatoria*. Desde la perspectiva de la filosofía de la ciencia, DCBD sigue la tradición del método científico de Galileo. El DCBD regresa al método científico del siglo XIX en el que los datos generan la teoría; lo que puede ser peligroso. Cuando se trata de seguir este procedimiento, se sabe que el proceso puede estar repleto de riesgos, pero a pesar de todo esto, puede ser útil en muchas situaciones. Este último enfoque para la metodología científica, es algunas veces denominado, *ciencia exploratoria* [27].

El término MD, que se ha asumido como sinónimo de DCBD, ha tenido algunas interpretaciones despectivas. En [23] el autor expone que el término MD no es nuevo para la comunidad estadística y que se ha usado para describir el

proceso de hurgar en los datos con la esperanza de identificar patrones. Por supuesto que tiene una connotación despectiva, esto se debe a que una búsqueda exhaustiva seguramente resultará en patrones de algún tipo –por definición, los datos que no son uniformes tienen diferencias que pueden interpretarse como patrones. El problema es que muchos de estos patrones simplemente serán producto de fluctuaciones aleatorias fugaces y no representarán alguna estructura subyacente. Definitivamente, es claro que el objetivo del análisis de datos no es modelar estos patrones aleatorios fugaces, sino modelar las verdaderas estructuras subyacentes en los datos y que dan origen a patrones consistentes y replicables.

Para los estadísticos, entonces, el término MD inicialmente expresaba el sentido de la esperanza ingenua navegando vanamente en las realidades frías de la casualidad. Sin embargo, como lo expresan en [9], DCBD es una actividad legítima mientras se comprenda cómo hacerla correctamente; sin duda, se debe evitar llevarla a cabo pobremente, esto es, sin considerar los aspectos estadísticos del problema.

1.8 Retos del proceso DCBD

De acuerdo a [32] no hacen faltan más desafíos en DCBD. Muchos de ellos son requerimientos para "más de lo mismo": manipular más registros, más variables, considerar un rango más amplio de hipótesis, alcanzar mejores precisiones y obtener los mismos resultados pero más rápido. Sin embargo, cuando se inicia el trabajo para mejorar algo, resulta que los problemas pueden ser muy sutiles. Datos multi-relacionales, distribuidos, incrementales y heterogéneos representan desafíos mucho más grandes. Esta sección se concreta a presentar algunas de las dimensiones que requieren mejorarse en DCBD y se concentra en algunos de los problemas.

Varios aspectos de los datos pueden ser complejos. Una tabla de datos única, la fuente más popular de datos, podría contener un gran número de registros, muchas dimensiones (atributos), podría usar atributos de muchos tipos y con números grandes de valores por atributo. La mayoría de los métodos de descubrimiento fueron desarrollados para operar sobre una tabla única de tamaño moderado. Existe ya un progreso significativo, pero podemos buscar mejoras adicionales en los métodos de análisis.

Para conjuntos de datos muy grandes, que incluyen millones de tuplas y varios cientos de atributos, la reducción es importante. Aunque las soluciones de alto desempeño pueden ampliar en alguna medida las aplicaciones de los métodos DCBD partiendo de los límites habituales de los sistemas comunes, con frecuencia son necesarios métodos de selección de características y de muestreo para proporcionar descubrimientos interactivos con eficiencia de tiempo. La interactividad de los sistemas de descubrimiento es importante por la naturaleza exploratoria del proceso DCBD. La reducción de variables es también importante para asegurar resultados de descubrimiento claros.

Otra fuente de grandes desafíos viene de la relación entre los datos y los dominios del mundo real que representan. En estadística y en bases de datos se denominan fuentes secundarias de datos para el propósito de descubrimiento (o análisis secundario). Los datos son secundarios porque se recolectan

principalmente para un propósito diferente al del descubrimiento de conocimiento del dominio, por ejemplo, para soportar un proceso del negocio. La adecuación estadística de los objetos representados en los datos con frecuencia es limitada y muchas veces no se encuentran las variables importantes. Deben desarrollarse nuevos paradigmas estadísticos porque los métodos estadísticos asumen comúnmente muestras distribuidas idéntica e independientemente. Para el caso de las grandes bases de datos secundarias obtenidas por un periodo de tiempo, estos supuestos no se cumplen.

Las poblaciones a las que se pueden legítimamente generalizar las regularidades descubiertas en las bases de datos pueden evaluarse con la ayuda de datos externos, como por ejemplo, datos de censos. Este es un resultado de investigación abierto que requiere atención sistemática.

Cuando se considera también ese conocimiento que es generado mediante búsquedas masivas, observamos que los resultados aleatorios, es decir, las regularidades que corresponden a fluctuaciones aleatorias, deben distinguirse de los resultados estadísticamente significativos. Los métodos de la estadística pueden contribuir a evitar que los métodos de la MD traten regularidades aleatorias como reales.

Las búsquedas a gran escala, algo que es central para los métodos de MD, incluyen comparaciones múltiples. En cada ciclo de la búsqueda, se generan muchas hipótesis alternativas para las que se estiman puntuaciones de calidad. Las puntuaciones se usan cuando se selecciona la mejor hipótesis para refinamiento posterior. Estos procedimientos son responsables de tres patologías: errores de selección de atributos, sobreajuste y exceso de búsquedas. Los enfoques estadísticos como el ajuste Bonferroni, prueba de aleatorización y validación cruzada pueden mitigar estos problemas.

Cuando se aplica el muestreo para reducir los datos y generar resultados aproximadamente correctos, podríamos tener problemas al no atender regularidades débiles pero que todavía pueden ser reales. Dado el amplio rango de hipótesis estadísticas consideradas por los sistemas de descubrimiento, aquí se encuentra un gran desafío que puede conducir al crecimiento de la estadística y a la aplicación de los métodos de muestreo.

Recientemente la estadística está desarrollando modelos más complejos y flexibles. Pero especialmente para conjuntos de datos de dimensión alta se ha observado que los métodos simples, como Bayes Naive o vecinos más cercanos, producen más alto desempeño que aquellos modelos complejos. Además, puede resultar necesario un nuevo paradigma para tratar con la dimensionalidad [15]. Deben derivarse las condiciones sobre cuándo son más eficientes los métodos de búsqueda simples que los más poderosos, cuándo las hipótesis más poderosas conducen a mejores resultados, o qué modelos simples deben combinarse en lugar de buscar un único modelo complejo mejor.

Tukey y otros protagonistas del análisis exploratorio de datos enfatizan con frecuencia el principio de robustez [16]. Aplicado a DCBD, esto significa que los resultados del descubrimiento no deben diferir tan sensiblemente respecto a alteraciones pequeñas en los datos, el lenguaje de descripción o los valores seleccionados de las variables dependientes. Por ejemplo, las modificaciones pequeñas en el grupo de datos objeto de análisis, que comúnmente se definen

solo vagamente, no deben conducir a un conjunto totalmente diverso de reglas derivadas. La robustez también es requerida con respecto a supuestos estadísticos inválidos, por ejemplo, sobre la distribución de las variables. El interés principal en DCBD ha sido la exactitud. El fortalecimiento del papel de la robustez en la investigación de descubrimiento es un desafío principal en DCBD.

Los lenguajes de las hipótesis tienen raíces en la lógica y la estadística. La mayoría de los lenguajes atributivos usados para expresar las reglas y los árboles de decisión, tienen un poder expresivo limitado. Algunas veces los enfoques de primer orden basados en lenguajes pueden ser útiles, incluyendo cuantificadores y variables para distinguir diferentes instancias de una o varias clases de objetos [36]. Es un problema abierto y sustancial: cuáles subconjuntos del lenguaje de primer orden son suficientemente expresivos y aún eficientes para varios problemas DCBD.

Los subconjuntos del lenguaje de la lógica de primer orden son importantes para analizar bases de datos multi-relacionales que incluyen varias clases de objetos. Los enfoques para tratar con datos multi-relacionales, especialmente para la múltiple medición de datos, también se originan del campo estadístico, por ejemplo, el desarrollo de métodos de modelación multinivel.

La inducción constructiva está cercanamente relacionada a este problema. Aquí, se construyen dinámicamente variables adicionales durante la búsqueda que sean más apropiadas para describir las regularidades en los datos. Especialmente para el tiempo y espacio relacionado con los datos, dichas variables derivadas pueden ser útiles cuando incluyen términos descriptivos basados en medias, pendientes u otra serie de indicadores. Este problema también incluye métodos de selección de características y agregación.

La integración de varios aspectos de relevancia, por ejemplo, significancia estadística, novedad, simplicidad y utilidad, es otro problema. El conocimiento descubierto puede ser incluido en la base de conocimientos del dominio de un sistema y explotado en el proceso continuo de descubrimiento de conocimiento. Los métodos de descubrimiento se usan para aprender de los usuarios monitoreando y analizando sus reacciones a los resultados descubiertos y presentados para evaluar la novedad de la relevancia. La significancia sustantiva debe evaluarse para tratar con la relevancia de las hipótesis.

Las formas de integración de los sistemas DCBD con otros sistemas como son los sistemas de bases de datos, paquetes estadísticos, sistemas de visualización, sistemas de soporte para las decisiones y sistemas de administración de conocimiento deben ser preparadas. La necesidad de integrar los sistemas DCBD con los sistemas de bases de datos es obvia, ya que la mayoría de los datos son administrados por sistemas de bases de datos. Deben determinarse tipos de consultas estadísticas que puedan ejecutarse eficientemente, de modo que los métodos DCBD para bases de datos muy grandes puedan diseñarse y que sólo utilicen consultas ejecutables eficientemente. Cuando un algoritmo de búsqueda delega la siguiente generación de consultas estadísticas a un servidor de bases de datos que son conjuntamente respondidas dependiendo del procesamiento eficiente de la base de datos, por ejemplo, en un paso único sobre los datos, empleando optimizadores especializados para grupos de consultas, el desarrollo y

mantenimiento de los métodos de minería serán más fáciles y más portables, robustos, escalables e implementaciones paralelas serán soportadas.

La integración de bases de datos puede también soportar el desarrollo de algoritmos de MD espaciales y distribuidos explotando los datos que están distribuidos ya sea por razones de negocios o por escalabilidad.

Capítulo 2. Modelos de procesos estandarizados para DCBD

La búsqueda de patrones en conjuntos de datos tiene una larga tradición en el ámbito académico (en principio en el área estadística y más recientemente en la inteligencia artificial), de allí surgen métodos y procesos como DCBD. Pero la reciente necesidad de la industria por explotar el potencial de sus enormes acumulaciones de datos digitales ha impulsado a vendedores de tecnologías y organizaciones de consultoría a crear metodologías o procesos para el uso de las herramientas computacionales disponibles que implantan los algoritmos propios de MD, de allí que se hable comúnmente del proceso de MD e incluso se trate a DCBD como si fuera la misma cosa.

El uso industrial de la minería de datos requiere mucho más que la aplicación de sofisticadas técnicas como redes neuronales o árboles de decisión sobre tablas de datos. Por esa razón, en esta sección se presenta a la MD como uno de los pasos del proceso DCBD y a su vez éste como un proceso que consta de diferentes fases y en cada una de ellas ubica determinadas técnicas.

En este trabajo, se presentan dos procesos de DCBD, el llamado CRISP-DM (proceso estándar entre industrias para DCBD) [34], definido por un grupo de compañías con amplia trayectoria en el uso de DCBD y el modelo de proceso DCBD de Two Crows [5]. Ambos modelos son muy semejantes y su consideración obedece a que son los que más se apegan a la propuesta original del proceso DCBD y desde este punto de vista, el modelo CRISP-DM rescata mejor esta visión inicial del proceso; además de contar con una documentación detallada para cada una de las etapas.

2.1 El modelo de procesos CRISP- DM

A continuación se presenta el proceso CRISP-DM, el cual consta de las siguientes fases [34]:

1. Comprensión del problema
2. Comprensión de los datos
3. Preparación de los datos
4. Modelación
5. Evaluación de los resultados
6. Despliegue de los resultados

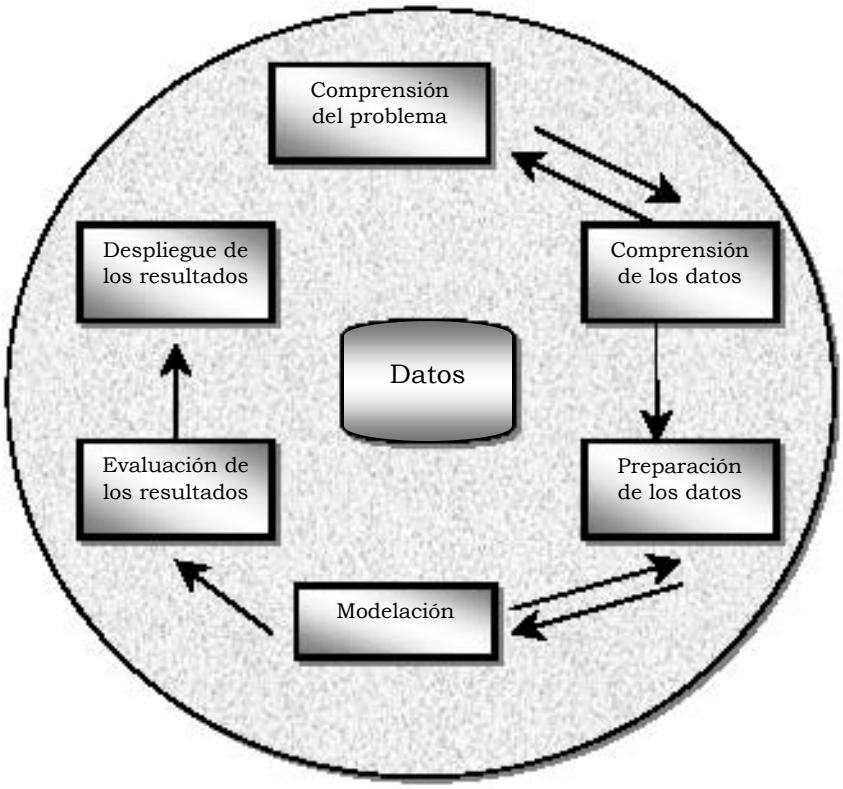

Figura 2.1. Diagrama del proceso de MD CRISP-DM

2.1.1. Comprensión del problema

Los objetivos de esta fase son:

- Determinación de los objetivos: El primer paso y uno de los más importantes es entender la necesidad de hacer minería de datos, determinando cuál es el problema que se desea resolver, para que se convierta en el objetivo del proceso de MD. Los problemas pueden ser diversos: optimizar la respuesta del cliente ante una campaña publicitaria, prevenir el uso fraudulento de tarjetas de crédito, etc.
- Definición de los criterios de éxito: Una vez definido el problema, es necesario disponer de criterios de éxito para el proceso de MD. Esos criterios pueden ser objetivos cuantitativos, por ejemplo un mejor número de detecciones y desviaciones, una mayor tasa de respuesta de los clientes a una campaña publicitaria, etc. Los criterios pueden ser también subjetivos o de naturaleza cualitativa, en este caso, un experto en el dominio califica el resultado del esfuerzo de MD con respecto al conocimiento que se tiene del problema. Los resultados deben contener algunas nuevas percepciones acerca de las relaciones entre las variables del dominio del problema.
- Evaluación de la situación: Una vez definido el problema y sus criterios de solución, hay que tomar en cuenta los aspectos relacionados al problema, como: ¿Cuál es el conocimiento experto o

previo disponible acerca del problema?, ¿Se tienen datos suficientes para intentar resolver el problema?, ¿Se dispone de un glosario que permita aumentar la comunicación entre los expertos en el dominio del problema y los expertos en MD?, ¿Cuál es la relación costo beneficio del proceso de MD?, ¿es rentable?

- Determinación de las metas de la MD: Consiste en una traducción de los objetivos del proyecto en términos de la tecnología de MD. Por ejemplo:

Objetivo del proyecto	Meta de Minería de Datos
Incrementar las ventas	Determinar propiedades de los clientes con respecto a su poder de compra.
Prevenir el uso fraudulento de tarjetas de crédito	Encontrar patrones críticos en el uso fraudulento de tarjetas de crédito o construir un algoritmo que asegure la detección automática de fraudes.

Cuadro 2.1. Ejemplos de metas en términos de la MD

La definición del problema y la meta de la MD están directamente relacionados con la división básica de los tipos de problemas de MD (que serán discutidos en la fase de Modelación), a saber: descripción y resumen de los datos, clasificación, predicción, descubrimiento de asociaciones, análisis de dependencia y segmentación.

- Producción de un plan del proyecto: Finalmente, se crea un plan para el proyecto que describa los pasos a seguir y las técnicas empleadas en cada paso.

2.1.2. Comprensión de los datos

Posterior a la definición del problema que se desea resolver y de haber creado un plan para llevarlo a cabo, es importante enfocarse en el aspecto principal de la MD: los datos. Hay muchas cosas que se pueden aprender acerca de los datos antes de aplicar las técnicas de MD. Las actividades a desarrollar en esta fase son:

- Recolectar los datos iniciales: El primer paso es la adquisición de los datos iniciales y su preparación para futuro procesamiento. El proceso de adquisición de datos puede producir las siguientes salidas: listas de datos adquiridos, localización de datos y métodos a usar para su adquisición; problemas y soluciones relacionados a la adquisición de datos.
- Descripción de los datos: Luego de adquiridos, estos deben ser descritos, lo cual significa principalmente establecer el volumen de los datos (número de registros y campos por registro), identificación y significado de cada campo y la descripción del formato inicial de los datos.
- Exploración de los datos: Este paso no es obligatorio, pero si útil en muchos aspectos. El rol principal de la exploración de datos en esta fase es encontrar una estructura general para los datos. La exploración no está directamente relacionada con la solución al

problema (esa es una tarea para las técnicas de modelación de MD), sino que contempla la aplicación de pruebas estadísticas básicas que revelen propiedades en los datos recién adquiridos.

- Verificación de la calidad de los datos: Aquí se realizan revisiones sobre los datos para determinar la consistencia de los valores individuales de los campos, la cantidad y distribución de los datos faltantes, encontrar valores fuera de rango (que pueden representar ruido o un fenómeno nuevo e interesante). La idea en este punto es asegurar qué tan completos (completitud) y correctos (correctitud) son los datos. *Completitud* se refiere a la proporcionalidad y regularidad de los valores faltantes y *Correctitud* se refiere al descubrimiento de valores erróneos en los datos y su posible solución.

2.1.3. Preparación de los datos

Aunque el núcleo del proceso es la aplicación de las técnicas de modelación de MD y la evaluación de los modelos resultantes en base a sus valores predictivos o descriptivos, no debe disminuirse la importancia que tienen los esfuerzos en la preparación de los datos. La fase de preparación de los datos está dividida en:

- Selección de datos: Un subconjunto de los datos adquiridos en las fases previas es seleccionado basado en criterios también establecidos en fases anteriores: calidad de los datos (completitud y correctitud), limitaciones en el volumen o en los tipos de datos que están relacionadas con las técnicas de MD preseleccionadas.
- Limpieza de los datos: Este paso complementa al anterior, también es uno de los que más tiempo consumen, debido a la enorme cantidad de técnicas que pueden aplicarse para optimizar la calidad de los datos con miras a la fase de modelación. Algunas técnicas son: Normalización de los datos, discretización de campos numéricos, tratamiento de valores ausentes, reducción del volumen de datos.
- Construcción de nuevos datos: Aquí se crean nuevas estructuras a partir de los datos seleccionados, por ejemplo: generación de nuevos campos a partir de dos o más ya existentes, creación de nuevos registros (muestras), fusión de dos tablas que contengan atributos diferentes para el mismo objeto, agregación de nuevos campos o nuevas tablas donde se resumen características de múltiples registros o de otros campos en nuevas tablas de resumen.
- Formateo de los datos**:** Este paso en la preparación de los datos, implica transformaciones sintácticas de los datos sin modificar su significado, esto con la idea de permitir o facilitar el empleo de alguna técnica de MD en particular. Algunas ejemplos son: reordenación de los campos y/o registros de la tabla (algunas herramientas de modelación requieren que los campos estén en cierto orden, las redes neuronales requieren que los registros estén ubicados aleatoriamente), ajuste de los valores de los campos a las

limitaciones de las herramientas de modelación (remover comas, tabuladores, caracteres especiales, máximos y mínimos para las cadenas de caracteres, etc.)

2.1.4. Modelación

Lo novedoso y abundante de las técnicas disponibles y de los algoritmos involucrados en la fase de modelación hacen de esta la fase más interesante del proceso de MD. Los pasos importantes en la fase de modelación son:

- Selección de la técnica de modelación: Al principio del proceso de MD se establece el problema a resolver y la meta de MD implicada, ahora es el momento de seleccionar una técnica de MD concreta. Cuando se escoge una técnica apropiada entre numerosas técnicas de modelación disponibles en MD se debe tener en cuenta el objetivo principal del proyecto y su relación con la principal división de las herramientas de MD de acuerdo al tipo de problema. La primera división de las técnicas de modelación de MD está hecha en base al tipo de tarea de descubrimiento de conocimiento que se desea: Predicción o Descripción. La siguiente tabla muestra algunas clases de tareas de modelación y las técnicas de MD adecuadas.

.

Tareas de modelación	Técnicas
Clasificación	Métodos de inducción de reglas, Árboles de Decisión, K vecinos más cercanos, razonamiento basado en casos.
Predicción	Análisis de regresión, Árboles de regresión, redes neuronales, K vecinos más cercanos.
Análisis de Dependencia	Análisis de Correlación, Análisis de regresión, Reglas de Asociación, Redes Bayesianas, programación con lógica inductiva.
Segmentación	Técnicas de Agrupación, redes neuronales, técnicas de visualización.

Cuadro 2.2. Diferentes tareas de modelación en MD

- Generación de pruebas para el modelo: Luego de construido un modelo, se debe generar un procedimiento o mecanismo para probar la calidad y validez del modelo. Por ejemplo, en una tarea supervisada de la MD como la clasificación, es común usar la tasa de error como medida de la calidad. En consecuencia, típicamente se separan los datos en dos conjuntos, uno de entrenamiento y otro de prueba, para luego construir el modelo basado en el conjunto de entrenamiento y medir la calidad del modelo generado con el conjunto de prueba.
- Construcción del modelo: Una vez que la técnica de modelación ha sido seleccionada, se procede a aplicarla sobre los datos previamente preparados para generar un modelo. Todas las técnicas de modelación tienen un conjunto de parámetros que determinan las características del modelo a generar. La selección de los parámetros óptimos para la técnica de modelación es un proceso iterativo y se

basa exclusivamente en los resultados generados. Estos deben ser interpretados y su rendimiento justificado.

- Evaluación del modelo: Una vez que los modelos son generados, estos son interpretados de acuerdo al conocimiento preexistente del dominio y los criterios de éxito preestablecidos. Los expertos en el dominio del problema juzgan los modelos dentro del contexto del dominio y expertos en MD aplican sus propios criterios (seguridad del conjunto de prueba, perdida o ganancia de tablas, etc.)

2.1.5. Evaluación de los resultados

En las fases previas (sobre todo en la de modelación), la evaluación se refiere a la exactitud y generalidad del modelo obtenido, mientras que en esta fase se involucra la evaluación del modelo con respecto a los objetivos del proyecto. En esta fase se debe decidir si hay o no razones para construir un modelo eficiente (relación costo-beneficio), si es aconsejable probar el modelo en un problema real. Además de los resultados directamente relacionados con el objetivo del proyecto, ¿es aconsejable calificar el modelo con relación a otros objetivos diferentes a los originales?, esto podría revelar información adicional. El paso importante en esta fase es:

- Revisión del Proceso: se refiere a calificar al proceso completo de DCBD con la idea de identificar elementos que pudieran mejorarse. Por último, en esta fase se toma una decisión acerca de futuras fases. Si se ha determinado que las fases hasta este momento han generado resultados satisfactorios, podría decidirse pasar a la fase de despliegue de resultados, si no, podría decidirse por otra iteración desde la fase de preparación de datos o de modelación con otros parámetros. Podría ser incluso que en esta fase se decida partir de cero con un nuevo proyecto de DCBD.

2.1.6. Despliegue de los resultados

En esta fase se define una estrategia para desplegar los resultados del proceso DCBD.

- Monitoreo y Mantenimiento: Si los modelos resultantes del proceso de MD son desplegados en el dominio del problema como parte de la rutina diaria, es aconsejable preparar estrategias de monitoreo y mantenimiento para ser construidas sobre los modelos. La retroalimentación generada por el monitoreo y mantenimiento pueden indicar si el modelo está siendo utilizado apropiadamente.
- Reporte final: Es la conclusión del proyecto de DCBD. Resume los puntos importantes del proyecto, la experiencia ganada y explica los resultados producidos.

2.2 El modelo de procesos TWO CROWS

Este modelo presentado por Two Crows Corporation, también llamado por ellos Minería de Datos para el Descubrimiento de Conocimiento, toma muchas cosas de su propia experiencia y de los procesos de DCBD y

CRISP-DM. Por ejemplo, es también un proceso iterativo e interactivo y tiene fases parecidas a las de CRISP-DM. Las fases de este proceso son:

1. Definición del problema
2. Construcción de la base de datos
3. Exploración de los datos
4. Preparación de los datos para la modelación
5. Construcción de modelos
6. Evaluación del modelo
7. Despliegue de modelos y resultados.

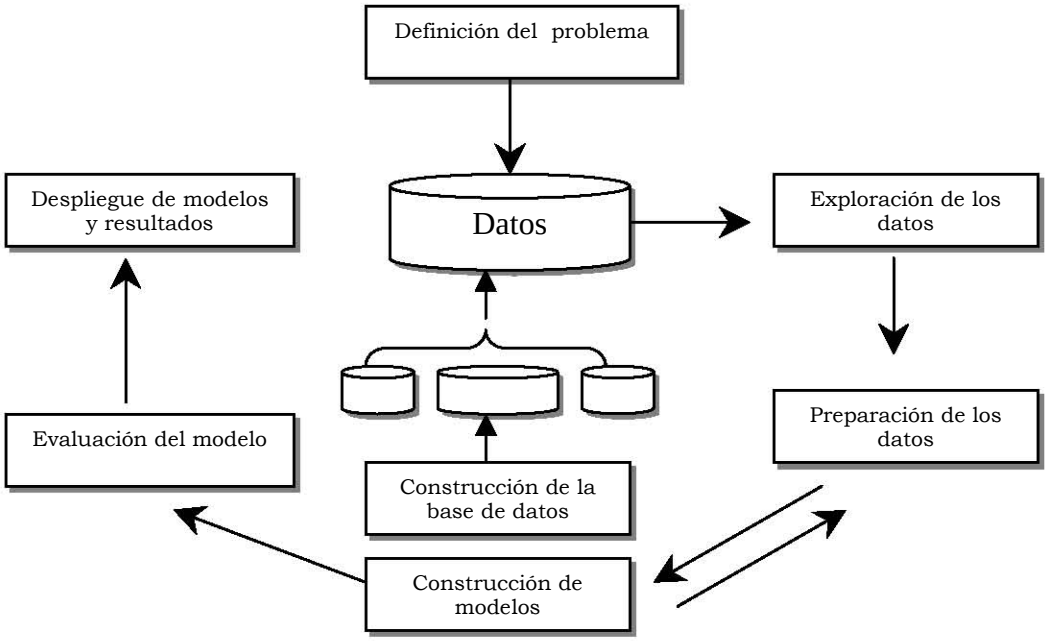

Figura 2.2. Diagrama del proceso de MD TWO-CROWS

2.2.1. Definición del problema

El primero y más importante de los prerrequisitos para el descubrimiento de conocimiento es comprender los datos y el negocio. Sin esta comprensión, ningún algoritmo, por sofisticado que sea, proveerá resultados confiables. Sin este conocimiento, no será posible identificar el problema que se desea resolver, preparar los datos para las técnicas de Minería de Datos o interpretar correctamente los resultados. Para hacer el mejor uso de la MD, deben tenerse claros los objetivos. Por ejemplo, puede ser que se desee incrementar la respuesta a una campaña publicitaria por correspondencia. Dependiendo de los objetivos específicos, tales como "incrementar la tasa de respuesta" o "incrementar el valor de la respuesta" se creará un modelo diferente.

2.2.2. Construcción de la base de datos

Esta fase y las siguientes tres son el corazón de la preparación de los datos, conjuntamente, incluyen la mayor parte del tiempo y el esfuerzo de un proceso de Minería de Datos.

Los datos que van a ser estudiados deben estar reunidos en una base de datos. Nótese que esto no necesariamente implica el uso de un sistema manejador de base de datos. Dependiendo de la cantidad de datos, de su complejidad y del uso que se les pretenda dar, un archivo plano o una hoja de cálculo podrían ser suficientes.

En general no es buena idea emplear el repositorio de datos corporativo para MD. Es recomendable crear un nuevo repositorio para los datos que van a ser analizados. Las técnicas de MD producirán abundantes y frecuentes accesos al repositorio de datos, lo que podría generar problemas. Seguramente se estarán haciendo modificaciones al repositorio de datos corporativo. Además, probablemente se pretenda añadir a los datos de la organización los de otras fuentes externas, como por ejemplo encuestas, o se pretenda incluir nuevos campos de datos.

Otra razón para disponer de un repositorio independiente de datos para el análisis es que posiblemente el repositorio de datos corporativo no soporte la clase de exploraciones que las técnicas de MD requieren. De hecho, muchas organizaciones prefieren usar sistemas manejadores de bases de datos de propósito especial que den mucho mejor soporte a los requerimientos de la MD.

Para crear esta nueva base de datos se deben realizar actividades como:

- Recolección de datos: Se identifica la fuente de los datos que se van a emplear en el análisis. Posiblemente los datos necesarios nunca han sido recolectados o se necesiten datos externos de bases de datos públicas (tales como censo o clima) o privadas (tales como datos sobre uso de tarjetas de crédito). Un reporte de recolección de datos muestra las propiedades de las diferentes fuentes de datos como por ejemplo: fuente (interna o vendedor externo), propietario, persona u organización responsable de su mantenimiento, costo, estructura de almacenamiento, tamaño en tablas, en registros, en bytes, soporte físico del almacenamiento (CD-ROM, cinta, servidor, etc.), requerimientos de seguridad y restricciones de uso.
- Descripción de los datos: Aquí se describe el contenido de cada archivo o tabla de la base de datos. Entre las propiedades que se enumeran en el Reporte de Descripción de datos están: Nombre de la tabla, número de campos, número/porcentaje de registros con datos ausentes, nombre de los campos y para cada campo el tipo de dato, definición, descripción, fuente del campo, unidad de medida, lista de valores, rango de valores, número/porcentaje de valores ausentes, frecuencia, relación con la clave primaria o foránea.

- Selección: Aquí se selecciona un subconjunto de los datos para ser procesados. No se refiere a una toma de muestras o a escoger variables predictivas, es más bien a una simple eliminación de datos irrelevantes o innecesarios. Otros criterios para excluir datos pueden ser restricciones en su uso, costos, o problemas de calidad de estos.
- Evaluación de la calidad de los datos y filtrado: Para obtener un buen modelo se necesitan buenos datos. Una evaluación de la calidad de los datos identifica características de los datos que afectarán la calidad del modelo. En esencia se pretende asegurar no solo la correctitud y consistencia de los datos, sino también que estos correspondan a mediciones del mismo fenómeno en la misma manera. Hay varios tipos de problemas de calidad de los datos como valores incorrectos, valores correctos colocados en lugares equivocados o valores ausentes.
- Integración y consolidación: Es ahora cuando se toman los datos de distintas fuentes y se crea una única base de datos para alojarlos, lo cual requiere conciliar las diferencias entre distintos valores de datos de varias fuentes. Hacer esto de manera deficiente es una de las mayores fuentes de problemas de calidad de datos. Hay muchas maneras en que los datos pueden ser definidos y empleados en distintas bases de datos. Algunas inconsistencias son fáciles de solucionar (como distintas direcciones para un mismo cliente), otras son más sutiles como el uso de homónimos o sinónimos, que un mismo cliente tenga varios nombres o números de identificación, o que se empleen dólares USA o canadienses.
- Construcción de metadatos: En esencia es crear una base de datos sobre la base de datos. Está basado principalmente en los reportes de descripción de los datos y provee información que será usada en la creación física de la base de datos y que servirá también para analizar los datos y comprender los modelos generados.
- Carga de la base de datos: En la mayoría de los casos los datos a ser analizados deben estar ubicados en una base de datos independiente, y dependiendo de su cantidad y complejidad, puede ser complicado y requerir la participación de expertos.
- Mantenimiento de la base de datos: Una vez creada, la base de datos necesita cuidados, debe ser respaldada periódicamente, su desempeño debe ser monitoreado, y ocasionalmente requiere ajustes que exigen espacio de almacenamiento o aumento del rendimiento. Estas tareas involucran a profesionales en sistemas de información.

2.2.3. Exploración de los datos

En esta fase se emplean numerosas técnicas de visualización de datos, de búsqueda de relaciones entre variables y otras medidas para la exploración de los datos. La meta es identificar los campos con mayor potencial predictivo y cuales variables con valores derivados pueden ser útiles. Bases

de datos con muchos campos hacen que esta tarea sea ardua y consuma tiempo.

2.2.4. Preparación de los datos para la modelación

Este es el último paso de la preparación de los datos antes de construir modelos. Hay cuatro partes importantes en esta fase:

- Selección de Variables: En un caso ideal se toman todas las variables disponibles, alimentando con ellas a los algoritmos de Minería de Datos y dejándolos encontrar las mejores predicciones. En la práctica esto no funciona bien. Una de las razones es que el tiempo empleado para construir un modelo incrementa con la cantidad de variables. Otra razón es que incluir variables ciegamente puede llevar a la creación de modelos erróneos. Un error muy típico es emplear una variable de predicción que sólo puede ser conocida si se conoce el valor de la variable a predecir, por ejemplo, algunas veces se ha incluido inadvertidamente la fecha de nacimiento en un algoritmo que pretende predecir la edad de un grupo de personas. Aunque muchos algoritmos de Minería de Datos ignoran los campos irrelevantes, no es bueno depender para todo de los algoritmos. Incluir el número de cédula como variable predictiva es en el mejor de los casos inútil y en el peor puede reducir drásticamente la calidad del modelo.
- Selección de registros: Igual que en el caso anterior, sería ideal poder emplear todos los datos disponibles, pero esto podría tardar demasiado tiempo o requerir un hardware más potente. Por lo tanto, es buena idea tomar muestras de los datos cuando la base de datos es demasiado grande. En la mayoría de los casos no habrá perdida de calidad pero se debe estar seguro de haber tomado las muestras al azar.
- Construcción de nuevas variables: Con frecuencia es necesario crear nuevas variables predictivas derivadas de datos crudos. Por ejemplo, en la predicción del riesgo de crédito se usa con más frecuencia la tasa deuda-ingresos como variable predictiva que los valores individuales de deuda e ingresos.
- Transformación de variables: La tarea de Minería de Datos seleccionada puede dictar el cómo se representan los datos. Por ejemplo, los campos pueden ser ajustados para que entren en un rango particular (en muchos casos, el rango [0,1]). Muchos árboles de decisión empleados para clasificar requieren que variables continuas sean discretizadas o divididas en clases como “Alto, Medio, Bajo”.

2.2.5. Construcción de modelos

La construcción de modelos es un proceso iterativo. Será necesario explorar múltiples modelos alternativos hasta encontrar el más útil a la organización. Lo aprendido durante la creación de modelos puede llevar a

modificar nuevamente los datos incluso a cambiar el objetivo del proyecto. Una vez que se ha decidido qué tarea de MD va a efectuarse, se debe escoger un tipo de modelo para representar los resultados (como un árbol de decisión, una red neuronal, algún otro método propietario, etc.). La selección del tipo de modelo tendrá influencia en la preparación de los datos. Una vez listos los datos, se puede entrenar al modelo.

La creación de modelos predictivos requiere un protocolo de entrenamiento y validación bien definido. Este tipo de protocolo es llamado a veces entrenamiento supervisado. En esencia, consiste en entrenar al modelo con una porción de los datos y validarlo luego con otra porción. De no hacerse de esta manera, se pueden obtener modelos sobreestimados que sólo pueden predecir correctamente para el conjunto de datos procesados.

La validación se puede hacer de varias maneras. La validación Simple consiste en entrenar al modelo con una porción grande de los datos y dejar una más pequeña (5% al 30%) para validar. Si se dispone de un conjunto de datos pequeño, se puede hacer una Validación Cruzada, que consiste en separar los datos en dos grupos, entrenar al modelo con el primer grupo y validarlo con el segundo, luego se entrena al modelo con el segundo grupo y se valida con el primero, por último se entrena al modelo con todos los datos y se usa el promedio de los errores de los modelos anteriores. De hecho, el método más frecuentemente usado es el de la Validación Cruzada con N grupos, que sigue el mismo procedimiento pero para más de dos grupos.

Basado en los resultados de la construcción del modelo, se puede decidir si crear otros modelos empleando la misma técnica con parámetros diferentes o intentar con otra técnica o algoritmo.

2.2.6. Evaluación del modelo

Después de construir un modelo se deben evaluar sus resultados e interpretar sus significados. Debe tenerse presente que la confiabilidad calculada para el modelo sólo aplica para los datos sobre los que se realizó el análisis.

Las matrices de confusión son muy empleadas en problemas de clasificación y consisten de una tabla que indica cuantas clasificaciones se han hecho para cada tipo, la diagonal de la tabla representa las clasificaciones correctas.

Predicción	Clase A	Clase B	Clase C
Clase A	45	2	3
Clase B	10	38	2
Clase C	4	6	40

Cuadro 2.3 Matriz de confusión

Véase que en este ejemplo, se han hecho 123 predicciones correctas de un total de 150 (82% de efectividad). Además, observando la tabla se puede tener una idea de donde no está prediciendo bien el modelo.

Otros métodos gráficos pueden ayudar en otros tipos de modelos a evaluar los resultados. Por ejemplo las graficas de barras.

2.2.7. Despliegue de modelos y resultados

Una vez que el modelo ha sido construido y validado puede ser empleado en una o dos maneras importantes: La primera de ellas consiste en que un analista recomiende acciones basadas simplemente en la observación del modelo y sus resultados. La segunda consiste en aplicar el modelo a diferentes conjuntos de datos, por ejemplo, para marcar ciertos registros según su clasificación o asignarles puntuaciones tales como la probabilidad de acción (ejemplo, la probabilidad de respuesta favorable a una campaña de publicidad directa por correo).

A veces los modelos son parte del proceso del negocio, tales como análisis de riesgo, autorizaciones de crédito o detección de fraudes. En esos casos el modelo es incorporado a una aplicación de negocios.

Capítulo 3. Una aplicación específica de DCBD. Análisis previo.

3.1. Comprensión del problema

Antecedentes

Durante 1980 y principios de los 90's, en el Reino Unido y en otros países, los investigadores del tema eficacia escolar, se enfocaron en la producción y el uso de indicadores de desempeño, que casi siempre se obtenían de los logros académicos demostrados por los estudiantes en pruebas estandarizadas externas. A raíz de esto, se desarrolló un gran debate, conducido principalmente por consideraciones políticas, sobre lo apropiado de usar estas mediciones para realizar clasificaciones de escuelas. Este debate sobre la clasificación de escuelas a partir únicamente del *desempeño bruto* que obtienen sus alumnos, promovió diversas iniciativas para la evaluación de la calidad de las mismas. Existen criterios de evaluación diferentes en diversos círculos educativos; pero correcta o incorrectamente, es común percibir estos resultados como el parámetro tradicional del éxito educativo [20]. El conocimiento recopilado de la literatura sobre el tema de la eficacia escolar contribuye mucho en este trabajo debido a que se enfoca en las diferencias en el logro académico entre y dentro de las escuelas; siendo el objetivo primordial obtener conocimiento con respecto a las relaciones entre variables explicatorias y respuesta, utilizando los modelos apropiados [19].

Actualmente, existen dos marcos de trabajo básicos para la interpretación de los logros obtenidos por las escuelas [30]. La forma más fácil y que además está relacionada con el modelo estándar (noción absoluta de eficacia), es comparar el *desempeño bruto* de unas escuelas contra otras; con un promedio local o nacional, para luego hacer declaraciones relacionadas con la eficacia de cada escuela. Sin embargo, dado que los alumnos no son asignados aleatoriamente a las escuelas, estos modelos corren el riesgo de gratificar a las escuelas por la obtención de buenos resultados, siendo estos logros principalmente un reflejo del tipo de estudiante que admiten, en lugar de lo que realmente hacen por ellos. En contraste, el modelo actual con mayor aceptación (noción relativa de eficacia) y sus variantes presentan los resultados de las pruebas en el contexto del tipo de estudiante que atiende cada institución y por tanto permiten comparaciones más justas. Este último enfoque de "valor agregado" es ampliamente defendido por los estudiosos del tema.

Existe una considerable literatura sobre los métodos para comparar escuelas y otras instituciones con base en el logro académico de sus alumnos. Entre los requerimientos mínimos para realizar comparaciones institucionales válidas están el realizar análisis basados en datos al nivel de los individuos, ajustados por características de entrada[1] y utilizando técnicas de modelación Multinivel [1].

En [1] y en la discusión en torno al mismo, surgieron algunos problemas: los puramente técnicos de realizar estimaciones de grandes conjuntos de datos y

[1] Se denomina característica de entrada a variables relevantes que explican parte de la variación en la variable respuesta y que son medidas ya sea al nivel del estudiante o al nivel de la escuela.

que se ha resuelto eficazmente con el desarrollo de nuevos programas computacionales [31]. Los restantes hacen énfasis en la consideración de características relevantes, tanto a nivel de alumnos como de escuelas, que puedan utilizarse para ajustar el *desempeño bruto*; además de la naturaleza multivariada de los resultados de la escuela y del tipo de interpretación que se realice de estos.

En consecuencia, dada la relativa facilidad de acceso a fuentes de información secundaria derivadas del ámbito educativo en México, concretamente de evaluaciones estandarizadas externas cuya finalidad es la selección de aspirantes a instituciones de Educación Media Superior (EMS); además de la disponibilidad de datos sobre el origen de estos aspirantes y otras características de interés (hábitos de estudio, involucramiento de los padres en la actividad escolar, etc.), es posible la realización de estudios de carácter exploratorio, que aborden temáticas relacionadas con las comentadas en párrafos anteriores. A la luz de estos trabajos, es interesante observar, a través de la modelación de estas estructuras jerárquicas, qué tanto difieren las escuelas en sus logros después de ajustarlos considerando las características de los hogares de origen de sus estudiantes (características sociofamiliares) y el género, con la intención de realizar comparaciones más justas entre las instituciones. Aunado a lo anterior, también resulta relevante la exploración de los efectos del origen social de los estudiantes en cada una de las escuelas, en la medida que esto nos permite obtener una visión general de cómo se distribuye el conocimiento entre sus estudiantes.

Es importante destacar que entre los datos que se encuentran disponibles no existe algún indicador de logro inicial que nos proporcione información sobre cómo ingresaron los estudiantes que presentaron la prueba en el 2001 a sus respectivas secundarias de origen, por lo que no será posible realizar un análisis de eficacia escolar acorde a los requerimientos mínimos que proponen los expertos para este tipo de estudios. Sin embargo, cabe señalar en este punto, que además de enfocarnos a realizar ajustes en el desempeño bruto para obtener el desempeño neto y poder realizar comparaciones en igualdad de condiciones entre las escuelas, se propone un criterio subjetivo para identificar escuelas eficaces con la información que se encuentra disponible.

Determinación de los objetivos del problema

Los objetivos principales de este estudio son:

1. Explorar la relevancia de las características sociofamiliares y de género de los estudiantes para explicar las diferencias del logro en la prueba.
2. Explorar la homogeneidad de la composición sociofamiliar de los egresados de cada escuela.
3. Explorar las diferencias entre las escuelas en términos de su desempeño neto.
4. Explorar el criterio de eficacia escolar, propuesto por el experto en el dominio, considerando las características sociofamiliares de los estudiantes y el enfoque de la distribución social del conocimiento.

Criterios de éxito del proyecto

Debido a que se analizará una fuente de datos secundaria y no específicamente diseñada para contestar determinadas preguntas de investigación del ámbito en cuestión, este estudio será exploratorio y no se tomarán los resultados como concluyentes, sino como indicativos de relaciones existentes entre las variables de interés y que a su vez podrían utilizarse para formular hipótesis de futuras investigaciones. Por tanto, el criterio de éxito será subjetivo en el sentido de que serán los expertos en el dominio los que juzguen la importancia práctica de los resultados esperados y aquellos no esperados que surjan en el curso del análisis.

Evaluación de la situación

A lo largo de las casi cuatro décadas de historia de los estudios de eficacia escolar el elevado número de publicaciones sobre el tema pone de manifiesto el interés que sigue despertando, así como la evolución y desarrollo que dichos estudios vienen experimentando en muchos aspectos relevantes, para superar las muchas dificultades que se plantean en un tema tan complejo. Es importante la contribución que el movimiento de escuelas eficaces ha realizado para identificar y conocer la importancia de los factores de eficacia escolar y la investigación continúa su curso, pero existen un buen número de interrogantes, preocupaciones y críticas que surgen sistemáticamente en el contexto educativo. Desde distintos ámbitos se han abordado algunos de los problemas más importantes y actualmente se encuentran prestigiosos investigadores centrados en estos estudios, intentando encontrar nuevas alternativas, técnicas o metodológicas, para evaluar la eficacia escolar.

Sin entrar al debate sobre la causalidad en las ciencias sociales, debe reconocerse que actualmente existe una creciente tendencia a recomendar y discutir la necesidad de contar con diseños experimentales que permitan extraer conclusiones más certeras. En general, se acepta que los estudios de las escuelas eficaces no pueden realizar inferencias causales debido a que el diseño no es apropiado para este objetivo. En [38], por ejemplo, señalan que las inferencias causales serían aceptables sólo bajo condiciones experimentales que aseguren:

- ✓ La asignación aleatoria de los alumnos a las escuelas,
- ✓ La asignación aleatoria de las escuelas a los tratamientos (políticas) que teóricamente se supone generan diferencias (por ejemplo, liderazgo, clima escolar, enfoque didáctico, etc).

Por otro lado, en [19], el autor ha sintetizado cuatro requisitos mínimos que este tipo de investigaciones cuasi-experimentales debieran cumplir:

1. Ser un estudio longitudinal, de tal forma que las diferencias iniciales en los estudiantes y los subsecuentes eventos en las escuelas puedan ser considerados;
2. Enfocar el análisis con un apropiado modelo multinivel, de tal forma que las inferencias estadísticas sean válidas y en particular, sea explorada la existencia de "eficacias diferenciales" ;
3. Disponer de cierta replicación del estudio en otros momentos y en otros lugares a los efectos de fundamentar la replicabilidad; y

4. Disponer de cierta explicación plausible del proceso por el cual las escuelas devienen en eficaces.

Además, el autor también señala que de todos los estudios más frecuentemente citados en la bibliografía especializada, sólo el informe de Mortimore *et al.* (1988) cumple con estos requisitos; seguido parcialmente por Rutter *et al.* (1979). Esta recurrente falta de rigor metodológico está en la base de las agudas críticas que se han hecho a quienes trabajan en esta perspectiva en el mundo anglosajón [3]

En resumen, como se mencionó anteriormente, dado que se analizará una fuente secundaria de datos y no se dispone de información específica sobre las escuelas de origen (tamaño de la escuela, tipo de escuela, etc.), salvo las variables composicionales que resulten de interés y que se puedan construir a partir de sus estudiantes egresados (contextualización), el presente estudio es de carácter exploratorio y enfocado a contestar las preguntas que se formularon inicialmente en los objetivos, obviamente sin intentar proporcionar resultados concluyentes.

Es también importante señalar, que se cuenta con una gran cantidad de datos correspondientes a la aplicación de una prueba estandarizada en el año 2001, misma que tiene como propósito fundamental complementar los criterios de admisión de los estudiantes a determinadas instituciones de educación media superior públicas en el D.F. y la zona conurbada del Estado de México. No todas las escuelas están representadas en la información disponible y existen algunas en la base de datos que tienen un número reducido de sustentantes. Por lo anterior, es necesario precisar que este estudio es sólo un ejercicio de aproximación que debe tomarse con cautela porque no arroja datos exhaustivos. Su universo incluye aproximadamente 1371 escuelas secundarias del área metropolitana, de las que se tienen suficientes datos lo que permite en cierto sentido afirmar algo sobre ellas.

Terminología

La metodología CRISP-DM propone la generación de dos glosarios: uno relacionado con términos del dominio del problema y otro que contemple términos del área de la MD que se utilicen a lo largo de todo el ciclo de vida del proyecto. Esto con la finalidad de establecer un puente para facilitar la comunicación entre los expertos en el dominio y los de MD. Al final del documento se puede acceder a este banco terminológico que se generó en el proyecto.

Determinación de las metas de la minería de datos

En este apartado se presenta la meta del proyecto en términos técnicos.

Cuadro 3.1. Objetivos y metas de la MD.

Objetivo del proyecto	Meta de la minería de datos
Construir tipologías de escuelas, equitativamente comparables, con base en el desempeño neto en la prueba.	➢ Construir un modelo multinivel que ajuste el desempeño bruto en la prueba considerando características de entrada relevantes al nivel del estudiante.

El plan del proyecto

Este proyecto se planeó para realizarse en seis meses y la distribución de tareas a lo largo de este periodo se muestra en el cuadro 3.2. Considerando que no es un proyecto financiado por la industria o la academia, no hay recursos específicamente asignados a cada fase, por lo que se realiza con los recursos básicos: hardware y software disponible.

Fase	Abril	Mayo	Junio	Julio	Agosto	Sept.
1.Comprensión del problema - Determinar los objetivos - Evaluar la situación - Determinar las metas de la minería de datos - Producir un plan del proyecto.						
2. Comprensión de los datos - Colectar los datos iniciales - Describir los datos - Explorar los datos - Verificar la calidad de los datos						
3. Preparación de los datos - Seleccionar los datos - Depurar los datos - Construir los datos - Integrar los datos - Formatear los datos						
4. Modelación - Seleccionar la técnica de modelación - Generar el diseño de prueba - Construir el modelo - Evaluar el modelo						
5. Evaluación - Evaluar los resultados - Revisar el proceso						
6.Presentación - Planear la presentación - Producir el reporte final - Revisar el proyecto.						

Cuadro 3.2. Planeación del proyecto

Evaluación inicial de herramientas y técnicas

Las herramientas disponibles que se utilizaron para las etapas de comprensión y preparación de los datos son los paquetes estadísticos: Statistica versión 6.0 y SPSS versión 12, ya que cuentan con la suficiente capacidad para trabajar con bases de datos grandes y generar resultados en un tiempo razonable; obviamente esto también depende de la capacidad del hardware disponible.

Para alcanzar las metas del proyecto, en la parte que se considera como de minería de datos, se ajustó un modelo multinivel. Dicho proceso se realizó con el paquete estadístico HLM.

3.2 Comprensión de los datos

El papel principal de la exploración de los datos en esta fase es encontrar una estructura general. La exploración no está directamente relacionada con la solución al problema (esa es una tarea para las técnicas de modelación de la

minería de datos), sino que envuelve la aplicación de pruebas estadísticas básicas que revelen propiedades en los datos recién adquiridos: si se tienen campos nominales se crean tablas de frecuencias y para los campos numéricos, se grafica su distribución y se buscan dependencias.

Un número significativo de preguntas de la hoja de registro contempla respuestas que se encuentran en una escala ordinal, por ejemplo, en el caso de los hábitos de estudio, los aspirantes deben determinar aproximadamente con qué frecuencia realizan ciertas formas de estudio (casi nunca, pocas veces, muchas veces, casi siempre). Por otro lado, hay preguntas que requieren respuestas que están dadas en una escala numérica discreta, por ejemplo, el número de hermanos. De acuerdo al tipo de escala en que se miden las respuestas de cada una de las preguntas es que se determinó la descripción estadística adecuada. La mayoría de las variables relevantes cuenta con una descripción básica en términos de tablas de frecuencias y gráficos (histogramas o sectores) que reflejen ciertos patrones en esta primera aproximación.

En cuanto a los detalles importantes en esta primera exploración de los datos, cabe mencionar que se consideró la *No respuesta* en el conteo de las tablas de frecuencia, con la finalidad de determinar la cantidad y distribución de los datos faltantes para cada pregunta. Además, se presentó una columna adicional con el porcentaje ajustado denominado *porcentaje válido*, sin considerar la *No respuesta*. Para el caso de las preguntas en escala numérica discreta, se considera una columna adicional que registra el porcentaje acumulado (considerando el porcentaje válido como base). El cuadro 3.3 muestra un ejemplo de este tipo de análisis. Se incluyó también un apartado con estadísticas descriptivas del desempeño en la prueba estandarizada, por asignaturas y global, con el fin de caracterizar la distribución de estos puntajes en términos de las medidas de tendencia central y dispersión más comunes.

Aunado a lo anterior, se exploró la relación de cada variable de la encuesta con la puntuación global en la prueba (logro), para lo que se incluyeron estadísticas descriptivas (media y desviación estándar) para cada una de las categorías de la variable en cuestión, esto permite tener una primera aproximación de las relaciones entre las posibles variables predictoras con la variable respuesta (logro en la prueba). El cuadro 3.4 presenta un ejemplo.

Es importante mencionar que dada la extensión de los reportes exploratorios generados en el proceso de análisis y revisados en su momento por el experto en el dominio para la consideración de los indicadores relevantes para su inclusión en la etapa de MD, no se incluirán todos los detalles de los cuadros producto de este análisis. En el anexo A y B se incluyen solo algunos cuadros con información sobre algunas de las variables más relevantes.

Número de hrs. estudio	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
0	3465	1.5	1.7	1.7
1	29145	12.3	13.9	15.6
2	43536	18.3	20.8	36.4
3	33836	14.2	16.2	52.6
4	19464	8.2	9.3	61.9
5	22624	9.5	10.8	72.7
6	10764	4.5	5.1	77.8
7	9784	4.1	4.7	82.5
8	8170	3.4	3.9	86.4
9	3522	1.5	1.7	88.1
10	10574	4.5	5.0	93.1
11. Más de 10	14587	6.1	7.0	100.0
Total	**209471**	**88.2**	**100.0**	
No respuesta	28109	11.8		
Total	237580	100.0		

Cuadro 3.3. Tabla de frecuencias de las horas a la semana dedicadas al estudio fuera del horario escolar

Horas de estudio a la semana	Puntuación media	Desviación estándar
No respuesta	44.3	13.8
0	44.5	13.3
1	43.1	12.4
2	44.0	12.7
3	45.6	13.1
4	47.7	13.4
5	49.0	13.6
6	50.1	13.6
7	50.6	13.9
8	50.2	13.6
9	51.6	14.0
10	50.8	13.8
Más de 10	52.6	14.2

Cuadro 3.4. Horas de estudio a la semana y desempeño bruto en la prueba

Recolectar los datos iniciales

Para la realización de este trabajo se consideró una base de datos derivada de la aplicación de una prueba estandarizada de selección de aspirantes a EMS, donde se registra el desempeño global; así como también para las diferentes asignaturas evaluadas. Además, se cuenta con los datos del cuestionario aplicado, que contiene información importante relacionada con aspectos socioeconómicos, hábitos de estudio, expectativas, etc.

Estos datos han sido cargados en el paquete estadístico Statistica versión 6.0 para su procesamiento con fines exploratorios. Posteriormente, después de disponer los datos relevantes en un formato adecuado, se cargaron al paquete HLM, para la construcción del modelo. En el anexo A se muestran algunas tablas de distribución de frecuencias para ilustrar instancias del proceso.

Descripción de los datos

La finalidad de esta tarea es describir los datos que se han adquirido, incluyendo: el formato, la cantidad; por ejemplo, el número de registros y campos en cada tabla, la identidades de los campos etc. En este trabajo se analizan los datos que provienen de:

1. La prueba estandarizada aplicada en el 2001 en el área metropolitana de la Ciudad de México.
2. El cuestionario de datos personales y socioeconómicos, conocido como hoja de registro, que deben responder los sustentantes.

La base de datos original consta de 237,580 registros con 146 campos para cada registro. En el anexo C se presenta la descripción de la estructura de la base de datos original.

Exploración de los datos

La tarea de exploración de los datos tiene como objetivo abordar las preguntas enfocadas a la MD, que puede lograrse empleando consultas, métodos de visualización y reportes. Estos pueden incluir: la distribución de atributos claves, por ejemplo, el atributo objetivo (variable dependiente) en una tarea de predicción; relaciones entre pares o números pequeños de atributos; resultados de agregaciones simples, etcétera.

Análisis de los niveles de desempeño global según características sociofamiliares

Basados en los enfoques teóricos que necesariamente deben tomarse en cuenta en este tipo de análisis, se decidió considerar los indicadores relacionados con los aspectos económicos y culturales de los hogares de origen del estudiante. En términos generales, son los indicadores considerados como representativos de las características de los hogares de origen de los estudiantes (características sociofamiliares) los que discriminan más en el desempeño y presentan el mismo sentido de la relación con el logro en la prueba que en otras investigaciones similares [10][11][12].

Por ejemplo, la escolaridad de la madre es la que genera la mayor discriminación en el logro, es decir, a mayor escolaridad mayor puntuación en la prueba (con una diferencia promedio de 20 respuestas correctas entre las categorías extremas), seguido del ingreso familiar mensual, que presenta también una relación positiva, a mayor ingreso familiar mayor puntuación en la prueba (con una diferencia promedio de 15 respuestas correctas entre las categorías extremas). El resto de los indicadores, como es el caso del hábito de lectura, que también muestra una relación positiva con el logro, tiene una diferencia promedio de 7 respuestas correctas; el nivel de hacinamiento en el hogar, presenta una relación negativa con el logro (a mayor número de personas viviendo en casa, le

corresponde una menor puntuación) con una diferencia promedio de 5 respuestas correctas. Aunado a lo anterior, se considera también la variable género del sustentante, dada la estabilidad de sus efectos en otras investigaciones. A continuación se presenta el cuadro que resume lo descrito anteriormente :

Variable	Categoría más desfavorable	Categoría más favorable	Diferencia en el puntaje
Nivel educativo materno	**No sabe leer ni escribir**	**Posgrado**	
	53.1	73.0	19.9
Ingreso familiar mensual	**Menos de $1,000**	**$20001 ó más**	
	52.6	67.9	15.3
Hábito de lectura	**Lee 3 hrs. o menos**	**Lee más de 3 hrs.**	
	58.8	65.9	7.1
Nivel de hacinamiento (número de personas que viven en su casa)	**Más de 4 personas**	**4 personas o menos**	
	56.6	61.9	5.3
	Mujer	**Hombre**	
Sexo	58.4	63.6	5.2

Cuadro 3.5. Promedio de respuestas correctas de acuerdo a las categorías extremas de las variables seleccionadas

Otras variables de interés

Entre las variables restantes, de las que presentan una relación positiva con el logro académico, es decir, a mayor valor de la variable le corresponde un mejor desempeño, ordenadas de acuerdo al nivel de discriminación entre las categorías extremas, son:

Variable	Muy desfavorable	Muy favorable	Diferencia en respuestas correctas
Los padres promueven que tomen sus propias decisiones sobre lo que pasa en la escuela	54.3	64.2	9.9
Horas de estudio a la semana	57.5	64.6	7.1
Los padres respetan sus opiniones sobre lo que ocurre en la escuela	57.3	62.9	5.6
Hábito de estudio: Estudiar principalmente con los apuntes de clase	57.3	62.6	5.3
Intenciones de estudiar la ES	56.5	61.7	5.2
Involucramiento de los padres en la actividad escolar Me apoyan cuando tengo algún problema en la escuela	59.1	62.2	3.1
Calificación de la formación que recibieron de la escuela	58.7	61.1	2.4
Los padres los felicitan o premian cuando les va bien en la escuela	58.2	63.6	5.4
Utilizar enciclopedias, diccionarios y atlas	57.4	62.8	5.4

Cuadro 3.6. Variables relacionadas positivamente con el logro académico

Otras variables presentan una relación negativa con el logro (a mayor valor de la variable, menor desempeño en la prueba), de manera semejante, ordenadas de acuerdo al nivel de discriminación, son:

Variable	Muy desfavorable	Muy favorable	Diferencia en el puntaje
Número de hermanos	55.2	64.9	9.8
Utilizar monografías que venden en las papelerías	55.9	64.6	8.7
Estudiar principalmente con los apuntes de los compañeros	53.7	61.9	8.2
Estudiar en equipo con sus compañeros de clase	55.7	63.8	8.1
Los profesores califican injustamente a sus alumnos	55.2	61.9	6.7
Lugar que ocupa entre sus hermanos	56.9	63.3	6.4
Los profesores organizan la clase tomando en cuenta la opinión e intereses de los alumnos.	57.7	63.9	6.2
Los profesores castigan injustamente a sus alumnos	55.9	61.7	5.8
Presentación de exámenes extraordinarios	56.5	61.7	5.2
Solicitar asesoría a sus maestros	57.9	62.8	4.9
Solicitar apoyo a sus padres o hermanos	58.1	62.5	4.4
Los profesores promueven en clase un ambiente amistoso y de confianza	59.7	61.5	1.8

Cuadro 3.7. Variables relacionadas negativamente con el logro en la prueba

Verificación de la calidad de los datos

Es importante señalar que la base de datos original consta de 237,580 registros, donde 25,881 de los mismos contienen nula información del cuestionario o más del 80% de datos ausentes, por lo que de entrada se tomó la decisión de eliminar este subconjunto de datos de análisis posteriores debido a que se consideró presentaban un alto porcentaje de no respuesta.

En términos generales, los datos correspondientes a la hoja de registro captaron el 86.4% de respuestas válidas y no se detectaron valores erróneos o fuera de rango en los datos.

3.3 Preparación de los datos

Selección de los datos

Con la finalidad de ajustar el desempeño bruto de los aspirantes y realizar comparaciones en igualdad de condiciones entre las escuelas, la selección de las variables para ajustar el modelo, se realiza a partir de la revisión del estado del arte en la investigación socioeducativa relacionada con este tipo de estudios y en la evidencia empírica presente en los datos disponibles, considerando como variables relevantes las características de los hogares de origen, en términos culturales y económicos. Aunado a lo anterior, se considerará también la variable género porque es otra variable que ha mostrado tener efectos significativos en el logro, lo que de alguna manera evidencia que al considerarla no se estarían

modelando fluctuaciones aleatorias fugaces, sino una verdadera estructura subyacente.

En lo que respecta a los registros, el criterio a seguir es considerar la totalidad de los casos que tienen respuestas válidas en las variables seleccionadas para realizar el ajuste y aquellas escuelas donde se tengan 30 ó más estudiantes disponibles para el análisis. Después de aplicar estos criterios para obtener el subconjunto de datos, objeto de análisis, se exploró que las distribuciones de frecuencias de las variables de interés no se alteraran de manera significativa, es decir, que se conservara semejante la representatividad de las categorías para cada variable.

Nombre de la variable	Estado de la variable
Nivel educativo materno	Construcción de un índice mediante Análisis factorial que represente los conceptos de capital cultural y económico de la familia del aspirante [6].
Hábitos de lectura	
Ingreso familiar mensual	
Nivel de hacinamiento (número de personas que viven en su casa)	
Sexo	Variable dicotómica, donde 0=mujer, 1=hombre

Cuadro 3.8. Variables seleccionadas para ajustar el desempeño bruto

Limpieza de los datos

El tratamiento de las siguientes variables se hizo con la finalidad de describir posteriormente cada una de las tipologías de escuelas que se obtengan en términos de estas variables y ver si existen algunas diferencias que puedan aportar conocimiento valioso sobre el tipo de estudiante que atienden principalmente estas escuelas. El tratamiento específico que se aplicó a cada variable de interés se describe a continuación:

1. Se dicotomizaron algunas variables en función del nivel de discriminación en el logro en la prueba estandarizada y en otras se agruparon algunas categorías en un intento por construir otras que tuvieran cierta importancia práctica para el estudio.

Variable	Código	Descripción de los códigos
Horas de estudio a la semana fuera del horario escolar	0 1	Tres horas o menos de estudio Más de tres horas de estudio
Presentación de exámenes extraordinarios	0 1	Sí presentó extraordinarios No presentó extraordinarios
Repetición de año escolar en la secundaria	0 1	Sí repitió No repitió
Gusto por las asignaturas	0 1	Categorías agrupadas: Me gustó poco y no me gustó. Categorías agrupadas: Me gustó mucho y no me gustó.
Preparación en las asignaturas	0 1	Categorías agrupadas: muy mala y mala. Categorías agrupadas: Muy buena y buena.

Variable	Código	Descripción de los códigos
Intención de seguir estudios de ES	0 1	Categorías agrupadas: No y no sabe. Sí tiene intención de seguir estudiando la ES.
Instituciones de ES a las que les gustaría ingresar	1 2 3	Normal, Inst. Tecnológico(SEP), UPN, U.Tecnológica, otras. UAM, UNAM, U.P.Estatal IPN, Universidad Privada
Actividades que más les gusta realizar	0 1	Entretenimiento (Deporte, cine, pasear y platicar con amigos, etc.) Desarrollo de habilidades (Jugar juegos de mesa, leer, otra)
Calificación de la formación que recibieron	0 1	Categorías agrupadas: Mala y deficiente. Categorías agrupadas: Excelente y buena.
Cuántos hermanos tiene	0 1	Más de 1 hermano. Hijo único ó 1 hermano.
Lugar que ocupan entre sus hermanos	0 1	2º, 3º, etc. 1º
Edad de los padres	0 1	Categorías agrupadas: 30 o menos y más de 60 años. Categorías agrupadas: Entre 31 y 60 años.
Con quién vive actualmente	0 1	Ambos padres y hermanos. Otro caso
Ocupación de los padres	1(Baja) 2(Media) 3(Alta)	No trabaja, labores del hogar, construcción, serv. personales, comerciante, trabajador en servicios personales, trabajador en oficios o por su cuenta. Jubilado o pensionado, obrero, empleado técnico o administrativo. Directivo o funcionario, ejercicio de la profesión por su cuenta, empleado en el ámbito profesional.

Cuadro 3.9. Preparación de las variables

2. Se transformó la variable respuesta (logro en la prueba) a una puntación en el rango de 0 a 10, utilizando la normalización min-max, con la siguiente fórmula:

$$V' = \frac{v - \min_A}{\max_A - \min_A}\left(nuevo\_\max_A - nuevo\_\min_A\right) + nuevo\_\min_A$$

   La normalización min-max realiza una transformación lineal en los datos y mantiene las relaciones entre los valores de los datos originales.

3. Para efectos de este estudio, no se contemplaron las escuelas con menos de 30 estudiantes y aquellos registros que no contienen información en las variables que se seleccionaron como factores relevantes al nivel del aspirante para realizar los ajustes en el desempeño bruto.

Construcción de nuevos datos

Índice sociofamiliar

Con el objetivo de obtener un índice sintético que represente las características sociofamiliares de cada aspirante, se realizó un Análisis Factorial utilizando como método de extracción de factores los Componentes Principales.

Considerando el estudio [10], la hipótesis con la que se analizó la extracción de factores fue que debía obtenerse una estructura factorial con dos dimensiones: una que representara las variables de capital cultural y la otra las de capital económico junto con el hacinamiento. La hipótesis de dos dimensiones se vio refutada debido a que ambas especies de capital en nuestros datos también parecen estar bastante correlacionadas.

	Peso en el factor de capital familiar global	Comunalidad
Escolaridad de la madre	.81	.65
Ingreso familiar mensual	.77	.59
Hábitos de lectura	.34	.11
Número de personas que habitan en la casa	-.43	.19
Método de extracción: Análisis de Componentes Principales. 1 componente extraído.		
Porcentaje de varianza explicada=38.4%		

Cuadro 3.10. Análisis Factorial de las variables de estratificación sociofamiliar matriz de componentes rotados, comunalidades y test de KMO

Como puede observarse en el cuadro anterior, la solución factorial obtenida es unifactorial, porque los dos indicadores más representativos del aspecto cultural y económico tienen las cargas más altas en el primer factor, lo que para nuestro estudio también refuta la hipótesis de partida e indica un espacio social en el cual los diversos campos conceptuales están fuertemente correlacionados. Las correlaciones entre las variables ingresadas son lo suficientemente fuertes como para que la técnica extraiga un solo factor que resume la posición de clase de las familias. La medida que se obtiene es por tanto un índice que refleja el volumen de capital de la familia del aspirante en dos de las dimensiones más importantes: económica y cultural [10].

Las variables que tienen mayor peso en el factor son la escolaridad de la madre y el ingreso familiar, las dos con un sentido positivo: cuanto mayor es la escolaridad de la madre y los ingresos, más alto es el índice sociofamiliar. Para el caso de los hábitos de lectura, la correlación con el factor es menor pero positiva.

El número de personas que habitan en la casa del aspirante está relacionado negativamente con el factor, lo que indica que un nivel más alto de índice sociofamiliar está acompañado de una disminución del número de personas en la casa.

Del cuadro 3.10 se observa que todas las cargas factoriales son mayores que ± 0.30, lo que se considera está en el nivel mínimo aceptable. Así, cuanto mayor sea el tamaño absoluto de la carga factorial, más importante resulta la carga al interpretar la matriz factorial. Dado que la carga factorial es la correlación entre la variable y el factor, el cuadrado de la carga es la cuantía de la varianza total de la variable de la que da cuenta el factor; de esta forma, el factor extraído explica más del 50% de la varianza de las dos primeras variables y un porcentaje más reducido de las restantes, pero se decidió conservar estas últimas dados los sólidos fundamentos teóricos disponibles [10].

Finalmente, el porcentaje de varianza explicada por el análisis es más reducido del que corresponde a una buena solución factorial ya que se reduce al 38% de lo observado.

Índice de hábitos de estudio

Para representar los hábitos de estudio en un solo índice, se construyó una escala sumada con los indicadores que representan los hábitos de estudio de los estudiantes. Posteriormente, se normalizó esta escala en un rango de 0 a 10 con el método de normalización *min-max*. Para determinar qué elementos incluir en la escala, se realizó un Análisis de Confiabilidad de todos los elementos y como resultado se obtuvo que la eliminación de tres de ellos aumentaba su confiabilidad. En el cuadro 3.11 se presentan los indicadores incluidos en la escala.

Indicadores de hábitos de estudio
Al iniciar, identifico lo que necesito estudiar y hago un plan de trabajo
Estudio principalmente con mis apuntes de clase
Estudio principalmente con el libro de texto de la materia.
Utilizo enciclopedias, diccionarios y atlas
Realizo resúmenes y/o cuadros sinópticos
Resuelvo ejercicios para reafirmar lo estudiado
Solicito apoyo a mis padres o hermanos
Solicito asesoría a mis maestros
(Coeficiente Alfa de Cronbach estandarizado=0.65)[2]

Cuadro 3.11. Indicadores que conforman el índice de hábitos de estudio

Cabe mencionar que este índice de hábitos no presenta ni siquiera una correlación moderada con la variable respuesta (r=0.06), esto se debe a que buena parte de los elementos que conforman la escala no contribuyen a discriminar entre los buenos y malos resultados en la prueba, por lo que se recomienda que una vez obtenidas las tipologías de escuelas, se exploren, de manera individual aquellos indicadores que parecen estar más relacionados positivamente con los resultados en la prueba (Ver cuadro 3.6).

Índice del nivel de involucramiento de los padres en la actividad escolar

De manera semejante al caso anterior, para el grupo de indicadores correspondientes al nivel de involucramiento de los padres en la actividad escolar, se construyó otra escala sumada considerando:

Forma de actuar de los padres respecto a la actividad escolar
Me ayudan en mis tareas escolares
Me felicitan o premian cuando me va bien en la escuela
Me apoyan cuando tengo algún problema en la escuela
Respetan mis opiniones sobre lo que ocurre en la escuela
Promueven que tome mis propias decisiones sobre lo que pasa en la escuela (Coeficiente Alfa de Cronbach estandarizado=0.68)

Cuadro 3.12. Indicadores que conforman el índice del nivel de involucramiento de los padres

[2] El coeficiente Alfa de Cronbach es un estadístico comúnmente utilizado para informar qué tan confiable es la medida que se ha construido a partir de la combinación de las variables de origen. Su rango varía de 0 a 1. Los valores más altos indican mayor confiabilidad y los más bajos menor o nula confiabilidad.

Este índice correlaciona de manera positiva con el logro en el examen, es decir, a mayor involucramiento de los padres mejores resultados en la prueba, aunque dicha correlación es baja (r=0.17). De manera semejante al caso de la escala de hábitos, se hace la recomendación de explorar las tipologías de escuelas con los indicadores más discriminantes de manera individual.

Índice de las formas de actuar de los profesores en el salón de clase

El grupo de indicadores relacionados con el comportamiento o formas de actuar de los profesores se trató de manera similar a los anteriores y se consideraron los siguientes indicadores:

Forma de actuar de los maestros en clase
Se dedican la mayor parte del tiempo de la clase a trabajar con los alumnos
Se esfuerzan para que los alumnos comprendan lo tratado en clase
Ayudan a los alumnos en el desarrollo de sus trabajos en clase Realizan evaluaciones regularmente Promueven el trabajo en equipo entre los alumnos
Organizan la clase considerando la opinión e intereses de los alumnos
Promueven la participación de todos los alumnos durante la clase
Promueven en clase un ambiente amistoso y de confianza
(Coeficiente Alfa de Cronbach estandarizado=.74)

Cuadro 3.13. Indicadores que conforman el índice de comportamiento de los profesores

Este índice presenta una correlación negativa y casi nula con el logro en la prueba (r= - 0.04).

Índice de alimentación

Finalmente, en lo que respecta a la frecuencia con que se consumen determinados tipos de alimentos, se construyó igualmente una escala sumada, que posteriormente se normalizó en un rango de 0-10 con el método *min-max*. Aquí se utilizaron todos los elementos de la escala. La correlación entre el índice de alimentación y el logro en la prueba es positiva pero baja (r=0.16).

Código de escuela

Con el propósito de disponer los datos de manera adecuada para la construcción del modelo con el paquete HLM, se generó una variable nueva cuya composición es más simple (un número consecutivo y único) comparado con la clave original de la escuela, este código servirá para ligar los archivos correspondientes a los dos niveles de análisis del modelo multinivel. El requisito que debe cumplir el código de identificación de la escuela es que debe tener la misma amplitud para todas las escuelas y los archivos deben ordenarse con respecto a esta clave única. El siguiente es un pequeño ejemplo de un archivo de nivel-2 (nivel de escuelas), con el código de escuela ordenado y una variable composicional “indiceprom”:

Id_escuela	indiceprom
0001	2.45
0002	-0.70
. .	
1371	1.80

Generación de archivos para la modelación

La preparación y generación de los archivos indispensables para la construcción del modelo con el paquete HLM, se realizó en el paquete estadístico SPSS, debido a que HLM no contempla este tipo de facilidades para la preparación de los datos. A continuación, se describe el procedimiento:

1. Se generó un primer archivo con los datos de los aspirantes (denominado archivo de nivel-1), en términos de las variables seleccionadas como relevantes para realizar el ajuste del desempeño bruto; así como también el código de escuela correspondiente a cada aspirante.
2. Se generó un segundo archivo con el código de la escuela y la variable generada como producto de la contextualización en términos de las características sociofamilares de los aspirantes (denominado archivo de nivel-2). Para obtener esta variable composicional, se consideró el promedio del índice de capital familiar global de los egresados de cada escuela, es decir, cada escuela cuenta con un índice sociofamiliar promedio.

Formateo de los datos

Los dos archivos creados de nivel-1 y nivel-2 se ordenaron con respecto al código creado para la escuela ya que es un requisito indispensable del paquete HLM para la importación de archivos y la generación de estadísticas suficientes. El paquete únicamente reconoce archivos ASCII y otros formatos de paquetes estadísticos como son: SPSS, SYSTAT y SAS. La siguiente información corresponde a un ejemplo de archivo de estadísticas suficientes:

```
                  LEVEL-1 DESCRIPTIVE STATISTICS
 VARIABLE NAME        N       MEAN         SD         MINIMUM      MAXIMUM
      SEXO        160432       0.50       0.50          0.00         1.00
    GLOBAL        160432       4.77       1.35          0.30         9.60
    INDSOC        160432      -0.00       1.00         -2.36         4.17

                  LEVEL-2 DESCRIPTIVE STATISTICS
 VARIABLE NAME        N       MEAN         SD         MINIMUM      MAXIMUM
  MEDIAIND          1371      -0.07       0.48         -1.03         2.72
```

Capítulo 4. Modelación y resultados

4.1 Modelación

Selección de la técnica de modelación

Antecedentes

Después de quince años de discusión sobre el método óptimo, hoy es generalmente aceptado que para abordar satisfactoriamente este tipo de estudios, sea solo para propósitos de realizar ajustes en el desempeño y derivar clasificaciones de escuelas o para los estudios de eficacia escolar, se requieren modelos estadísticos multinivel o jerárquico-lineales [19].

El uso de este modelo requiere de un proceso de decisiones de especificación fuertemente orientado por la teoría, en el caso particular de los estudios de escuelas eficaces, por la propia conceptualización de escuela eficaz que se tenga. De hecho, las diferencias más importantes entre los distintos modelos que se pueden ajustar radican en el supuesto teórico que tienen por detrás, tal como se mostrará a continuación.

En términos generales, estos modelos trabajan con información del alumno y de la escuela sin fusionar o suprimir ambos niveles (sin agregar o desagregar). De esta manera, la modelación permite mantener la distinción teórica entre lo macro (la escuela y su entorno) y lo micro (los procesos individuales) sobre el nivel de aprendizaje de un alumno. Lo cual abre nuevas potencialidades al estudio de la eficacia escolar. Analíticamente se pueden ajustar ecuaciones de factores determinantes por separado, alcanzando una comprensión más elaborada de cómo y sobre qué tiene incidencia la organización escolar. Una escuela puede ser clasificada como "eficaz" en dimensiones que antes no eran posibles de hacerse; por ejemplo, introduciendo el requisito combinado de Edmonds [6].

Descripción de la técnica de Modelación Multinivel

El análisis Multinivel es una técnica estadística para estudiar datos que tienen patrones de variabilidad muy complejos, como son los datos de fuentes anidadas, es decir, con estructura de agrupamiento jerárquico. Es un método ideal para analizar datos de estudiantes agrupados- en salones de clase, salones agrupados en escuelas, de escuelas anidadas en ciudades, etcétera. Cuando se analizan datos de este tipo, es importante considerar que no sólo existe variabilidad entre los estudiantes, sino también entre los salones de clase, escuelas, ciudades, etc. Se puede llegar a conclusiones equivocadas si se ignoran estas fuentes de variabilidad.

El principal modelo de análisis multinivel es el modelo jerárquico lineal, que es una extensión del modelo de regresión lineal múltiple, donde se incorporan coeficientes aleatorios a todos los niveles de la jerarquía. En el caso de los datos de los estudiantes, tanto éstos como los salones, las escuelas, las ciudades y los estados son tratados como variables aleatorias.

La idea fundamental en el modelo jerárquico es que la variabilidad se puede separar en variabilidad dentro de los grupos y variabilidad entre los grupos. En el caso de alumnos, escuelas, ciudades y estados, la varianza total de los datos –por ejemplo de las calificaciones de los alumnos- se puede separar en varianza entre estados, varianza entre ciudades, varianza entre escuelas y varianza dentro de las escuelas. Esta variabilidad dentro y entre los grupos se considera también como variable aleatoria. La diferencia que existe entre un modelo jerárquico lineal y uno de regresión lineal múltiple es que la ecuación que define al primero de ellos contiene términos de error (o residuales) a todos los niveles de la estructura de anidamiento.

El modelo jerárquico lineal más simple –llamado nulo – es aquel en el que se emplea una variable dependiente y no existen variables predictoras. Contiene grupos aleatorios y variación aleatoria dentro de los grupos, se puede expresar como un modelo donde la variable dependiente Y_{ij} es la suma de la media global β_0, más el efecto aleatorio (o residual) de cada grupo, u_{0j}, más el efecto aleatorio para cada individuo, e_{ij}. En caso de que se consideren dos niveles de la jerarquía, la ecuación que lo describe es:

$$Y_{ij} = \beta_0 + u_{0j} + e_{ij} \qquad (1)$$

Donde el índice i identifica a los individuos dentro de un grupo y el índice j identifica a los grupos.

Es un modelo con tres parámetros: β_0, u_{0j} y e_{ij}. Se asume que las variables aleatorias u_{0j}, e_{ij} tienen media 0, son independientes y tienen varianza:

$$Var(u_{0j}) = \tau_{00} \qquad (2)$$

$$Var(e_{ij}) = \sigma^2 \qquad (3)$$

La importancia de este modelo es que permite una primera división de la variabilidad de los datos en dos niveles: la varianza total de Y_{ij} se puede descomponer en la suma de las varianzas a los dos niveles:

$$Var(Y_{ij}) = Var(u_{0j}) + Var(e_{ij}) = \tau_{00} + \sigma^2 \qquad (4)$$

La covarianza entre dos individuos distintos i e i' en un mismo grupo j es igual a la varianza de la contribución de u_{0j}, que es compartida por esos dos individuos:

$$Cov(Y_{ij}, Y_{i'j}) = var(u_{0j}) = \tau_{00} \qquad (5)$$

Y su correlación es:

$$\rho(Y_{ij}, Y_{i'j}) = \tau_{00} / (\tau_{00} + \sigma^2) \qquad (6)$$

Este parámetro es el coeficiente de correlación intra-clase. Se puede interpretar de dos maneras: es la correlación entre dos individuos tomados al azar dentro de un mismo grupo tomado al azar, y también es la fracción de la varianza total que se debe al nivel de grupo. El coeficiente de correlación intra-clase es de suma importancia, pues permite saber si los datos en cuestión son susceptibles de análisis multinivel. Si existe una fracción de varianza significativa a nivel de grupo, el análisis amerita un enfoque multinivel. Si no existe, el análisis se puede realizar con regresión lineal ordinaria. En otras palabras, si τ_{00}

es significativamente mayor que 0, el modelo jerárquico lineal sería un modelo más apropiado para el análisis de los datos que la regresión lineal ordinaria.

Razones para ajustar un modelo multinivel

Aunque su uso más destacado ha sido para identificar el objetivo más general de conocer cuáles son las características de la organización escolar que inciden sobre los aprendizajes, un enfoque un poco menos complejo radica en su aplicación para realizar ajustes, considerando características de entrada relevantes de los estudiantes (obviamente, relacionadas con el logro), en el desempeño bruto y construir una clasificación de las escuelas para efectos de compararlas más equitativamente. Su utilización en el campo de los estudios de eficacia escolar, tiene grandes ventajas. Entre las razones señaladas por la bibliografía, se pueden señalar las siguientes cuatro:

- Con un modelo multinivel es posible especificar de forma correcta y completa, la compleja interdependencia entre las características sociofamiliares individuales de los alumnos y las características contextuales. Al hacer posible un modelo que considera ambas unidades de análisis, se evitan dos sesgos que se cometen al usar el método de mínimos cuadrados que son: a) desconocer los agrupamientos de alumno y escuela; y b) suprimir la heterogeneidad del alumnado como cuando se "agregan los datos" al nivel de escuela.
- Desglosa la variación del aprendizaje, en la porción que corresponde a los atributos individuales del alumno y la porción atribuible a la escuela, cuantificación que responde a la pregunta más general sobre cuánto importa la escuela. Esta distinción levanta la restricción que impone el supuesto de ausencia de auto-correlación en el modelo de regresión lineal múltiple. Lo cual resulta congruente con la teoría; esto es que los niños aprenden en procesos grupales y que por lo tanto se espera que los aprendizajes de los alumnos de un mismo grupo sean semejantes en alguna proporción.
- La descomposición de la varianza en los aprendizajes permite representar más adecuadamente la parte no sistemática del modelo. En el método de mínimos cuadrados, se especifica una parte sistemática donde se incluyen todos los términos relativos a las variables sociofamliares y una parte aleatoria con un único término de error, sea a nivel de la escuela, sea a nivel del alumno. Se mostró ya que este término de error no tiene interpretación directa cuando se ajusta al nivel de los alumnos, porque es una mezcla no distinguible entre los residuos generados a nivel individual y aquellos generados al nivel de las escuelas. Este problema es resuelto en el modelo multinivel, incluyendo en su especificación mínima, representar dos términos de error en la ecuación.
- El uso del modelo permite establecer si la escuela tiene efectos diferenciales sobre sus alumnos dependiendo de cuáles sean sus

características sociofamiliares, étnicas o de género. En comparación, el método de mínimos cuadrados hace el supuesto de que los efectos del capital cultural, por ejemplo, son homogéneos y constantes a través de todas las escuelas. Al levantarse tal restricción, se conquista una nueva dimensión para el análisis empírico que está presente en el enfoque "combinado de calidad y equidad". Con esta modelación se está en condiciones de identificar si una escuela distribuye los aprendizajes entre sus estudiantes de forma más igualitaria, además de generar un "piso común" de aprendizajes más alto. Como se verá a continuación, no es frecuente que este tipo de consideraciones sea incluido en los análisis.

Formas en que ha sido aplicado en análisis multinivel

1. La estimación multinivel bajo el supuesto de uniformidad

Los conceptos de "eficacia relativa" y de "eficacia longitudinal" no consideran como nota esencial de una escuela eficaz la modalidad con que los conocimientos se distribuyen socialmente dentro del alumnado de una escuela. En ese sentido, se puede asumir el supuesto de que los efectos de la escuela son constantes para todo su alumnado y que las únicas diferencias entre las mismas radican en el promedio de aprendizajes o "piso común" que generan para sus estudiantes.

Si se adopta la notación clásica de los modelos multinivel propuesta por [35], lo cual implica escribir las ecuaciones respectivas para cada nivel, este modelo de efectos uniformes se expresaría así:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij} \quad (7)$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01} C_j + u_{0j} \quad (8)$$

$$\beta_{1j} = \gamma_{10} \quad (9)$$

Donde:

Y_{ij} es el nivel de aprendizaje del i-ésimo alumno en la j-ésima escuela

β_{0j} es el promedio de aprendizajes en la j-ésima escuela

γ_{00} es el nivel de aprendizajes promedio para todos los grupos de la muestra analizada

X_{ij} es un vector de características sociofamiliares del i-ésimo alumno

C_j es un vector de características del contexto de la escuela

β_{1j} es el efecto del vector X estimado para la j-ésima escuela.

γ_{10} es el efecto del vector X estimado a través de todas las escuelas de la muestra

γ_{01} es el efecto del vector C sobre el promedio de la escuela

e_{ij} es la diferencia entre el estimado y el observado para el estudiante

u_{0j} es el efecto único de la escuela sobre los aprendizajes.

La uniformidad de los efectos de la escuela se traduce en "fijar" el coeficiente que representa cómo se distribuyen socialmente los aprendizajes dentro de cada escuela. La variabilidad entre escuelas queda restringida a la diferencia entre los promedios o pisos comunes de aprendizaje que cada escuela proporciona a su alumnado. De acuerdo a esta especificación, el efecto de la escuela sobre los aprendizajes se identifica directamente despejando u_{0j} de la ecuación (8):

$$u_{0j} = \beta_{0j} - \gamma_{00} - \gamma_{01} C_j \qquad (10)$$

Si se desea también se puede expresar este modelo de efectos uniformes mediante la ecuación combinada, tal como lo hacen en [34]:

$$Y_{ij} = \gamma_{00} + \gamma_{01} C_j + \gamma_{10j} X_{ij} + e_{ij} + u_{0j} \qquad (11)$$

Debe notarse que el modelo puede extenderse para que X_{ij} pueda representar un vector amplio de características individuales, tales como el sexo, la condición de actividad laboral, aspiraciones educacionales, disposiciones culturales, auto-valoraciones, etc. Del mismo modo, se pueden incluir dentro de las características contextuales todo el conjunto de características del entorno de la escuela sobre los que no se puede teóricamente suponer que sean manipulables por la gestión de la escuela.

2. El modelo de distribución social del conocimiento

Es razonable suponer y así se ha discutido mucho durante los años noventa, que las escuelas eficaces no son eficaces de la misma forma para cualquier tipo de estudiantado. Podría resultar que los varones aprovecharan más que las mujeres los efectos de una escuela eficaz; que los alumnos de alto capital más que los de bajo capital cultural; que los blancos más que los indígenas, y así sucesivamente. Una situación idónea sería que la escuela fuese eficaz para todo tipo de alumnado, que lograra compensar las diferencias iniciales y tener un efecto semejante en los mismos, a este enfoque se le ha denominado en la literatura como la exploración de la "distribución social del conocimiento" [12].

En consecuencia, conjuntamente al ajustar el promedio o "piso común" de aprendizajes para la j-ésima escuela, podría la investigación interesarse en cómo se distribuyen los aprendizajes dentro de la escuela y establecer por ejemplo, si las desigualdades de clase social se incrementan, se mantienen iguales o disminuyen. Tal preocupación se especifica permitiendo que el efecto de clase social sobre el aprendizaje varíe entre las escuelas. Formalmente se agrega un término a la ecuación (7) para obtener:

$$Y_{ij} = \gamma_{00} + C_j \gamma_{01} + X_{ij} \gamma_{10j} + e_{ij} + u_{0j} + u_{1j} * X_{ij} \qquad (12)$$

Donde:

$u_{1j} * X_{ij}$ es el efecto variable entre escuelas de los antecedentes sociofamiliares.

Observe que la anterior especificación de los efectos de la escuela ha ampliado conceptualmente su rango de forma tal que ahora ε_j (parte aleatoria) ha sido representado por la adición de dos términos:

$$\varepsilon_j = u_{0j} + u_{1j} * X_{ij} \quad (13)$$

Estimadores Bayesianos

Desde el punto de vista estadístico existe una ventaja adicional a favor de este método de modelación y de los distintos modelos: la introducción de los estimadores Bayesianos. Esto permite corregir de forma más eficiente las peculiaridades que se observan en las distribuciones heterocedásticas de los residuos cuando son generadas por datos agregados con distintos tamaños de muestra de alumnos evaluados por escuela. Tal como lo señalan en [39]:

> "Los residuos empíricos de Bayes estimados bajo un modelo jerárquico lineal proveen de un indicador estable para juzgar el desempeño de una escuela individual. Estos estimadores empíricos bayesianos tienen ventajas distintivas sobre los métodos previos. Aquellos a) toman en cuenta la pertenencia a los grupos [escuelas] aún cuando el número de grupos sea muy grande, y b) producen estimadores relativamente estables aún cuando los tamaños muestrales por escuela sean modestos".

Los paquetes estadísticos (HLM, MLwin) utilizan estimadores bayesianos para todos los parámetros y por ende, u_{0j}, aunque por ejemplo, se pueda disponer de estimaciones mínimo-cuadráticas para los primeros pasos del análisis y realizar comparaciones apropiadas.

La propiedad de este algoritmo es que opera reduciendo o restringiendo la varianza de u_{0j}, de aquí que se le denominen estimadores reducidos ("*shrunken residuals*", [19]). La reducción se aplica al estimador mínimo cuadrático en forma inversamente proporcional a la falta de plausibilidad del valor observado en los residuos. Cuanto más confiable es el estimador OLS, menor es la reducción observada. En el extremo, los valores que han sido generados por muestras muy pequeñas de alumnos, son reducidos a cero. El supuesto por detrás es que los residuos de las escuelas siguen una distribución en el universo que debe ser estimada a partir de los estadísticos provistos por los promedios de las escuelas. Si dichos estadísticos provienen de muestras pequeñas, entonces es razonable que la estimación sea poco confiable. Específicamente, el estimador de los efectos de las escuelas bajo un modelo jerárquico-lineal es:

$$u^*_{0j} = \lambda_j u_{0j} \quad (14)$$

$$\lambda_j = \tau_{00} / (\tau_{00} + \sigma^2 / n_j) \quad (15)$$

Donde:

τ_{00} es la varianza entre escuelas con respecto a la variable dependiente

σ^2 es la varianza (homocedástica) a nivel de los alumnos

n_j es el número de alumnos de la j-ésima escuela

Sin embargo, estas estimaciones no están libres de un potencial sesgo que puede denominarse "profecía autocumplida" [39]. Tal como se ha mostrado, por definición, la falta de confiabilidad del efecto de la escuela provoca que el estimador tienda a acercarse al valor esperado con base a las características del

alumnado. Por ejemplo, una escuela inusualmente eficaz que atiende niños con origen social muy bajo, tendrá una estimación sesgada hacia la media global, hacia el valor que típicamente muestran otras escuelas con un alumnado similar desde el punto de vista sociofamiliar. Si este fuera el caso, el procedimiento estadístico operaría una suerte de profecía estadística autocumplida, en el que los efectos de la escuela serían subestimados para algunas escuelas.

Generación de pruebas para el modelo.

La finalidad de este estudio es la construcción de tipologías de escuelas considerando el Desempeño Neto (DN) de sus egresados y la descripción de las mismas en términos de otras variables relevantes, para explorar si existen patrones interesantes. Como se ha mencionado a lo largo de este trabajo, la inclusión de las variables consideradas en el ajuste del modelo se basó en los resultados obtenidos en investigaciones realizadas a nivel nacional e internacional; así como de la evidencia empírica presente en los datos, que revelan que son precisamente las características sociofamiliares las que muestran un efecto mayor en el logro.

Las tendencias actuales de producción de clasificaciones de escuelas (rankings), de acuerdo al desempeño de sus estudiantes en pruebas estandarizadas, muestran un claro avance en el sentido de considerar características relevantes del origen de los estudiantes con la finalidad de ajustar estos desempeños brutos y realizar comparaciones en igualdad de condiciones. Sin duda, la acumulación de varias aplicaciones de estas pruebas estandarizadas y la realización continua de este tipo de análisis, revelaría la estabilidad de los resultados obtenidos por las escuelas a lo largo del tiempo y permitiría establecer mecanismos de monitoreo continuos y permanentes. Esto debido a que las escuelas son organizaciones que pueden verse afectadas por un sinnúmero de factores y esto a su vez impactar en el desempeño de sus estudiantes, de tal forma que haya cambios significativos en las clasificaciones de un año a otro.

Para efectos de establecer un diseño de prueba para demostrar formalmente los beneficios de la modelación multinivel, se pueden realizar dos pruebas: primero eliminar puntos de datos simples y revisar la predicción del ajuste del modelo para el resto de los datos. Después, se pueden eliminar grupos (escuelas) y realizar el mismo procedimiento. Para cada paso de validación cruzada, se comparan los estimadores multinivel [17].

En este estudio se realizó un diseño de prueba simple para los parámetros más importantes del modelo, que consistió en dividir el conjunto en dos partes con una proporción similar de escuelas. Para la selección de las escuelas de cada conjunto, se aplicó un muestreo aleatorio de escuelas con el paquete Statistica y del análisis de las dos muestras se obtuvieron los siguientes resultados:

Modelo		Muestra 1	Muestra 2	Conjunto completo
Componentes de varianza (Modelo Nulo). Variable dependiente: Logro en la prueba.	Nivel-1	86.6%	88.4%	87.5%
	Nivel-2	13.4%	11.6%	12.5%
Modelo ajustado	γ_{00}	4.467673	4.444638	4.456250
	γ_{10} (Sexo)	0.405850	0.418682	0.412191
	γ_{20} (indsoc)	0.255756	0.252075	0.253947
	Deviance	264935.4353	260010.0536	524927.90052
Componentes de varianza (Modelo Nulo) Variable dependiente: índice de sociofamiliar	Nivel-1	79.1%	80.0%	79.6%
	Nivel-2	20.9%	20.0%	20.4%

Cuadro 4.1 Estimadores de parámetros del modelo multinivel

Los resultados del cuadro 4.1 reflejan estimadores de parámetros muy semejantes para ambas muestras y a su vez también muy cercanos a los estimadores del conjunto completo de escuelas. Es importante mencionar que este diseño de prueba tiene limitaciones para probar los efectos mencionados en [17], pero dadas las restricciones de tiempo, se decidió considerarla para reflejar únicamente la estabilidad de los estimadores de parámetros más importantes del modelo.

Construcción del modelo

Análisis multinivel bajo el enfoque de uniformidad

En este trabajo, resulta muy importante considerar el conocimiento previo generado por investigaciones relacionadas realizadas en los países anglosajones, de donde son originarias estas perspectivas; así como también aquellas llevadas a cabo en América Latina, donde todavía se considera incipiente el desarrollo de este tipo de estudios. Después de revisar la vasta cantidad de investigaciones realizadas a nivel internacional, donde se ha mostrado que los factores que influyen significativamente en los aprendizajes y que se deben considerar como ajustes antes de realizar cualquier comparación entre escuelas o individuos, son precisamente aquellos que tienen que ver con las características sociofamiliares de los estudiantes.

De manera semejante, se han reportado diferencias en los niveles de logro por género, que no parecen ser fluctuaciones aleatorias fugaces, razón por la que se incluyó también como variable de ajuste en nuestro modelo. Debe recordarse que únicamente se consideran estas variables con la finalidad de ajustar los desempeños brutos y construir las tipologías de escuelas a partir del desempeño neto, bajo el supuesto de uniformidad comentado en secciones anteriores. Es importante mencionar en este punto, que el modelo utilizado como base para construir las tipologías de escuelas no contempla la variable composicional

producto de la contextualización del índice sociofamiliar. Este análisis será objeto de un apartado adicional donde se describirán los hallazgos relacionados con el efecto moderador que tiene esta variable, tanto para explicar la variación en los promedios de aprendizaje de las escuelas, como también aquellos relacionados con el efecto del género y del origen sociofamiliar.

A continuación, se presentan dos modelos nulos que fueron útiles para tomar las decisiones metodológicas más importantes en el análisis. La diferencia entre ambos modelos radica en la variable dependiente. El primero considera como variable dependiente el índice sociofamiliar y la finalidad del mismo es probar la homogeneidad intra-escuela, en términos del origen social de sus estudiantes, para tomar la decisión de la metodología que se usará para construir las tipologías. El segundo modelo tiene como objetivo separar la varianza correspondiente al nivel de los estudiantes (nivel-1) y la varianza al nivel de las escuelas (nivel-2); así como también justificar el uso de un Análisis Multinivel.

Análisis del supuesto de homogeneidad intra-escuela

En el método de contextualización por estratos familiares, uno de los supuestos teóricos que subyace a buena parte de las decisiones estadísticas que se toman, es que las escuelas son organizaciones que presentan niveles importantes de homogeneidad sociocultural en la composición de su alumnado. Si se contara con suficiente evidencia empírica para apoyar este supuesto, podría realizarse un Análisis Cluster de las escuelas y construir las tipologías a partir del índice que refleja precisamente las condiciones económicas y culturales de los hogares de origen previa agregación de este índice al nivel de escuela (contextualización).

Con la finalidad de probar este supuesto, se realizó una prueba estadística mediante la utilización de un modelo multinivel, en el que la variable dependiente es el índice sociofamiliar construido al nivel del aspirante y en el que el valor del coeficiente de correlación intra-clase (ρ) para el modelo vacío funge como estadístico de homogeneidad [10][39].

Fuente de variación	Porcentaje
Nivel de escuela	20.4%
Nivel individual	79.6%
Varianza total	100.0%

Cuadro 4.2. Análisis Multinivel HLM, "modelo vacío"

A partir del cuadro 4.2, es claro que el coeficiente de correlación es relativamente bajo y por tanto implica que dos individuos pertenecientes a una misma escuela de origen no son muy semejantes en cuanto a sus características sociofamiliares. Se esperaría que este supuesto de homogeneidad sociocultural en el alumnado fuera compatible en aquellas sociedades en las que el funcionamiento del sistema educativo y el patrón de asentamiento residencial de las clases sociales conlleve una fuerte diferenciación social entre escuelas. Por el contrario, resulta muy discutible su validez en las sociedades que han generado predominantemente escuelas policlasistas [10], lo que parece ser el caso de las escuelas analizadas.

Modelo de Componentes de Varianza para la variable respuesta: Logro académico

Este modelo considera como variable dependiente el logro en la prueba. Se contemplan dos niveles: el nivel 1, constituido por todos los individuos (sustentantes) y el nivel 2, formado por las escuelas.

Fuente de variación	Porcentaje
Nivel de escuela	12.5%
Nivel individual	87.5%
Varianza total	100.0%

Cuadro 4.3. Análisis Multinivel HLM, "modelo vacío"

El coeficiente de correlación intra-clase es relativamente bajo, pero significativamente diferente de 0, lo que indica que en el área metropolitana las escuelas son relativamente semejantes en términos del desempeño en la prueba de sus egresados. Además, es importante mencionar que este valor está considerado dentro del rango de valores obtenido en otros estudios internacionales similares [19].

Los estudios realizados a nivel internacional sugieren que aproximadamente entre 10% y 18% de la variación en los aprendizajes de los estudiantes se debe a las diferencias entre las escuelas. Una cantidad adicional de variación, aproximadamente del 50%, se atribuye a las diferencias entre los salones de clase dentro de las escuelas. Es decir, que aproximadamente el 60% de la variación en el desempeño de los estudiantes se encuentra entre las escuelas o salones de clase. El 40% restante se debe a características individuales de los estudiantes y al entorno del que provienen [2]. De lo anterior se desprende que del 87.5% atribuido a los estudiantes en este estudio, aproximadamente un 50% podría explicarse si se contara con información relevante al nivel del salón de clase.

En resumen, el análisis de los dos modelos anteriores proporciona el soporte necesario para la decisión de qué metodología utilizar para lograr los objetivos del estudio. En primera instancia, los datos revelan que las escuelas no son lo suficientemente homogéneas en cuanto a la composición sociocultural de su alumnado, por lo que no se recomienda utilizar únicamente la "agregación" (contextualización) del índice sociofamiliar al nivel de la escuela, debido a que se estaría escondiendo una importante diversidad que podría generar resultados engañosos. El segundo modelo nulo nos da la pauta para justificar nuestro análisis multinivel, es decir, tenemos un coeficiente de correlación intra-clase relativamente bajo, pero significativamente diferente de cero (12.5%).

Metodología para la construcción de las tipologías de escuelas

La construcción de las tipologías de escuelas se realizará a partir del ajuste de un modelo multinivel, que consiste de dos niveles. El nivel-1, correspondiente a los estudiantes y el nivel-2, conformado por las escuelas incluidas en el análisis. El análisis de las escuelas se desarrolla en cuatro etapas:

1. Sólo a manera de exploración, en esta primera etapa se establece el comportamiento de las escuelas sobre el Desempeño Bruto (DB) que se estima a partir de los puntajes crudos que obtienen los sustentantes agrupados por escuela de procedencia. Las escuelas fueron clasificadas en categorías, de acuerdo a su discrepancia con el desempeño promedio.
2. En la segunda etapa se incorpora un ajuste a estos puntajes, bajo el supuesto de uniformidad descrito en párrafos anteriores, considerando fijos los efectos del índice sociofamiliar y del género para determinar el DN. Nuevamente se clasifican las escuelas en categorías después de controlar el efecto de estas variables.
3. En esta etapa se describen las tipologías de escuelas obtenidas del paso 2, en términos de algunas características de sus estudiantes (hábitos de estudio, expectativas de continuar sus estudios de ES, involucramiento de los padres, etc.).
4. Adicionalmente, en esta etapa se presenta el análisis del modelo considerando el efecto de la variable composicional "índice sociofamiliar promedio" para explicar la variación del logro promedio de las escuelas, del efecto de la variable género y del origen social.

Descripción de resultados

1. Estimación del desempeño bruto (DB)

La ecuación (1) puede utilizarse para modelar el DB por escuela, donde Y_{ij} se sustituye por la calificación cruda del sustentante i de la escuela j.

De esta forma, es posible clasificar todas las escuelas j de acuerdo a la diferencia que hay entre la media general β_0 y la media estimada del grupo β_0+u_{0j} de manera equivalente, de acuerdo con su residual u_{0j}, aunque en realidad se produce una lista ordenada de los grupos; las escuelas, en este caso. El cuadro 3.17 contiene la clasificación de las escuelas y el número que cae dentro de cada categoría, considerando el DB.

En [19], el autor propone comparar estos residuales tomando en cuenta sus intervalos de confianza asociados, con lo que es posible producir una clasificación en categorías de grupos estadísticamente equivalentes. En este trabajo, se consideraron los intervalos de confianza del 95% para probar la equivalencia de cada par de residuales. En otras palabras, dos escuelas tienen desempeño similar si sus intervalos correspondientes se intersectan.

El criterio seguido para clasificar las escuelas fue separar el conjunto de escuelas cuyos intervalos se encontraban por arriba del promedio en dos clases: los grupos cuyos intervalos de confianza intersectan al mejor grupo de la categoría y los que no la intersectan. Del mismo modo, la categoría que se encuentra debajo del promedio se separó en dos clases: las escuelas cuyos intervalos intersectan al peor grupo de la categoría y los que no la intersectan. Finalmente, la categoría promedio, está conformada por todos aquellos intervalos

que intersectan el promedio. Esto produce una clasificación final de cinco categorías.

Desempeño Bruto	Frecuencia	Porcentaje	Porcentaje acumulado
Muy bajo	218	15.9	15.9
Bajo	191	13.9	29.8
Promedio	593	43.3	73.1
Alto	329	24.0	97.1
Muy Alto	38	2.8	99.9
Máximo	2	.1	100.0
Total	1371	100.0	

Cuadro 4.4. Clasificación de las escuelas por su DB.

Del cuadro 4.4, se observa que de la totalidad de escuelas, sólo un 26.9% muestra un DB que está por arriba del promedio.

2. Estimación del desempeño neto

Como se mencionó anteriormente, el desempeño bruto no refleja necesariamente la calidad del servicio o el nivel del logro de una escuela específica. Estos resultados pueden deberse a otros factores y no necesariamente sean dependientes de la escuela. Hay escuelas que ofrecen una mejor calidad de preparación académica a los alumnos en desventaja social y económica, en contraste con otras donde sus alumnos logran aparentemente mejores resultados en el examen, pero que buena parte de los mismos se deben a su origen sociocultural y no necesariamente a la calidad de la institución a la que asisten.

Con la finalidad de hacer las comparaciones más justas entre las escuelas, se definió un método para estimar el DN de una escuela. A partir del conjunto de variables disponibles en nuestros datos, se consideró el nivel de escolaridad de la madre y los hábitos de lectura del aspirante relacionados con el concepto de capital cultural. Aunado a lo anterior, el ingreso mensual familiar y el nivel de hacinamiento, se relacionaron con el concepto de capital económico, estas cuatro variables se combinaron en un índice sintético construido a partir de la aplicación de un análisis factorial a los datos (índice sociofamiliar).

Las diferencias de género se consideran también. Esta variable, en varias investigaciones, ha mostrado tener un efecto significativo en el desempeño. En general, los hombres tienden a desempeñarse mejor que las mujeres. El género se incorpora también como variable categórica explicatoria.

Los efectos fijos estimados se restan del puntaje total del aspirante, dependiendo de su caso particular. El uso de un modelo como éste permite hacer el análisis de las escuelas en igualdad de condiciones, suponiendo que todas tienen el mismo tipo de estudiantes.

El modelo de la ecuación (1) se modifica agregando el efecto debido al género (una variable *dummy* que toma valor 0 para mujeres y 1 para hombres) y el efecto debido al capital familiar global del aspirante, que es un índice

construido de las variables que representan la condición cultural y económica de los hogares de origen. Estos efectos se especifican como fijos y únicamente el logro académico promedio varía a través de las escuelas, lo que se refleja en las ecuaciones siguientes:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X1_{ij} + \beta_{2j} X2_{ij} + e_{ij} \quad (16)$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad (17)$$

$$\beta_{1j} = \gamma_{10} \quad (18)$$

$$\beta_{2j} = \gamma_{20} \quad (19)$$

La clasificación de las escuelas de acuerdo con su desempeño neto se llevará a cabo de la manera como se explicó anteriormente, es decir, comparando los residuales u_{0j} y considerando sus intervalos de confianza del 95% asociados. El cuadro 4.5 muestra cada una de las clasificaciones y el número de escuelas que caen en cada categoría.

Desempeño Neto	Frecuencia	Porcentaje	Porcentaje acumulado
Muy bajo	197	14.4	14.4
Bajo	220	16.0	30.4
Promedio	573	41.8	72.2
Alto	345	25.2	97.4
Muy Alto	34	2.5	99.9
Máximo	2	.1	100.0
Total	1371	100.0	

Cuadro 4.5. Clasificación de las escuelas por su DN

Del cuadro anterior se desprende, que si se realizara el análisis considerando el DN con base en una comparación más equitativa de las escuelas, los resultados son ligeramente diferentes, aquí el 27.8% tiene un DN por arriba del promedio.

Resulta interesante cruzar ambas clasificaciones para explorar el comportamiento de las escuelas y ver cuántas logran reclasificarse con base en su DN. El cuadro 4.6 muestra la clasificación de las escuelas analizadas por su DB y DN, de acuerdo con el resultado global del examen.

		Desempeño Neto						Total
		Muy Bajo	Bajo	Promedio	Alto	Muy Alto	Máximo	
Desempeño Bruto	Muy bajo	195	20	3	0	0	0	218
	Bajo	2	179	10	0	0	0	191
	Promedio	0	21	555	17	0	0	593
	Alto	0	0	5	323	1	0	329
	Muy Alto	0	0	0	5	33	0	38
	Máximo	0	0	0	0	0	2	2
Total		197	220	573	345	34	2	1371

Cuadro 4.6. Clasificación de las escuelas por su DB y DN.

Del cuadro anterior podemos observar, por ejemplo, que de las 218 escuelas que caen en la categoría Muy baja considerando el DB, 23 alcanzan a

reclasificarse en categorías más altas: 20 en la categoría Baja y solo tres alcanzan la clasificación Promedio. Otro ejemplo es la categoría Alta, donde cinco se reclasifican una categoría abajo; una en la categoría Muy Alta y el resto conserva su clasificación.

3. Descripción de las tipologías de escuelas

Con la finalidad de explorar si existe algún patrón de comportamiento sobresaliente de cada una de las tipologías de escuelas, se analizarán en términos de las variables que se trabajaron en la etapa de preparación de los datos. El siguiente cuadro muestra el número y el porcentaje de aspirantes que caen en cada una de las categorías generadas.

Desempeño Neto	Frecuencia	Porcentaje	Porcentaje acumulado
Muy bajo	11950	7.4	7.4
Bajo	29028	18.1	25.5
Promedio	58918	36.7	62.3
Alto	56149	35.0	97.3
Muy Alto	3784	2.4	99.6
Máximo	603	.4	100.0
Total	160432	100.0	

Cuadro 4.7. Número de estudiantes por categoría

Las siguientes gráficas presentan los patrones de comportamiento de cada una de las categorías de escuelas considerando las variables que discriminan más entre las clases.

Variables que representan los hábitos de estudio

Las variables relacionadas con los hábitos de estudio que mostraron mayor poder de discriminación en el desempeño en la prueba son básicamente si utilizan sus apuntes para estudiar, el uso de diccionarios, enciclopedias, etc., y las horas que dedican al estudio.

El análisis de estos indicadores, a la luz de las tipologías generadas, produce resultados interesantes que revelan el porcentaje de sustentantes que tienen estos hábitos, por tipo de escuela. En general, se observa que el hábito de estudiar principalmente con sus apuntes de clase es el que predomina más en las cinco categorías, mostrando variaciones en los niveles porcentuales de sustentantes, es decir, entre 66% y 71% para las clases con desempeño a lo más Promedio, y más de tres cuartas partes para las escuelas con desempeño por lo menos Alto. El siguiente hábito es el uso de materiales de estudio (diccionarios, enciclopedias, etc.), donde las clases Muy baja a Promedio presentan un porcentaje que va del 48% al 57% de sustentantes que manifiestan tenerlo, esto contrastado con los niveles porcentuales presentados por las categorías más altas, que van de un 65% a un máximo del 77%. Finalmente, el porcentaje de estudiantes que dedican más de 3 horas al estudio también varía de una clase a otra, mostrando una diferencia porcentual del 30% entre las clases extremas

(Muy baja, Muy Alta o Máxima), donde el rango porcentual de las categorías a lo más Promedio va del 39% al 48% y las más altas, entre 56% y 69% (Ver figura 4.1).

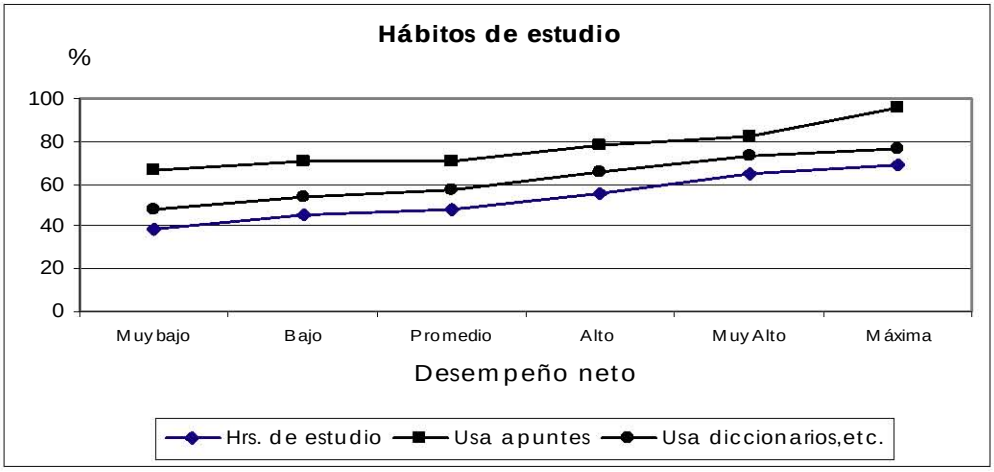

Figura 4.1 Clasificación de escuelas por su DN y hábitos de estudio

Variables que representan la trayectoria escolar en la secundaria de origen

La descripción de las tipologías en términos de la trayectoria de los sustentantes, relacionada principalmente con el porcentaje de estudiantes que reprueban y repiten año escolar, así como también con el promedio de la secundaria, presenta un patrón interesante. Por ejemplo, para el caso de la presentación de exámenes extraordinarios, este porcentaje varía relativamente poco en todas las clases, excepto para la clase con Máximo desempeño, que presenta apenas un 4.8%.

La repetición es otra variable de interés que se relaciona con su trayectoria en la secundaria y de la figura 4.2 se observa que son las clases con desempeño Muy bajo y Bajo, las que tienen el mayor porcentaje de estudiantes repetidores (mayor al 10%), y que éste va disminuyendo hasta ser casi nulo en las categorías altas. Finalmente, para el caso del promedio de secundaria, este parece mejorar a medida que mejora la clasificación y encontramos el promedio más alto en la categoría Máxima.

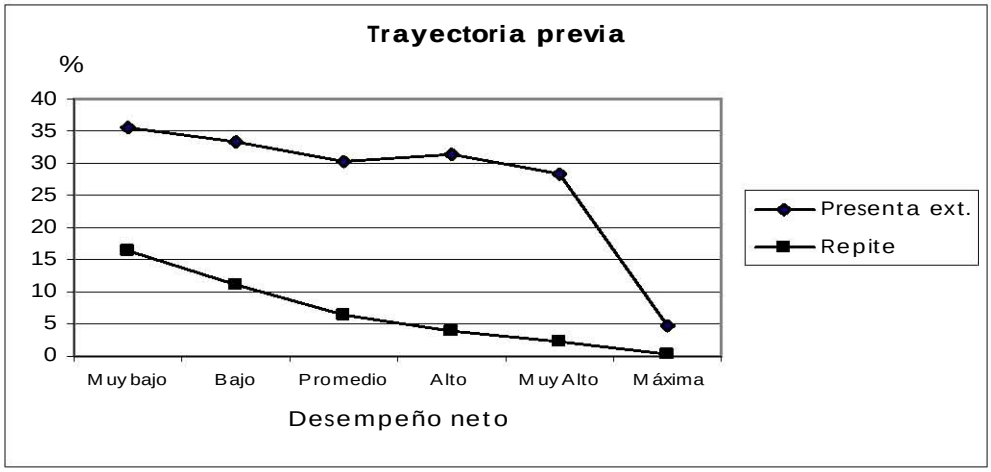

Figura 4.2 Clasificación de escuelas por su DN y Trayectoria previa

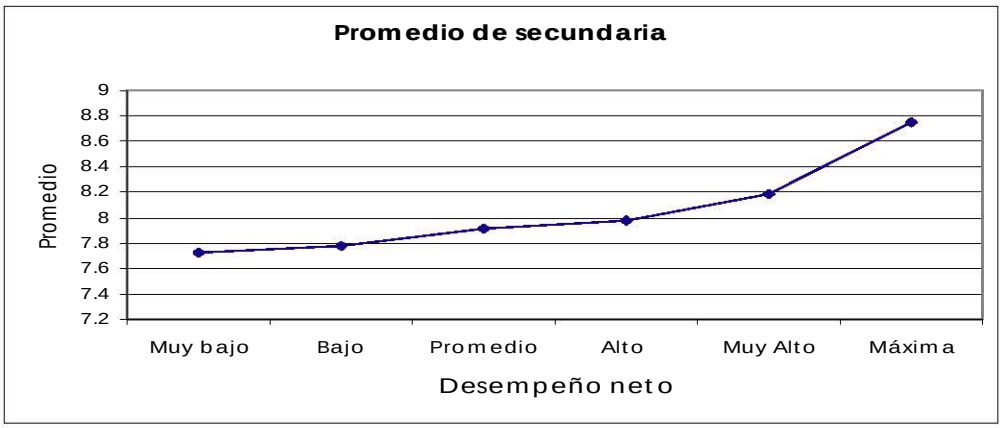

Figura 4.3 Clasificación de escuelas por su DN y Trayectoria previa

Involucramiento de los padres en la actividad escolar

En general, a partir de la gráfica siguiente puede observarse que el involucramiento de los padres en la actividad escolar que se relaciona con actitudes de confianza, respeto, promoción de cierta autonomía en las decisiones y el reconocimiento de logros escolares, son las que más discriminan en los resultados de la prueba y que marcan también las diferencias más fuertes entre las tipologías de escuelas generadas en nuestro análisis, concentrándose los porcentajes más altos en las clases con desempeño mayor al Promedio.

Es precisamente el respeto a las opiniones sobre lo que pasa en la escuela por parte de los padres lo que predomina más en todas las clases, con porcentajes que varían de un 74% para la clase Muy baja, hasta un 91% en la clase Máxima. Aunado a lo anterior, es la promoción de cierta autonomía permitiendo que tomen sus propias decisiones con respecto a la actividad escolar, la que también discrimina entre las clases, con niveles porcentuales del 58% para la clase Muy baja y entre 64%-67% para las clases Baja y Promedio. Las clases más altas muestran porcentajes del 75% o mayores. Finalmente, el reconocimiento de logros escolares se mantiene en niveles porcentuales menores al 60% en las clases a lo más Promedio y superiores al 65% para las clases más altas.

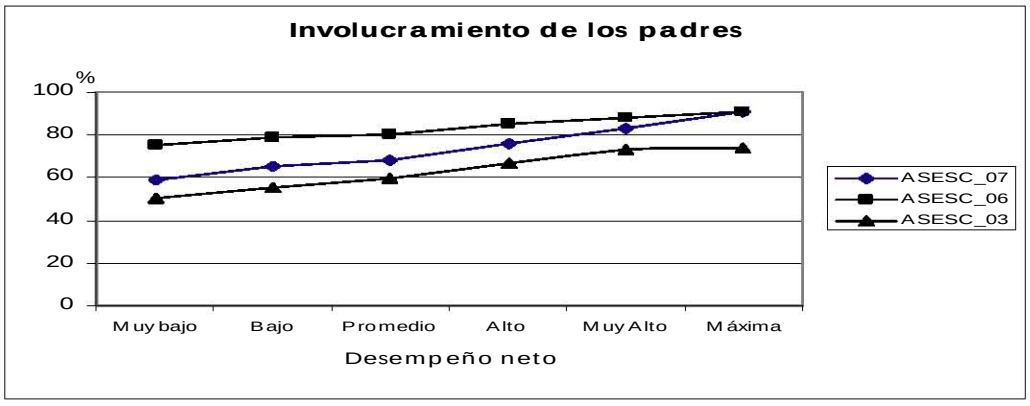

Figura 4.4 Clasificación de escuelas por su DN e involucramiento de los padres

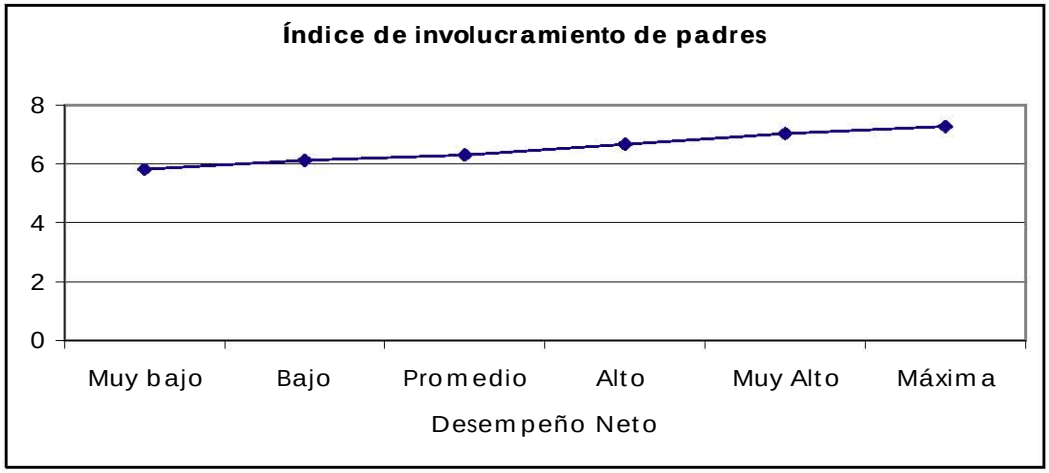

Figura 4.5 Clasificación de escuelas por su DN e índice de involucramiento de los padres

Descripción de los códigos utilizados en la gráfica:

ASESC_07= Promueven que tomen sus propias decisiones sobre lo que pasa en la escuela
ASESC_06= Respetan sus opiniones sobre lo que ocurre en la escuela
ASESC_03= Los felicitan o premian cuando les va bien en la escuela

Expectativas de estudio para el nivel superior

La figura 4.6 muestra cómo se comportan las expectativas de estudio para el nivel superior de los sustentantes en cada una de las clases. A partir de la gráfica se observa que todas son altas y que las categorías Muy baja, baja y promedio son muy semejantes y cercanas al 85%. Por otro lado, a partir del desempeño Alto, tiende a incrementarse el porcentaje, lo que significa que casi la totalidad de los aspirantes de este tipo de escuelas tienen la intención de continuar sus estudios de educación superior.

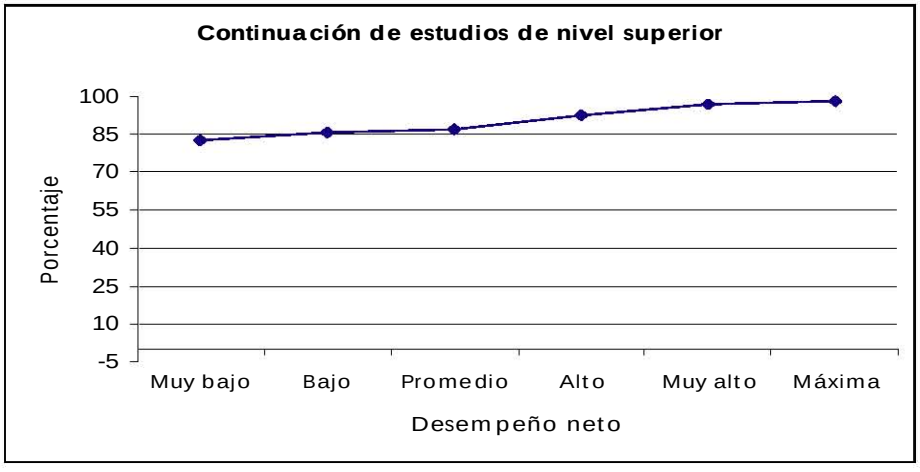

Figura 4.6 Clasificación de escuelas por su DN y expectativas

Condición laboral

La condición laboral de los sustentantes por tipo de escuela presenta variaciones importantes. Por ejemplo, la dos categorías más bajas muestran índices porcentuales que están por arriba del porcentaje global (11.7%) y la clase Promedio se encuentra muy cercana a este. Para el caso de las clases más altas, este porcentaje se reduce gradualmente hasta llegar a un 2.7% para la clase Máxima.

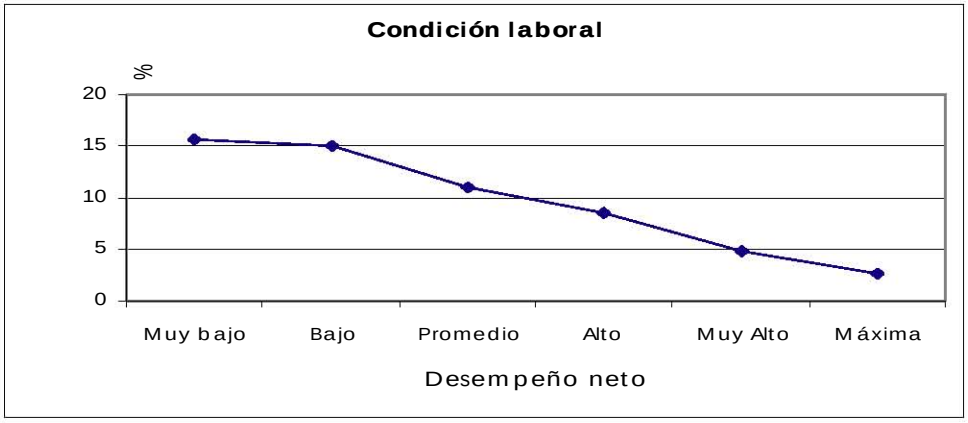

Figura 4.7 Clasificación de escuelas por su DN y condición laboral

Estructura familiar

Con respecto a la composición familiar, se observa que en las escuelas con desempeño por lo menos Alto, más de una tercera parte de los sustentantes (34%) señala ser hijo único o tener solo un hermano. Para el caso de las categorías Promedio, Baja y Muy baja, los porcentajes son: 25.9%, 22.8 y 19.4%, respectivamente. El porcentaje de estudiantes primogénitos también varía entre las tipologías, siendo del orden del 31% en la categoría Muy baja, incrementándose gradualmente, hasta llegar a porcentajes del 42% y 52%, respectivamente para las dos clases más altas.

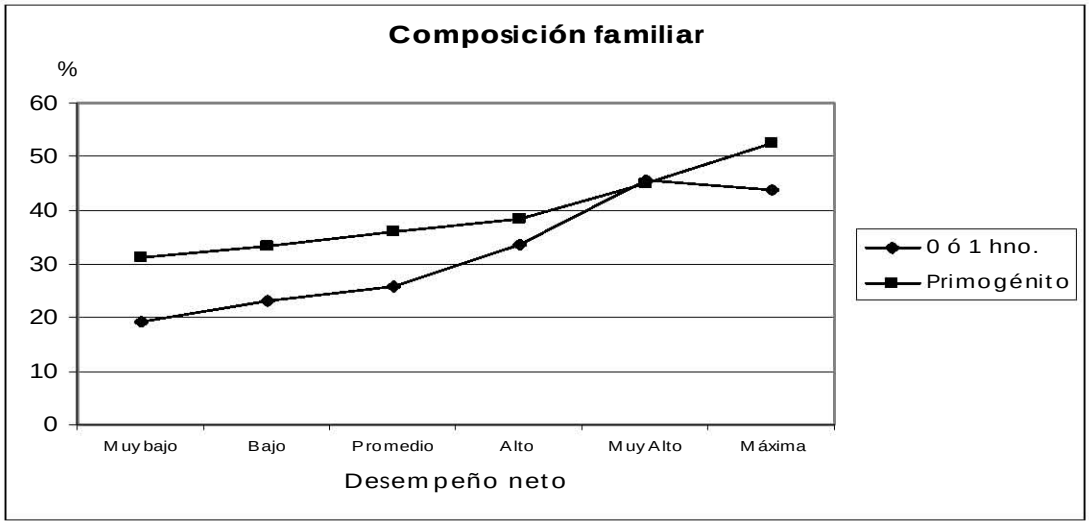

Figura 4.8 Clasificación de escuelas por su DN y composición familiar

Ocupación de los padres

En cuanto a la ocupación de las madres, se observa que el porcentaje de aspirantes que señalan que sus madres tienen un status ocupacional alto, es mayor entre las clases más altas (Muy Alta y Máxima), superior al 20%, esto contrastado con el 5% o menor para las dos clases más bajas. El porcentaje de aspirantes con madres de status ocupacional bajo va disminuyendo conforme mejora la clasificación. El status ocupacional medio varía relativamente poco entre las clases.

Para el caso de los padres, se observa un comportamiento similar pero más claro. Es decir, el porcentaje de status ocupacional alto sigue una tendencia positiva, va aumentando conforme mejora la clasificación, pero de una forma más rápida, partiendo de un 13.7% para la clase Muy baja, hasta llegar a un 44.7% para la clase Muy alta. De manera semejante para el status ocupacional bajo, se disminuye el porcentaje, del 55% para el desempeño Muy bajo hasta el 36.7% para el Muy alto.

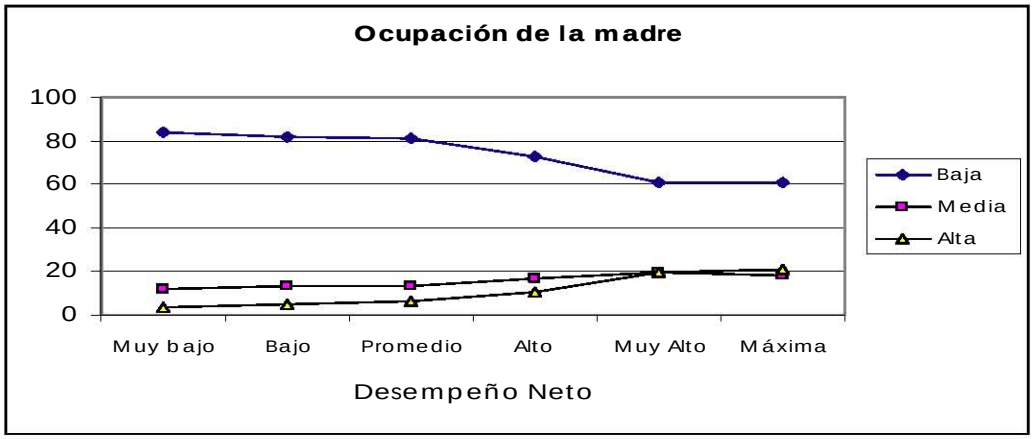

Figura 4.9 Clasificación de escuelas por su DN y ocupación de la madre.

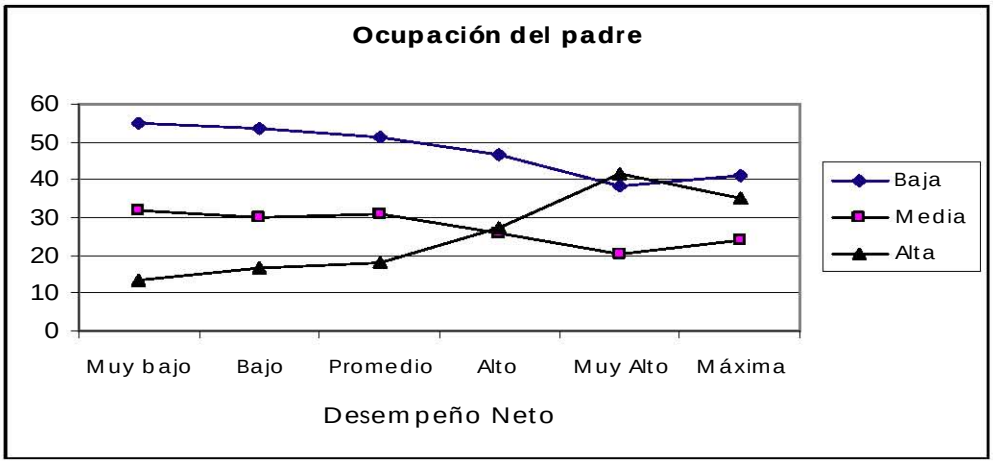

Figura 4.10 Clasificación de escuelas por su DN y ocupación del padre

Índice de alimentación

Con respecto al índice de alimentación, resulta claro que conforme mejora la clasificación el índice es más alto, lo que indica que son precisamente los

estudiantes de las escuelas con desempeño por lo menos Alto, los que tienen acceso y consumen con más frecuencia todo tipo de alimentos.

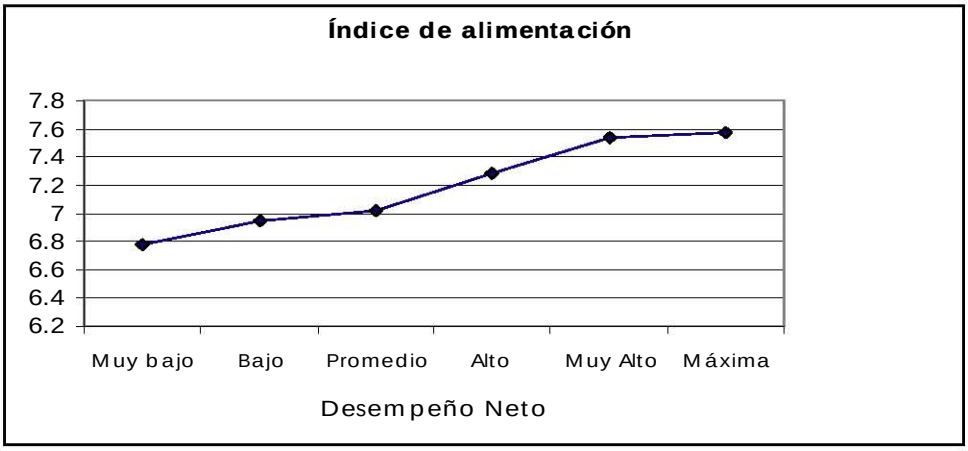

Figura 4.11 Clasificación de escuelas por su DN e índice de alimentación

Índice del comportamiento de los profesores

El índice que refleja las formas de actuar y el comportamiento de los profesores en el aula no parece discriminar en los desempeños netos. Esto ya se había evidenciado en secciones anteriores y se hace también evidente en esta descripción conjunta de clases y desempeño neto.

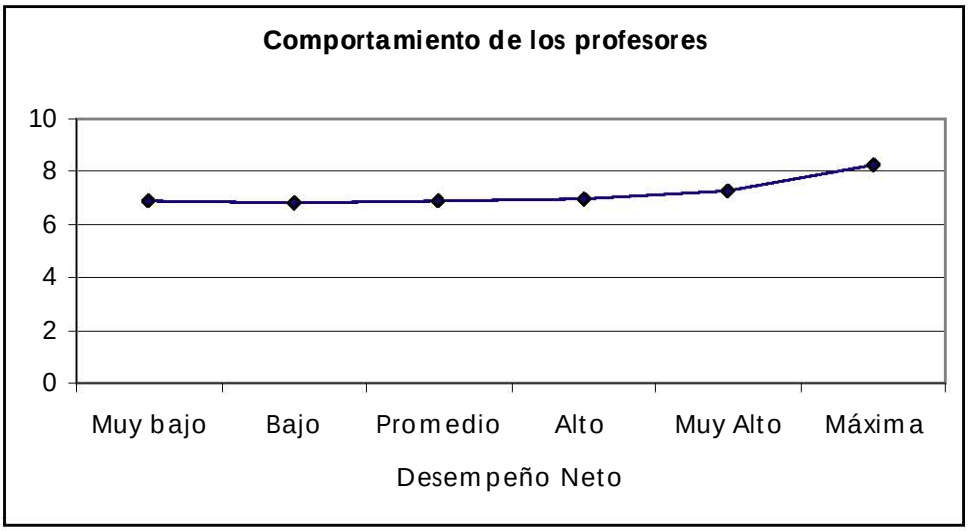

Figura 4.12 Clasificación de escuelas por su DN y comportamiento de los profesores

4. Análisis del modelo con la variable contextual: "índice sociofamiliar promedio"

En el estudio de la eficacia escolar generalmente se consideran mediciones realizadas al nivel de la escuela, por ejemplo, el tamaño, el tipo de escuela (rural o urbana), clima escolar, etc. El objetivo es determinar el efecto que tienen estas variables en la explicación de la variación en la relación entre la variable

respuesta y la (s) predictora (s); así como también la variación del intercepto a través de las escuelas.

Dentro del conjunto de datos disponible, no se encontraron mediciones realizadas directamente al nivel de la escuela, por lo que se consideró únicamente la construcción de una variable composicional que se denominó: "índice sociofamiliar promedio", generada a partir de agregar, en términos del promedio, el índice sociofamiliar individual.

La finalidad de contextualizar en términos del índice sociofamiliar es la obtención de un estimador del efecto que tiene la composición sociocultural promedio del alumnado de una escuela, en el logro promedio de la misma. Además, interesa también observar si hay un efecto moderador de esta variable composicional en las diferencias que existen entre el logro y las variables correspondientes al género y el origen social. Es decir, se podría suponer que si una escuela tiene una composición sociocultural alta de su alumnado, esto podría contribuir a que las diferencias en el logro entre hombres y mujeres se reduzca.

La especificación de un modelo multinivel, considerando el género y el índice sociofamiliar como variables al nivel individual y permitiendo la variación de sus efectos a través de las escuelas (coeficientes aleatorios); además de la inclusión de la variable "índice sociofamiliar promedio" derivada de la contextualización, produce los siguientes resultados:

Efecto fijo	Coeficiente	Error estándar
Para el intercepto, β_0:		
Intercepto2, γ_{00}	4.459	0.0099
Mediaindsoc, γ_{01}	0.764	0.0237
Para pendiente Sexo, β_1:		
Intercepto2, γ_{10}	0.409	0.007
Mediaindsoc, γ_{11}	0.044	0.015
Para pendiente Indsoc, β_2:		
Intercepto2, γ_{20}	0.256	0.004
Mediaindsoc, γ_{21}	-0.034	0.009

Cuadro 4.8. Estimación de efectos fijos

Efecto aleatorio	Desviación estándar	Coeficiente
U_o	0.319	0.102
Pendiente Sexo, U_1	0.069	0.005
Pendiente Indsoc, U_2	0.071	0.005
R	1.22	1.502

Cuadro 4.9. Estimación de los componentes de varianza

Todos los efectos resultaron significativos, pero únicamente el efecto de la variable composicional para explicar la variación del logro académico promedio de las escuelas es alto y positivo, lo que significa que a medida que la composición sociocultural de la escuela es más alta, el logro promedio de la escuela aumenta. Los efectos moderadores para explicar la variación en los coeficientes de regresión, correspondientes al género y al índice sociofamiliar individual, son pequeños, lo que significa por ejemplo, que esta variable no contribuye mucho a moderar la relación entre el género y el logro, es decir, el hecho de tener una

composición sociocultural alta del alumnado, no contribuye a reducir significativamente las diferencias en el desempeño entre hombres y mujeres. La misma situación se presenta para el caso del índice sociofamiliar.

Es importante mencionar que con la inclusión en el modelo de la variable "índice sociofamiliar promedio", al nivel de las escuelas, se explica un 57% de la variación de los interceptos (logro académico promedio) a través de las escuelas, lo que refleja el impacto que tienen las desigualdades sociales en los aprendizajes.

Análisis multinivel bajo el enfoque de la distribución social del conocimiento

De acuerdo a lo mencionado en secciones anteriores, es razonable suponer que las escuelas eficaces no lo sean de la misma forma para cualquier tipo de alumno. Por ejemplo, es probable que los varones aprovechen más que las mujeres los efectos de una escuela eficaz; o que los de alto capital sociofamiliar más que los de bajo capital, etc., lo ideal sería que la escuela lograra compensar estas diferencias iniciales en su alumnado y tener un efecto semejante en todos ellos, a este enfoque se le denomina "distribución social del conocimiento".

Para explorar este enfoque de la distribución social del conocimiento con los datos disponibles, se consideró el modelo utilizado para producir las tipologías de escuelas, modificando únicamente el *status* de la pendiente correspondiente al índice sociofamiliar, que pasó de ser fija a aleatoria, es decir, ahora su efecto varió de escuela a escuela. Los estimadores de estos coeficientes se obtuvieron por el método de Mínimos Cuadrados Ordinarios (OLS), generados también por el paquete HLM.

De acuerdo a la revisión realizada de las investigaciones sobre la Eficacia Escolar, es generalizada la recomendación de contar con indicadores del logro académico inicial y del logro al final del nivel educativo que se pretenda analizar, esto para estimar el "valor agregado" correspondiente al efecto de la escuela. Conscientes de esta situación y únicamente con la finalidad de explorar este fenómeno con la información que se encuentra disponible, se decidió utilizar como criterio para considerar si una escuela es eficaz que su desempeño sea por lo menos igual al logro académico promedio de todas las escuelas (γ_{00}) y su efecto similar para todo el alumnado, es decir, que sus desempeños sean aproximadamente iguales y por lo menos también iguales a lo que se espera de ellos considerando su origen sociofamiliar. Para este fin, se analizaron aquellas escuelas que tienen un logro académico promedio mayor o igual al estimador promedio del intercepto para todas las escuelas y un efecto casi nulo del índice sociofamiliar. Aunado a lo anterior, se exploró el desempeño del alumnado dentro de cada una de estas escuelas para verificar si este logro se encontraba por lo menos en el desempeño esperado de acuerdo a su origen sociofamiliar.

Con la finalidad de facilitar este análisis, se construyeron cinco grupos de estudiantes de acuerdo al índice sociofamiliar utilizando el Análisis Cluster. Posteriormente, se calculó el logro promedio para cada uno de estos tipos de estudiantes dentro de las escuelas, para contrastarlas finalmente con el desempeño global de estas mismas categorías y observar si "todos ganan" en este enfoque de calidad y equidad de las escuelas, es decir, si todos están por lo menos en el logro promedio que se espera de acuerdo a su origen sociofamiliar.

El siguiente cuadro presenta el desempeño global promedio para cada tipo de estudiante de acuerdo a su origen sociofamiliar. Este desempeño global promedio se utilizará como referente para contrastar el desempeño de esta tipología de estudiantes dentro de las escuelas y verificar si efectivamente "todos ganan" (desempeño por lo menos igual a su correspondiente promedio global) en las escuelas analizadas.

Origen sociofamiliar	Indicador	Desempeño Global (escala 0-10)
Muy bajo	Media	4.30
	N	35289
Bajo	Media	4.63
	N	56035
Promedio	Media	4.96
	N	39395
Alto	Media	5.24
	N	21451
Muy Alto	Media	5.57
	N	8262
Total	Media	4.77
	N	160432

Cuadro 4.10. Desempeño global por origen sociofamiliar

Del conjunto global de escuelas analizadas, sólo 70 (5.2%) cumplieron con el criterio de tener un intercepto mayor o igual al promedio de todas las escuelas y una pendiente para el índice sociofamiliar casi nula. De este subconjunto, sólo en 15 escuelas sus estudiantes presentaron un desempeño igual o superior al desempeño esperado, es decir, sólo esta mínima parte de las escuelas sería considerada como eficaz y equitativa desde nuestra perspectiva de escuela eficaz. Sin duda, si se contara con variables medidas al nivel de la escuela, resultaría muy interesante explorar qué características tienen estas escuelas que logran los dos atributos de calidad y equidad en los aprendizajes.

Un dato adicional que se deriva del análisis de las 55 escuelas restantes, es el porcentaje de las mismas donde sus estudiantes, agrupados por categorías, no obtienen ni siquiera el logro esperado. Del siguiente cuadro, se observa que son precisamente los estudiantes de origen social, por lo menos promedio, los que presentan desempeños por debajo de lo que se espera de ellos, en más del 40% de este subconjunto de escuelas.

Origen social	Porcentaje
Muy bajo	4.3
Bajo	12.9
Promedio	42.9
Alto	51.4
Muy alto	52.9

Cuadro 4.11. Porcentaje de escuelas donde sus estudiantes no logran superar el promedio esperado

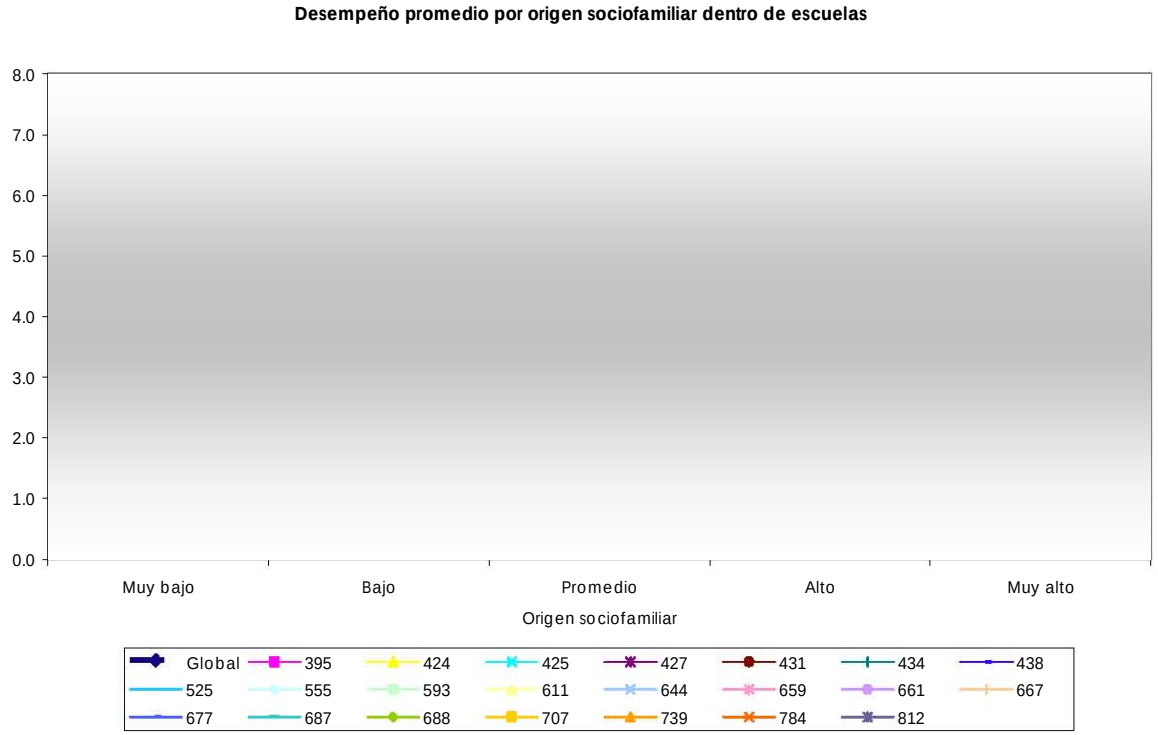

Figura 4.13. Un subconjunto de las 70 escuelas analizadas

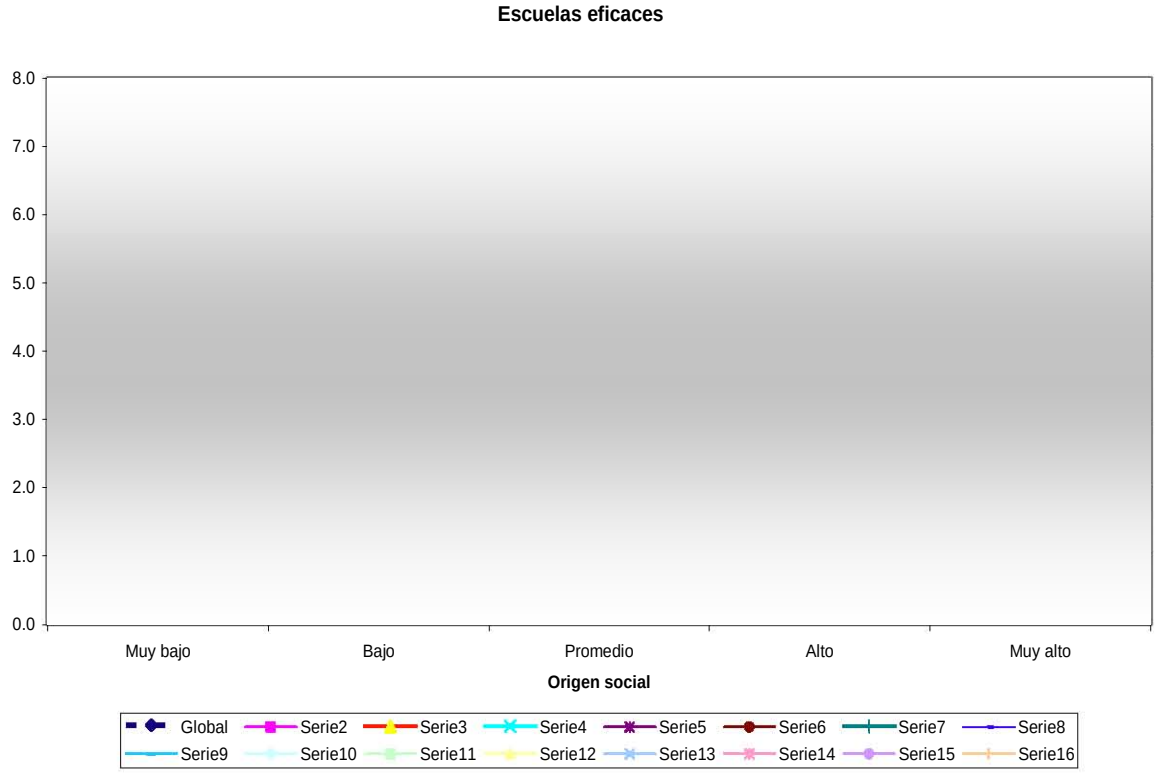

Figura 4.14. Subconjunto de escuelas eficaces

Origen social	Características
Muy bajo	Escolaridad de la madre a lo más de profesional técnico, con un porcentaje superior al 90% concentrado en las categorías 1 a 3 de escolaridad. El 97.6% de los ingresos familiares mensuales concentrados en las categorías de 1 a 4. El 10% tiene 4 ó menos personas viviendo en su casa. Solo Un 10.6% tiene hábitos de lectura.
Bajo	El 98.3% de las madres tiene a lo más una escolaridad de capacitación técnica (después de la secundaria) o menor. El 98.2% de los ingresos familiares concentrados en las categorías de 1 a 6. El 33.7% tiene 4 ó menos personas viviendo en su casa. Sólo un 25.6% tiene hábitos de lectura.
Promedio	El 98% de las madres tiene a lo más una escolaridad de capacitación técnica (después de la secundaria) o menor. El 97.8% de los ingresos familiares concentrados en las categorías de 1 a 9. El 44.5% tiene 4 ó menos personas viviendo en su casa. Sólo un 41.1% tiene hábitos de lectura.
Alto	El 94.7% de las madres tiene por lo menos una escolaridad de secundaria. El 90% de los ingresos familiares concentrados en las categorías de 4 a 15. El 49.9% tiene 4 ó menos personas viviendo en su casa. El 50.3% tiene hábitos de lectura
Muy alto	El 96.5% de las madres tienen escolaridad de capacitación técnica (posterior a la secundaria) o mayor. El 97.4% tiene ingresos familiares de 6 a 15. El 56.3% tiene 4 ó menos personas viviendo en su casa. El 64.6% tiene hábitos de lectura.

Cuadro 4.11 Descripción de las tipologías de estudiantes de acuerdo a las variables utilizadas para construir el índice sociofamiliar.

4.2 Evaluación de los resultados

Revisión del proceso

La aplicación de la metodología multinivel en este estudio ha permitido revelar nuevo conocimiento en términos de qué tan homogéneas son las escuelas en cuanto a la composición sociocultural de sus egresados y del porcentaje de variación en el desempeño de los estudiantes que es atribuible a las escuelas. Sin duda, el segundo resultado es congruente con aquellos obtenidos en estudios internacionales sobre esta línea de investigación y el primero tal vez resulte una nueva aportación sobre la diversidad de alumnos, en términos socioculturales, que atienden las escuelas del área metropolitana.

Considerando los objetivos de este estudio, sólo se han incluido como factores de ajuste las características sociofamiliares y de género, por lo comentado en secciones anteriores, pero es factible la inclusión de un mayor número de características del estudiante y de la escuela, con la sugerencia que hacen comúnmente los estudiosos del tema de apoyarse en algún modelo

conceptual que permita tomar decisiones en el curso del análisis, así como también para efectos de la interpretación de los resultados.

Aunado a lo anterior y como las tendencias internacionales lo muestran, el estudio de la eficacia escolar, dada su relevancia, es una línea de investigación que cada día tiene más adeptos. El hecho de que algunas escuelas sean más exitosas que otras, genera cuestionamientos sobre qué es la eficacia, cuáles factores contribuyen a lograrla y cómo se puede usar esta información como base para mejorar las escuelas y por ende los resultados de los estudiantes. Para realizar un estudio de este tipo, adicional a los datos que se tienen, se requiere contar con un indicador de logro inicial, así como también características medidas al nivel de las escuelas, lo que podría servir como referencia a las instituciones responsables de las evaluaciones externas, para considerar su inclusión en los bancos de información que mantienen y poder ampliar el alcance de futuras investigaciones basadas en estos datos.

Finalmente, como parte de las ventajas de la institucionalización de las evaluaciones estandarizadas externas en México y de la constante acumulación de estos resultados, concretamente en los términos de este estudio, sería interesante monitorear continuamente el desempeño de las escuelas del área metropolitana y en general, de todas las entidades de la Republica Mexicana, para explorar la estabilidad de los resultados que obtienen a lo largo del tiempo.

4.3 Despliegue de resultados

Reporte final

La realización del proyecto de descubrimiento de conocimiento en la base de datos de aspirantes a EMS se llevó a cabo considerando el conocimiento previo ganado a raíz del gran número de investigaciones internacionales que se han realizado sobre este tema y de la evidencia empírica presente en los datos disponibles.

Entre los resultados obtenidos más importantes se pueden enunciar los siguientes:

1. En general, de la información disponible, son las características sociofamiliares las que tienen un efecto mayor en el logro en la prueba. Estas características sólo abarcan una pequeña proporción de la variación que de acuerdo a investigaciones internacionales realizadas se atribuyen al individuo (aproximadamente 40%).
2. La descomposición de la varianza al nivel individual y de escuela revela que el porcentaje de variación, en términos del desempeño en la prueba, atribuible a las escuelas está dentro del rango de valores obtenidos en otros estudios internacionales de este tipo. El valor de 12.5% obtenido en el presente estudio indica que las escuelas son relativamente semejantes en cuanto al desempeño de sus estudiantes en la prueba. Del 87.5% restante, como se mencionó anteriormente, aproximadamente un 50% podría explicarse con variables medidas al nivel del salón de clases y un 40% al nivel individual, esto si se contara con mediciones importantes para justificar tres niveles de análisis (alumnos-salón de clases-escuela).

3. Para el caso de las escuelas del área metropolitana, la composición sociofamiliar del alumnado no es lo suficientemente homogénea como se esperaría en sociedades donde hay patrones de asentamiento residencial marcados y una fuerte diferenciación social. Esta composición sociocultural diversa genera efectos significativos a favor de los más aventajados socialmente, es decir, a mayor status social, mayor logro académico.

4. Con respecto a las tipologías de escuelas obtenidas, aproximadamente el 62% de los sustentantes provienen de escuelas con un Desempeño Neto a lo más Promedio y más de una tercera parte de escuelas (38%) con un Desempeño Neto por lo menos Alto.

5. Las tipologías de escuelas con DN más altos (por lo menos Alto), presentan el porcentaje mayor de estudiantes que tienen padres cuyas ocupaciones son medias o altas, un porcentaje muy reducido trabaja, tienen hábitos de estudio que se relacionan positivamente con el logro, trayectorias escolares en sus secundarias con muy bajos índices de repetición de año escolar, promedios de secundaria más altos, casi la totalidad tiene expectativas de continuar sus estudios de educación superior, su composición familiar en cuanto al número de hermanos y el lugar que ocupan entre ellos es favorable en el sentido de que provienen de familias poco numerosas, donde a lo más hay dos hijos o son los primogénitos.

6. Resalta el caso de la reprobación y la presentación de exámenes extraordinarios, donde el porcentaje de sustentantes que han reprobado no varía mucho en las primeras cinco tipologías de escuelas, reduciéndose significativamente en la clase con Desempeño Neto Máximo.

7. Del análisis de los datos bajo el enfoque de la distribución social del conocimiento, se derivan resultados interesantes. La aplicación del criterio para identificar las escuelas eficaces genera un conjunto muy reducido de 15 escuelas, donde, desde la perspectiva de escuela eficaz considerada, se puede hablar de calidad y equidad. El conjunto restante de 55 escuelas, aunque presenta efectos del índice sociofamiliar nulos y "pisos de aprendizaje promedio" de escuela mayores o iguales al promedio de todas las escuelas, tiene la desventaja de que los estudiantes con status sociofamiliar promedio, alto y muy alto están por debajo del logro promedio esperado.

8. El análisis del modelo incluyendo la variable contextual: "índice sociofamiliar promedio" explica aproximadamente un 57% de la variación en el logro académico promedio a través de las escuelas, lo que sin duda refuerza la tendencia mostrada en los resultados ya expuestos. Es decir, una mejor composición social del alumnado genera mayores "promedios de aprendizaje".

Conclusiones

El proceso de descubrimiento de conocimiento en bases de datos está emergiendo como una herramienta poderosa para el soporte en la toma de decisiones en el mundo de los negocios y su utilidad se expande continuamente para abarcar nuevos campos del conocimiento. La gran difusión de las historias exitosas en la aplicación de DCBD se ha encargado de atraer la atención de muchos investigadores e industrias y conforme transcurre el tiempo, el campo se va enriqueciendo en términos del número de modelos y algoritmos, lo que se refleja en la vasta cantidad de herramientas de minería de datos académicas y comerciales disponibles.

A pesar de lo anterior, el proceso de descubrimiento está todavía lejos de reconocerse como una práctica habitual, en lugar de ello, se considera como un campo muy especializado. La razón principal de esta visión limitada es la falta de una práctica bien entendida y versátil del proceso, esto se debe a que los proyectos de este tipo se realizan en forma *ad-hoc* y aislada, con posibilidades casi nulas de aprender y reutilizar la experiencia que se genera en su realización.

Los expertos en el tema coinciden que el campo se enriquecería significativamente con la creación de un modelo de procesos estándar que guíe el ciclo de vida completo del proceso de descubrimiento de conocimiento. Una metodología que sea ampliamente aceptada y que lo ubique como una práctica convencional. Desafortunadamente, la mayoría de las herramientas disponibles tienen su propio modelo de procesos y en la mayoría de los casos, están estrechamente ligados a sus respectivas plataformas y herramientas. Otro factor en contra es que estos esfuerzos encaminados hacia la estandarización del proceso de descubrimiento se encuentran dispersos en diferentes industrias y universidades.

Una de las propuestas de modelo de procesos es CRISP-DM (Cross-Industry Standard Process for DM), que nace de un consorcio formado por compañías destacadas (Daimler-Benz, SPSS y NCR), donde intentan plasmar el gran caudal de experiencia adquirida en este tipo de prácticas, lo que refleja que no se trata de una propuesta teórica o académica, alejada completamente de los escenarios reales que se enfrentan cuando se intenta resolver un problema de descubrimiento de conocimiento. Las expectativas relacionadas con este modelo están encaminadas a proporcionar una guía eficiente en el desarrollo del ciclo de vida del proceso de descubrimiento de conocimiento, para los analistas de datos tanto novatos como para aquellos más experimentados. Se estima que una buena parte de los profesionales dedicados a proyectos de descubrimiento de conocimiento utilizan la metodología CRISP-DM.

La metodología CRISP-DM es un modelo de procesos jerárquico. En el nivel principal, el proceso se divide en seis fases genéricas, que van desde la comprensión del problema hasta la presentación de los resultados de un proyecto. El siguiente nivel desglosa cada una de estas fases en varias tareas genéricas. En este nivel, la descripción es lo suficientemente genérica para cubrir todos los escenarios de DCBD. El tercer nivel especializa estas tareas para situaciones específicas. Por ejemplo, la tarea genérica podría ser la depuración de los datos y la tarea especializada podría ser la depuración de valores numéricos o categóricos. El cuarto nivel es la instancia del proceso, es decir, el registro de

acciones, decisiones y el resultado de una ejecución real. El modelo también describe relaciones entre diferentes tareas de DCBD. Proporciona una secuencia idealizada de acciones realizadas durante el proyecto, pero no intenta proporcionar rutas posibles a través de las tareas.

Existen varios grupos que están trabajando modelos de procesos para estandarizar el descubrimiento de conocimiento en bases de datos. Mientras que unos trabajan arduamente en desarrollar lenguajes de MD estándares (OLEDBDM, PMML); otros trabajan en proporcionar marcos de trabajo para los sistemas que soportan decisiones (TASF) o los específicos para herramientas (SEMMA), CRISP-DM provee de herramientas y modelos de procesos de DCBD para la industria neutral.

En la realización de este proyecto de descubrimiento de conocimiento se decidió utilizar la metodología CRISP-DM debido a que ofrece un enfoque de procesos claro, detallado y muy apegado a la propuesta original del proceso de descubrimiento de conocimiento, que en lo personal considero es una percepción completa de esta actividad. Aunado a lo anterior, la capacidad de generar, a partir de esta metodología general, modelos de procesos específicos que sean aplicables a proyectos de descubrimiento semejantes, representa una gran ventaja para las compañías y profesionales independientes dedicados a este tipo de consultorías, debido al potencial de estandarización de tareas, que permitiría la medición de tiempos y recursos indispensables, así como también de la predicción de costos para cubrir metas de descubrimiento específicas.

Finalmente, desde mi punto de vista, este trabajo representa una aportación valiosa desde la perspectiva del conocimiento extraído de la base de datos, que sin duda resultó de utilidad para el experto en el dominio en la comprensión de esta línea de investigación, además de sugerir ideas o posibles hipótesis para futuras investigaciones, en torno a la situación de algunas escuelas del área metropolitana, principalmente en términos del tipo de alumnos que atienden. A partir de los resultados generados es claro que, de las variables disponibles, es el origen sociofamiliar el que tiene mayor impacto en el logro académico y es un factor que debe ser considerado para la realización de comparaciones en igualdad de condiciones entre las escuelas.

Aunado a lo anterior, es importante destacar las ventajas que se obtuvieron al adoptar la metodología CRISP-DM para planear el proyecto y llevarlo a cabo con éxito, además de documentar este modelo específico que se espera sea de utilidad en proyectos futuros donde se tengan metas de descubrimiento semejantes.

Referencias bibliográficas

[1] Aitkin, M., and Longford, N. Statistical modelling in school effectiveness studies (with discussion). J. Roy. Statist. Soc., A, 149, 1-43. 1986.

[2] Cresswell, J. Schooling Issues DIGEST. School Effectiveness. Department of Education Science and Training. 2004/1.

[3] Coe, Robert y Fitz-Gibbon, Carol Taylor. School effectiveness research: criticisms and recommendations. Oxford Review of Education, 24(4), pp. 421-438. 1998.

[4] Coleman, J., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfield, F. & York, R. Equality of educational opportunity. Washington, DC, US Government Printing Office. 1996.

[5] Edelstein, Herbert. Introduction to Data Mining and Knowledge Discovery, Third Edition. Two Cows Corporation. USA. 1999.

[6] Edmonds, R. (1979). Effective Schools for the Urban Poor. Educational Leadership, 37, pp. 15-27.

[7] Fayyad, U., G. Piatetsky-Shapiro, G., Amith, Smyth, P., and Uthurusamy, R. Advances in Knowledge Discovery and Data Mining. Revista AAAI/MIT Press, pags. 1-36, Cambridge, 1996.

[8] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of the ACM, V. 39, no. 11, págs. 27-34, November 1996.

[9] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. From Data Mining to Knowledge Discovery: An Overview. In Advances in Knowledge Discovery and Data Mining, eds. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Menlo Park, CA: AAAI Press, pp. 1-34. 1996.

[10] Fernández, T. Determinantes sociales y organizacionales del aprendizaje en la Educación Primaria de México. Un análisis de tres niveles. Informe de Investigación para el Instituto Nacional para la Evaluación de la Educación (INEE) de México. México , D.F. (2003b).

[11] Fernández, T. Perfiles de las escuelas primarias eficaces de México. Informe final de Investigación para el Instituto Nacional para la Evaluación de la Educación (INEE) de México. (2003).

[12] Fernández, T. Métodos estadísticos de estimación de los efectos de la escuela y su aplicación al estudio de las escuelas eficaces. Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación. Vol 1, núm.2. 2003.

[13] Fitz-Gibbon, C. T. Multilevel Modelling in an Indicator System. Paper presented to the ESRC International Conference on Multilevel Methods in Educational Research, University of Edinburgh. 1989.

[14] Friedman, J. H. Data mining and statistics: what's the connection?.1997a.

[15] Friedman, J.H. On bias, variance, 0/1-loss, and the curse of dimensionality. Data Mining and Knowledge Discovery 1(1):55-77. 1997b.

[16] Friedman, J.H. and J. W.Tukey. A projection pursuit algorithm for exploratory data analysis. IEEE Trans. Comput. C-23: 881-889. 1974.

[17] Gelman, A. Multilevel (hierarchical) modeling: what it can and can't do. June, 2005.

[18] Goldstein, H. Multilevel mixed linear model analysis using iterative generalized least squares, Biometrika, 73, 1, 43-56. 1986.

[19] Goldstein, H. Methods in school effectiveness research. School effectiveness and school improvement. 8: 369-95. 1997.

[20] Gray, J. The quality of schooling: Frameworks for judgement, British Journal of Educational Studies, 38, 3, 204-223. 1990.

[21] Gray, J. Information visualization in DM and KD. Microsoft Research

[22] Hand, D. Statistics and Data Mining: Intersecting Disciplines. SIGKDD Explorations, 1999.

[23] Hand, D., Data Mining: Statistics and More?. The American Statistician, Vol. 52, No. 2. May 1998.

[24] Hand, D. Data mining –reaching beyond statistics. Research in Official Stat. 1(2): 5-17. Año 1998.

[25] Hox, J.J. Multilevel analysis: techniques and applications. Lawrence Erlbaum Associates, Publishers. Año 2002.

[26] H. Jiawei, K. Micheline, Data Mining: Concepts and Techniques. Editorial: Academic Press. Año 2001.

[27] James, S. The role of Bayesian and Frequentist Multivariate Modeling in Statistical Data Mining. CRC Press LLC. 2004.

[28] J.F. Hair, Jr.,R.E., Anderson, R. L., Tatham, W.C. Black, Análisis Multivariante, 5a. Edición. Editorial Prentice Hall Iberia, Madrid, 1999.

[29] Jesson, D. & Gray, J. Slants on slopes: Using multi-level models to investigate differential school effectiveness and its impact on pupils' examination results, School Effectiveness and School Improvement, 2, 3, 230-247. Año 1991.

[30] Jesson, D. Beyond the league tables, Education, 179, 9, 179-180. Año 1992.

[31] Kreft, J.G.G., DeLeeuw, J., and Kim,K.S. Comparing four different statistical packages for hierarchical linear regression: GENMOD, HLM,ML2,VARCL. UCLA Centre for research on evaluation, Los Angeles, California, USA. Año 1990.

[32] Klösgen, W., Zytkow, J.M., Knowledge Discovery in Databases: The purpose, necessity, and challenges. Handbook of Data Mining and Knowledge Discovery. Oxford University Press, 2002.

[33] M. Goebel, L. Gruenwald. A survey of data mining and knowledge discovery software tools. SIGKDD Explorations, june 1999.

[34] Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS), Rûdiger Wirth (DaimlerChrysler). CRISP-DM 1.0: Step-by-step data mining guide. Copyright © 1999, 2000.

[35] Piatetsky-Shapiro, G. Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop. AI Magazine 11(5):68-79, 1991.

[36] Piatetsky-Shapiro, G. Data Mining Coming of Age. Handbook of Data Mining and Knowledge Discovery. Oxford University Press. Año 2002.

[37] Quinlan, R. Learning logical definitions from relations. Machine learning 5(3): 239-266. 1990.

[38] Raudenbush, S. y Willms, D. The Estimation of School Effects. Journal of Educational and Behavioral Statistics, 20(4), pp 307-335. Año 1995.

[39] Raundenbush, S., Bryk, A. Hierarchical Linear Models. Applications and Data Analysis Methods. Second Edition. Año 2002.

Glosario de términos relacionados con el proceso DCDB

Agrupación (Clustering)

Los algoritmos de agrupación encuentran grupos de elementos que son semejantes. Por ejemplo, la agrupación podría ser usada por una universidad para agrupar estudiantes de acuerdo a su origen social, edad y tipo de trayectoria. Divide un conjunto de datos de modo que los registros con contenidos similares estén en el mismo grupo y que los grupos sean tan diferentes como sea posible. Debido a que las categorías no se especifican, este algunas veces se denomina como aprendizaje no supervisado.

Algoritmos genéticos (Genetic algorithms)

Método basado en computadora para la generación y prueba de combinaciones de parámetros posibles de entrada para encontrar el resultado óptimo. Usa procesos basados en conceptos de evolución natural como la combinación genética, mutación y selección natural.

Análisis exploratorio (Exploratory analysis)

Búsqueda en los datos para descubrir relaciones no detectadas previamente. Las herramientas de análisis exploratorio comúnmente auxilian al usuario en la creación de tablas y despliegues gráficos.

Análisis discriminante (Discriminant Analysis)

Método estadístico basado en máxima verosimilitud para determinar los límites que separan los datos en categorías.

Antecedente (Antecedent)

El primer elemento (o lado izquierdo) cuando se define una asociación entre dos variables, se denomina antecedente. Por ejemplo, en la relación "Cuando un estudiante es el primogénito en su familia, tiene un desempeño alto 50% de las veces"; en este ejemplo, "es primogénito" es el antecedente.

Aprendizaje (Learning)

Modelos de entrenamiento (estimación de sus parámetros) basados en los datos existentes.

Aprendizaje supervisado (Supervised learning)

La colección de técnicas donde el análisis utiliza una variable dependiente bien definida (conocida). Todas las técnicas de regresión y clasificación son supervisadas.

Aprendizaje No Supervisado (Unsupervised learning)

Colección de técnicas donde las agrupaciones de los datos se definen sin el uso de una variable dependiente. El análisis Cluster es un ejemplo.

Árbol de decisión (Decision tree)

Forma arbórea de representación de una colección de reglas jerárquicas que conducen a una clase o valor.

Árbol de regression (regression tree)

Árbol de decisión que predice valores de variables continuas.

Asociaciones (Associations)

Algoritmo de asociación que genera reglas que describen cómo ocurren los eventos conjuntamente. Por ejemplo, "Cuando un estudiante es el primogénito en su familia, tiene un desempeño escolar alto el 50% de las veces". Las relaciones se expresan comúnmente con intervalos de confianza.

Binning

Actividad de preparación de los datos que convierte datos continuos a discretos reemplazando un valor de un rango continuo con un identificador bin, donde cada bin representa un rango de valores. Por ejemplo, la edad podría ser convertida a bins como 20 ó menos, 21-40, 41-65 y más de 65.

Búsqueda Hill Climbing (Hill Climbing Search)

Técnica de optimización simple que modifica una solución propuesta por una pequeña cantidad y después la acepta si es mejor que la solución previa.

CART

Árboles de Clasificación y Regresión. CART es un método de división de las variables independientes en grupos pequeños y ajuste de una función constante a los conjuntos de datos pequeños. En árboles categóricos, la función constante es la que toma sobre un conjunto de valores pequeños finito (ejemplo, Y o N, bajo o medio o alto). En los árboles de regresión, el valor promedio de la respuesta es ajustar a conjuntos de datos conectados pequeños.

Clasificación (Classification)

Proceso de encontrar un conjunto de modelos (o funciones) que describan y distingan clases o conceptos de datos, con el propósito de ser capaces de usar el modelo para predecir la clase de objetos cuya etiqueta de clase sea desconocida.

Confianza (Confidence)

La confianza de la regla "B dado A" es una medida de lo probable que resulte que B ocurra cuando A ha ocurrido. Se expresa como porcentaje, donde 100% significa que B siempre ocurre si A ha ocurrido. Los estadísticos la refieren como la probabilidad condicional de B dado A. Cuando se usa con las reglas de asociación, el término confianza es observacional en lugar de predictivo.

Continuo (Continuous)

Los datos continuos pueden tener cualquier valor en un intervalo de números reales. Es decir, los valores no tienen que ser enteros. Continuo es lo opuesto a discreto o categórico.

CHAID

Algoritmo para ajustar árboles categóricos. Se basa en el estadístico Chi-cuadrada para dividir los datos en conjuntos de datos pequeños conectados.

Chi-Cuadrada (Chi-squared χ^2)

Estadístico que evalúa qué tan bien se ajusta un modelo a los datos. En MD, es más comúnmente usado para encontrar subconjuntos homogéneos para ajustar árboles categóricos como en el caso de CHAID.

Datos (Data)

Valores colectados a través de la observación o medición, comúnmente organizados para el análisis o toma de decisiones. En palabras más simples, datos son hechos, transacciones e imágenes.

DMG (Data Mining Group)

Consorcio de vendedores de MD para desarrollar estándares de MD. Han desarrollado "Predictive Model Markup Language" (PMML), que es un lenguaje que se basa en XML y que proporciona una forma rápida y fácil de definir modelos predictivos. PMML permite también compartir modelos entre diferentes vendedores de aplicaciones al proporcionar un método independiente de definición de modelos.

Datos atípicos (Outliers)

Técnicamente, los datos atípicos son datos que no provienen de la población supuesta –por ejemplo, un dato no numérico cuando se esperan datos con valores numéricos únicamente. El uso más casual se refiere a los elementos de datos que caen fuera de los límites que encierra la mayoría de los datos restantes en el conjunto.

Datos ausentes (Missing data)

Valores de los datos que pueden estar ausentes porque no fueron medidos, no respondidos, desconocidos o se perdieron. Los métodos de MD varían en la forma en que tratan los valores ausentes. Comúnmente, ignoran los valores ausentes u omiten los registros que contienen los valores ausentes o los reemplazan con la moda o media, o infieren los mismos a partir de los valores existentes.

Datos categóricos (Categorical data)

Los datos categóricos se ajustan a un número pequeño de categorías discretas (opuesto a continuos). Los datos categóricos pueden ser no ordenados (nominales) como el género o la ciudad, u ordenados (ordinales) como alta, media o baja, para el caso de las temperaturas, por ejemplo.

Datos de entrenamiento (Training data)

Conjunto de datos utilizado para estimar o entrenar un modelo.

Datos externos (External data)

Datos no colectados por la organización, por ejemplo los datos disponibles de un libro de referencia, una fuente gubernamental o una base de datos propietaria.

Datos internos (Internal data)

Datos colectados por la organización, por ejemplo los datos de las operaciones y los clientes.

Datos no aplicables (Non-applicable data)

Valores ausentes que serían lógicamente imposibles (ejemplo, hombres embarazados) o son obviamente no importantes.

Datos de prueba (Test data)

Conjunto de datos independiente del conjunto de entrenamiento, usado para poner a punto los estimadores de los parámetros del modelo (pesos).

Deducción (Deduction)

La deducción infiere información que es una consecuencia lógica de los datos.

Depuración, limpieza (Cleaning, cleansing)

Paso de preparación de los datos para la actividad de minería de datos. Por ejemplo, la eliminación del ruido en los datos previo a la etapa de minería de datos.

Descubrimiento de secuencias (Sequence discovery)

Lo mismo que asociación, excepto que la secuencia de tiempo de los eventos se considera también. Por ejemplo, 20% de la gente que compra un VCR compra también una cámara de video aproximadamente en cuatro meses".

Dimensión (Dimension)

Cada atributo de un caso u ocurrencia en los datos donde se aplica la minería. Almacenado como un campo en un registro de archivo plano o una columna de una tabla de base de datos relacional.

Discreta (Discrete)

Elemento de datos que tiene un conjunto finito de valores. Discreto es lo opuesto a continuo.

Entrenamiento (Training)

Otro término para la estimación de los parámetros de un modelo basado en el conjunto de datos disponibles.

Entropía (Entropy)

Forma de medir la variabilidad en forma diferente de la varianza estadística. Algunos árboles de decisión dividen los datos en grupos basados en una entropía mínima.

Estandarizar (Standardize)

Colección de datos numéricos se estandariza sustrayendo una medida de ubicación central (como la media o mediana) y dividiéndola por alguna medida de dispersión (como la desviación estándar, rango intercuartil o rango). Esto produce datos con un histograma de forma semejante a los datos originales, pero con los valores centrados alrededor del 0. Es útil algunas veces hacer esto con las entradas de redes neuronales y también las entradas en modelos de regresión. (Ver también *Normalizar*).

Exactitud (Accuracy)

Factor importante en la evaluación del éxito de la minería de datos. Cuando se aplica a los datos, la exactitud se refiere a la tasa de valores correctos en los datos. Cuando se aplica a los modelos, la exactitud se refiere al grado de ajuste

de las predicciones del modelo. Debido a que la exactitud no incluye la información de costo, es posible que un modelo menos exacto sea más efectivo en relación al costo. Ver también *precisión*.

Formato de los datos (Data format)

Forma de los datos en la base de datos.

Función de activación (Activation function)

Función usada por un nodo en una red neuronal para transformar los datos de entrada de cualquier dominio de valores en un rango finito de valores. La idea original fue aproximar la forma en que las neuronas se activan y la función de activación tomaba el valor 0 hasta que la entrada llegaba a ser grande y el valor saltaba a 1. La discontinuidad de esta función 0- ó -1 provocó problemas matemáticos y las funciones sigmoides (función logística) se utilizan ahora.

Grado de ajuste (Degree of fit)

Medida de qué tan bien se ajusta el modelo a los datos de entrenamiento.

Hoja (Leaf)

Nodo que no se puede dividir más -la agrupación terminal – en un árbol de decisión o clasificación.

Inducción (Induction)

Técnica que infiere generalizaciones de la información en los datos.

Interacción (Interaction)

Dos variables independientes interactúan cuando los cambios en el valor de una cambian el efecto sobre la variable dependiente de la otra.

Interfaz del Programa de Aplicación (API)

Cuando un sistema de software caracteriza una API, proporciona un medio para que los programas escritos fuera del sistema puedan interactuar con el sistema para realizar funciones adicionales. Por ejemplo, un sistema de software de minería de datos podría tener una API que permita que los programas escritos por el usuario realicen tareas como la de extraer datos, realizar análisis estadísticos adicionales, crear gráficos especializados, generar un modelo, o hacer una predicción a partir de un modelo.

Interfaz gráfica de usuario (GUI)

Interfaz gráfica de usuario.

Lógica difusa (Fuzzy logic)

La lógica difusa se aplica a conjuntos difusos donde la membresía en un conjunto difuso es una probabilidad, no necesariamente 0 ó 1. La lógica no difusa manipula resultados que son verdaderos o falsos.

Matriz de confusion (Confusion matrix)

Matriz que muestra las frecuencias de la clase real *versus* la predicha. Muestra no solo qué tan bien predice el modelo, sino también presenta los detalles requeridos para ver exactamente donde podrían no estar funcionando bien las cosas.

Minería de datos (Data Mining)

Actividad de extracción de información cuya meta es descubrir información oculta contenida en bases de datos. Usando una combinación de técnicas de análisis estadístico, de aprendizaje máquina, técnicas de modelación y tecnología de bases de datos, encuentra patrones y relaciones sutiles en los datos e infiere reglas que permiten la predicción de resultados futuros. Aplicaciones comunes incluyen la segmentación del mercado, perfiles de clientes, detección de fraudes y análisis de riesgos de créditos.

Mínimos cuadrados (Least squares)

Método común de entrenamiento (estimación) de los pesos (parámetros) de un modelo seleccionando los pesos para minimizar la suma de la desviación cuadrada de los valores predichos del modelo a partir de los valores observados de los datos.

Máxima Verosimilitud (Maximum likelihood)

Método de entrenamiento o estimación. El estimador de máxima verosimilitud de un parámetro es el valor de un parámetro que maximiza la probabilidad de que los datos provengan de la población definida por el parámetro.

Media (Mean)

Valor del promedio aritmético de una colección de datos numéricos.

Mediana (Median)

Valor del centro de una colección de datos ordenados. En otras palabras, el valor con el mismo número de elementos arriba y abajo de él.

Moda (Mode)

Valor más común de un conjunto de datos. Si más de un valor aparece el mismo número de veces, los datos son multimodales.

Modelo (Model)

Una función importante de la minería de datos es la producción de un modelo. Un modelo puede ser descriptivo o predictivo. Un modelo descriptivo ayuda en la comprensión de los procesos o comportamientos subyacentes. Un modelo predictivo es una ecuación o conjunto de reglas que hace posible predecir un valor no visto o no medido (la variable dependiente o de salida) a partir de otra(s), los valores conocidos (variables independientes o de entrada). La forma de la ecuación o reglas se sugiere por la minería de datos. Se utiliza alguna técnica de entrenamiento o estimación para estimar los parámetros de la ecuación o las reglas.

Muestreo (Sampling)

Creación de un subconjunto de los datos del conjunto completo. El muestreo aleatorio intenta representar el todo seleccionando la muestra por medio de mecanismos aleatorios.

Multiprocesamiento Simétrico (SMP)

Multiprocesamiento simétrico es una configuración de computadora donde muchas CPU's comparten un sistema operativo común, memoria principal y discos. Pueden trabajar en diferentes partes de un problema al mismo tiempo

Nodo (node)

Punto de decisión en un árbol de decisión. También, un punto en una red neuronal que combina la entrada de otros nodos y produce una salida por medio de la aplicación de una función de activación.

Normalización (Normalize)

Una colección de datos numéricos es normalizado cuando se sustrae el valor mínimo de todos los valores y se dividen por el rango de los datos. Esto produce datos con un histograma de igual forma pero con todos los valores entre 0 y 1. Es útil hacer esto para todas las entradas de las redes neuronales y también para entradas en otros modelos de regresión.

OLEDB DM (OLEDB for Data Mining)

Intento de Microsoft por agregar una API de MD a SQL. Soporta la mayoría de los algoritmos de MD populares. Simplifica la API de MD proporcionando una interfaz como la de SQL. Un modelo de MD es parecido a una tabla en MS-SQL; las consultas para agregar filas a una tabla también trabajan sobre modelos de MD.

Patrón (Pattern)

Un patrón puede ser una relación entre dos variables. Las técnicas de minería de datos incluyen el descubrimiento automático de patrones que haga posible detectar relaciones no lineales complicadas en los datos. Los patrones no son lo mismo que causalidad.

Podar (Pruning)

Eliminación de divisiones de nivel inferior o sub-árboles completos en un árbol de decisión. Este término se usa también para describir algoritmos que ajustan la topología de una red neuronal removiendo nodos ocultos (es decir, podando).

Precisión (Precision)

La precisión del estimador de un parámetro es una medida de qué tan variable sería el estimador sobre otros conjuntos de datos similares. Un estimador muy preciso es el que no varía mucho sobre diferentes conjuntos de datos. La precisión no mide la exactitud. La exactitud es una medida de qué tan cercano está el estimador del valor real del parámetro. La exactitud se mide por la distancia promedio sobre diferentes conjuntos de datos del estimador del valor real. Los estimadores pueden ser exactos pero no precisos, o precisos pero no exactos. Un estimador preciso pero inexacto con frecuencia es sesgado, con el sesgo igual a la distancia promedio del valor real del parámetro.

Prevalencia (Prevalence)

Frecuencia de aparición conjunta de la colección de elementos en una asociación, medida como porcentaje de todas las transacciones.

Procesamiento masivamente paralelo (MPP)

Configuración de computadora que es capaz de usar cientos o miles de CPU's simultáneamente. En un MPP cada nodo podría ser una simple CPU o una colección de CPU's SMP. Una colección de nodos SMP es algunas veces denominada cluster SMP. Cada nodo tiene su propia copia del sistema operativo, memoria y almacenamiento en disco y existe un mecanismo de intercambio de datos o procesos de modo que cada computadora puede trabajar en una parte diferente del problema. El software debe ser escrito para sacar partido de esta arquitectura.

Procesamiento Analítico En Línea (OLAP)

Las herramientas de procesamiento analítico en línea dan al usuario la capacidad de realizar análisis multidimensionales de los datos.

Procesamiento paralelo (Parallel processing)

Varias computadoras o CPU's ligadas de modo que cada una pueda calcular simultáneamente.

Pronosticabilidad (Predictability)

Algunos vendedores de minería de datos usan pronosticabilidad de asociaciones o secuencias para expresar lo mismo que confianza.

Prueba de error (Test error)

Estimador del error basado en la diferencia entre las predicciones de un modelo sobre un conjunto de datos de prueba y los valores observados en el conjunto de los datos de prueba cuando el conjunto de prueba no fue utilizado para entrenar el modelo.

R^2 (r-squared)

Número entre 0 y 1 que mide qué tan bien se ajusta un modelo a sus datos de entrenamiento. Uno es un ajuste perfecto; sin embargo, cero implica que el modelo no tiene capacidad predictiva. Se calcula como la covarianza entre los valores predichos y observados dividido por las desviaciones estándar de los valores predichos y observados.

Rango (Range)

La diferencia entre el valor máximo y el mínimo. Por otra parte, el rango puede incluir el mínimo y máximo, como en "el valor va de 2 a 8".

Red Neuronal (Neural network)

Técnica de modelación compleja no lineal basada en el modelo de una neurona humana. Una red neuronal se usa para predecir salidas (variables dependientes) de un conjunto de entradas (variables independientes) tomando las combinaciones de las entradas y después haciendo transformaciones no lineales de las combinaciones lineales usando una función de activación. Puede mostrarse teóricamente que dichas combinaciones y transformaciones pueden

aproximarse virtualmente a algún tipo de función de respuesta. Además, las redes neuronales usan grandes números de parámetros para aproximarse a algún modelo. Las redes neuronales son con frecuencia aplicadas para predecir resultados futuros basados en la experiencia previa. Por ejemplo, una aplicación de red neuronal podría usarse para predecir quién responderá a una publicidad directa.

Re-muestreo (Bootstrapping)

Conjuntos de datos de entrenamiento que son creados por re-muestreo con reemplazamiento del conjunto de entrenamiento original, de este modo los registros (datos) podrían aparecer más de una vez. En otras palabras, este método trata una muestra como si fuera la población completa. Con frecuencia, los estimadores finales se obtienen tomando el promedio de los estimadores de cada uno de los conjuntos de prueba bootstrap.

Retropropagación (Backpropagation)

Método de entrenamiento usado para calcular los pesos en una red neuronal a partir de los datos.

Ruido (Noise)

Diferencia entre un modelo y sus predicciones. Algunas veces los datos se consideran ruido cuando contienen errores. Por ejemplo, contienen muchos valores ausentes o incorrectos o cuando existen columnas extrañas.

SABD (DBMS)

Sistemas de Administración de Bases de datos.

SABDR (RDBMS)

Sistema de Administración de Bases de datos Relacional.

Sesgo (Bias)

En una red neuronal, el sesgo se refiere a los términos constantes en el modelo. (observe que sesgo tiene un significado diferente para la mayoría de los analistas de datos). Ver también precisión.

SEMMA (Simple, Explore, Modify, Model, Assess)

Modelo de procesos propuesto por SAS. Este modelo de procesos guía el uso del software SAS.

SPSS

SPSS propuso el modelo de procesos de las 5 A's (Assess, Access, Analyse, Act and Automate). Ahora SPSS es parte del grupo CRISP-DM.

Significancia (Significance)

Medida de la probabilidad del soporte que los datos proporcionan a un determinado resultado (usualmente de una prueba estadística). Si se expresa la significancia de un resultado como de .05, significa que existe sólo una probabilidad de .05 de que el resultado pudiera haber sucedido sólo por casualidad. Muy bajos niveles de significancia (menores que .05) con frecuencia se toman como evidencia de que el modelo de minería de datos debe aceptarse ya

que los eventos con una probabilidad muy baja raras veces ocurren. De esta forma, si el estimador de un parámetro en un modelo mostrara una significancia de .01 sería evidencia de que el parámetro debe estar en el modelo.

Sobreajuste (Overfitting)

Tendencia de algunas técnicas de modelación a asignar importancia a variaciones aleatorias en los datos declarándolas patrones importantes.

Soporte (Support)

Frecuencia de aparición conjunta de la colección de elementos en una asociación, medida como porcentaje de todas las transacciones.

TASF (The Analytic Solutions Forum)

Consorcio cuya misión es establecer un criterio de desempeño orientado a la solución y requerimientos de interoperabilidad dentro y entre clases de modelos de soporte en las decisiones como OLAP, la MD y la visualización de datos.

Topología (Topology)

Para una red neuronal, la topología se refiere al número de capas y el número de nodos en cada capa.

Transformación (Transformation)

Re-expresión de los datos, como la agregación, normalización, cambiar su unidad de medida o tomar el logaritmo de cada elemento de los datos.

Validación (Validation)

Proceso de prueba de los modelos con un conjunto de datos diferente del conjunto de entrenamiento.

Validación cruzada (Cross validation)

Método de estimación de la exactitud de un modelo de clasificación o regresión. El conjunto de datos se divide en varias partes, con cada parte usada para probar un modelo ajustado de las partes restantes.

Varianza (Variance)

La medida estadística de dispersión más comúnmente utilizada. El primer paso es elevar al cuadrado las desviaciones de cada elemento de los datos con respecto a su valor promedio. Después, se calcula el promedio de las desviaciones al cuadrado para obtener una medida de la variabilidad total.

Variable dependiente (Dependent variable)

Variables predicha por la ecuación o las reglas del modelo usando las variables independientes (de entrada o predictoras).

Variable independiente (Independent variable)

Variables que se utilizan en la ecuación o reglas del modelo para predecir la variable respuesta (dependiente).

Vecino más cercano (K-nearest neighbor)

Método de clasificación que clasifica un punto calculando las distancias entre el punto y los puntos en el conjunto de datos de entrenamiento. Después asigna el

punto a la clase que es más común entre sus vecinos más cercanos (donde k es un entero).

Herramientas de visualización (Visualization)

Despliegan gráficamente los datos para facilitar mejor la comprensión de su significado. Las capacidades gráficas van desde los gráficos de dispersión simple a representaciones multi-dimensionales complejas.

Glosario de términos relacionados con la Investigación Educativa

Capital cultural. Desde el punto de vista de la teoría del capital cultural de Bourdieu, está relacionado, entre otros, con el campo educativo. El concepto se operacionaliza considerando el nivel de escolaridad de la madre del estudiante y los hábitos de lectura (cuánto y qué lee).

Capital económico. Desde el punto de vista de la teoría del capital cultural de Bourdieu, el concepto de capital económico se refiere a los recursos que caracterizan la posición del agente en el espacio social. Por excelencia, el capital estaría constituido por el ingreso y las diversas titularidades que los agentes puedan fácilmente transformar en dinero. El concepto se operacionaliza considerando el ingreso familiar mensual y el nivel de hacinamiento.

Característica de entrada. Variables relevantes que explican parte de la variación en la variable respuesta y que son medidas ya sea al nivel del estudiante o al nivel de la escuela.

Contextualización sociocultural. Se construye mediante diferentes técnicas aunque todas utilizan información primaria registrada directamente de las familias de los estudiantes a través de encuestas aplicadas especialmente con este fin. El indicador de clase más utilizado de todos es el nivel educativo de la madre del alumno y en segundo lugar, una escala de posesiones de confort o equipamiento del hogar. En resumen, se trata de un control de los resultados académicos de los alumnos con algún indicador que resuma las características socioculturales de los hogares de origen. El método de contextualización por indicadores sociofamiliares es el que más consensos genera entre los investigadores y el que progresivamente está siendo adoptado por los sistemas de evaluación de aprendizajes.

Desempeño bruto. Estimación del desempeño de las escuelas a través del logro académico de sus estudiantes considerando únicamente los puntajes crudos.

Desempeño neto. Estimación del desempeño de las escuelas a través del logro académico de sus estudiantes controlados por características de entrada relevantes.

Escuelas eficaces (School effectiveness). Término usado para describir la investigación educativa enfocada en la exploración de las diferencias entre y dentro de escuelas. Su principal objetivo es obtener conocimiento con respecto a las relaciones entre factores predictores y respuesta.

Índice sociofamiliar. Índice construido mediante un análisis factorial, considerando los indicadores que representan las condiciones económicas y culturales de los hogares de origen. (Ver también, Capital cultural y Capital económico).

Logro académico. Nivel de desempeño logrado por un estudiante en una prueba de conocimientos estandarizada.

Anexos

Anexo A

Análisis Exploratorio de los datos de la hoja de registro

Tablas de distribución de frecuencias

Código	Escolaridad de la madre	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
1	No sabe leer y escribir	5335	2.2	2.7	2.7
2	Sabe leer y escribir (sin concluir primaria)	24290	10.2	12.2	14.9
3	Primaria	54300	22.9	27.4	42.3
4	Capacitación técnica (posterior a la primaria)	9260	3.9	4.7	47.0
5	Secundaria	45276	19.1	22.8	69.8
6	Capacitación técnica (posterior a la secundaria)	20694	8.7	10.4	80.2
7	Profesional técnico	9697	4.1	4.9	85.1
8	Bachillerato o preparatoria	15375	6.5	7.7	92.8
9	Normal (no licenciatura)	5282	2.2	2.7	95.5
10	Licenciatura	8042	3.4	4.1	99.6
11	Posgrado	883	0.4	0.4	100.0
	Total	**198434**	**83.5**	**100.0**	
	No respuesta	39146	16.5		
	Total	237580	100		

Tabla de frecuencias de la escolaridad de la madre

Código	Escolaridad	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje Acumulado
1	1. No sabe leer y escribir	2133	0.9	1.2	1.2
2	2. Sabe leer y escribir (sin concluir la primaria)	17567	7.4	9.5	10.7
3	3. Primaria	39806	16.8	21.6	32.3
4	4. Capacitación técnica (posterior a la primaria)	6387	2.7	3.5	35.8
5	5. Secundaria	46050	19.4	25.0	60.8
6	6. Capacitación técnica (posterior a la secundaria)	12276	5.2	6.7	67.5
7	7. Profesional técnico	7804	3.3	4.2	71.7
8	8. Bachillerato o preparatoria o vocacional	26563	11.2	14.4	86.1
9	9. Normal (no licenciatura)	4342	1.8	2.4	88.5
10	10. Licenciatura	18672	7.9	10.1	98.6
11	11. Posgrado	2676	1.1	1.5	100.0
	Total	**184276**	**77.6**	**100.0**	
	No respuesta	53304	22.4		
	Total	237580	100.0		

Tabla de frecuencias de la escolaridad del padre

Código	Ingreso familiar	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje Acumulado
1	Menos de $ 1000	20415	8.6	9.9	9.9
2	De $ 1001 a $ 2000	51611	21.7	25.0	34.9
3	De $ 2001 a $ 3000	42047	17.7	20.3	55.2
4	De $ 3001 a $ 4000	29346	12.4	14.2	69.4
5	De $ 4001 a $ 5000	18713	7.9	9.1	78.5
6	De $ 5001 a $ 6000	12504	5.3	6.1	84.6
7	De $ 6001 a $ 7000	7605	3.2	3.7	88.3
8	De $ 7001 a $ 8000	6771	2.8	3.3	91.6
9	De $ 8001 a $ 9000	4238	1.8	2.1	93.7
10	De $ 9001 a $ 10000	4332	1.8	2.1	95.8
11	De $ 10001 a $ 12500	3509	1.5	1.7	97.5
12	De $ 12501 a $ 15000	2507	1.1	1.2	98.7
13	De $ 15001 a $ 17500	1294	0.5	0.6	99.3
14	De $ 17501 a $ 20000	924	0.4	0.4	99.7
15	$ 20001 ó más	849	0.4	0.4	100.0
	Total	**206665**	**87.0**	**100.0**	
	No respuesta	30915	13.0		
	Total	237580	100		

Tabla de frecuencias del ingreso familiar mensual

Código	Número de personas	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje válido acumulado
1	1	207	0.1	0.1	0.1
2	2	3875	1.6	1.8	1.9
3	3	16981	7.1	8.1	10.0
4	4	51410	21.6	24.4	34.4
5	5	64124	27.0	30.5	64.9
6	6	34517	14.5	16.4	81.3
7	7	16728	7.0	7.9	89.2
8	8	8573	3.6	4.1	93.3
9	9	4841	2.0	2.3	95.6
10	Más de 9	9303	3.9	4.4	100.0
	Total	**210559**	**88.6**	**100.0**	
	No respuesta	27021	11.4		
	Total	237580	100.0		

Tabla de frecuencias del número de personas que viven la casa

Respuesta	Frecuencia	Porcentaje	Porcentaje
SÍ	24492	10.3	11.7
No	184649	77.7	88.3
Total	**209141**	**88.0**	**100.0**
No respuesta	28439	12.0	
Total	237580	100.0	

Tabla de frecuencias del número de estudiantes que trabajan

Anexo B

Indicadores estadísticos de tendencia central y dispersión

Asignatura	Media	Cuartiles		Desviación estándar
		Inferior	Superior	
Razonamiento verbal	47.66	33.33	62.50	18.73
Español	42.01	30.00	60.00	19.49
Historia	45.58	30.00	60.00	19.35
Geografía	45.27	30.00	60.00	19.66
Civismo	49.83	40.00	60.00	18.82
Razonamiento matemático	51.45	37.50	66.67	18.43
Matemáticas	42.50	30.00	60.00	20.84
Física	43.05	30.00	60.00	19.34
Química	43.56	30.00	60.00	20.37
Biología	46.73	30.00	60.00	19.62
Global	46.59	36.72	55.47	13.55

Estadísticas descriptivas del desempeño en el examen de admisión

Sexo	Media	Desviación
Mujeres	45.0	13.1
Hombres	48.8	13.8
Global	46.9	13.6

Desempeño global en el examen por género

Horas de estudio a la semana	Media	Desv. Estándar
No respuesta	44.3	13.8
0	44.5	13.3
1	43.1	12.4
2	44.0	12.7
3	45.6	13.1
4	47.7	13.4
5	49.0	13.6
6	50.1	13.6
7	50.6	13.9
8	50.2	13.6
9	51.6	14.0
10	50.8	13.8
Más de 10	52.6	14.2

Desempeño global y horas de estudio a la semana

Escolaridad de la madre	Madre		Padre	
	Media	Desv. Estándar	Media	Desv. Estándar
No respuesta	41.2	11.9	43.6	12.8
No sabe leer y escribir	40.8	12.1	40.4	11.9
Sabe leer y escribir (sin concluir la primaria)	44.8	12.8	44.1	12.6
Primaria	45.2	12.9	44.8	12.9
Capacitación técnica (posterior a la primaria)	47.1	13.8	45.7	13.5
Secundaria	46.8	13.2	46.4	13.1
Capacitación técnica (posterior a la secundaria)	49.6	13.6	48.6	13.5
Profesional técnico	50.8	13.7	50.0	13.7
Bachillerato o preparatoria o vocacional	51.0	13.8	49.6	13.7
Normal (no licenciatura)	52.4	14.0	51.3	13.6
Licenciatura	55.7	14.1	53.3	14.2
Posgrado	57.4	14.2	55.0	14.5

Estadísticas descriptivas del desempeño en el examen
por escolaridad de los padres

Número de personas que habitan la casa	Media	Desv. Estándar
No respuesta	43.6	12.9
1	47.5	13.4
2	49.3	13.9
3	48.6	13.8
4	48.8	13.9
5	47.2	13.5
6	45.8	13.2
7	44.9	12.9
8	44.4	13.0
9	44.1	12.6
Más de 9	43.1	12.6

Estadísticas descriptivas del desempeño en el examen
por nivel de hacinamiento

Ingreso familiar	Media	Desviación estándar
No respuesta	42.9	12.9
Menos de $ 1000	40.3	11.6
De $ 1001 a $ 2000	44.6	12.7
De $ 2001 a $ 3000	46.9	13.2
De $ 3001 a $ 4000	48.4	13.5
De $ 4001 a $ 5000	49.4	13.6
De $ 5001 a $ 6000	50.0	13.8
De $ 6001 a $ 7000	50.8	13.7
De $ 7001 a $ 8000	50.8	14.1
De $ 8001 a $ 9000	51.5	14.3
De $ 9001 a $ 10000	51.5	14.1
De $ 10001 a $ 12500	52.8	14.0
De $ 12501 a $ 15000	53.4	14.7
De $ 15001 a $ 17500	52.9	14.5
De $ 17501 a $ 20000	53.4	14.8
$ 20001 ó más	53.4	15.2

Estadísticas descriptivas del desempeño en el examen por ingreso familiar

Anexo C

Descripción de la estructura de la base de datos

El siguiente cuadro presenta la estructura de la base de datos original correspondiente a la hoja de registro. El contenido de la primera columna corresponde al nombre de la pregunta (nombre del campo), seguido por el tipo de campo (carácter, numérico, etc.). La tercera columna contiene la longitud del campo y la cuarta registra todos los valores válidos posibles para la pregunta. Finalmente, la quinta columna proporciona una descripción más extensa de las preguntas y de sus posibles respuestas.

Pregunta	Tipo	Longitud	Valor	Descripción
****** Pregunta 1 ******				Tus actividades al estudiar fuera del horario escolar
prexa_01	Carácter	1		**Al iniciar, identifico lo que necesito estudiar y hago un plan de trabajo**
			1	Casi nunca
			2	Pocas veces
			3	Muchas veces
			4	Casi siempre
prexa_02	Carácter	1		**Estudio principalmente con mis apuntes de clase**
			1	Casi nunca
			2	Pocas veces
			3	Muchas veces
			4	Casi siempre
prexa_03	Carácter	1		**Utilizo las monografías que venden en las papelerías**
			1	Casi nunca
			2	Pocas veces
			3	Muchas veces
			4	Casi siempre
prexa_04	Carácter	1		**Estudio principalmente con el libro de texto de la materia**
			1	Casi nunca
			2	Pocas veces
			3	Muchas veces
			4	Casi siempre
prexa_05	Carácter	1		**Utilizo enciclopedias, diccionarios y atlas**
			1	Casi nunca
			2	Pocas veces
			3	Muchas veces
			4	Casi siempre

Pregunta	Tipo	Longitud	Valor	Descripción
prexa_06	Carácter	1		**Realizo resúmenes y/o cuadros sinópticos**
			1	Casi nunca
			2	Pocas veces
			3	Muchas veces
			4	Casi siempre
prexa_07	Carácter	1		**Estudio principalmente con los apuntes de mis compañeros**
			1	Casi nunca
			2	Pocas veces
			3	Muchas veces
			4	Casi siempre
prexa_08	Carácter	1		**Resuelvo ejercicios para reafirmar lo estudiado**
			1	Casi nunca
			2	Pocas veces
			3	Muchas veces
			4	Casi siempre
prexa_09	Carácter	1		**Solicito apoyo a mis padres o hermanos**
			1	Casi nunca
			2	Pocas veces
			3	Muchas veces
			4	Casi siempre
prexa_10	Carácter	1		**Solicito asesoría a mis maestros**
			1	Casi nunca
			2	Pocas veces
			3	Muchas veces
			4	Casi siempre
prexa_11	Carácter	1		**Estudio en equipo con mis compañeros de clase**
			1	Casi nunca
			2	Pocas veces
			3	Muchas veces
			4	Casi siempre
hora_est	Carácter	2		**A la semana ¿cuántas horas dedicas al estudio fuera del horario escolar ¿**
			00	0
			01	1
			02	2
			03	3
			04	4
			05	5
			06	6
			07	7

Pregunta	Tipo	Longitud	Valor	Descripción
			08	8
			09	9
			10	10
			11	Más de 10
ext_sec	Carácter	1		**¿Presentaste algún examen extraordinario en la secundaria?**
			1	Sí
			2	No
rep_sec	Carácter	1		**¿Repetiste algún año escolar en la secundaria?**
			1	Sí
			2	No
****** Pregunta 5 *****				**Señala que tanto gusto tuviste por cada una de las asignaturas**
gus_mat	Carácter	1		**Matemáticas**
			1	Me gustó mucho
			2	Me gustó
			3	Me gustó poco
			4	No me gustó
gus_ifq	Carácter	1		**Introducción a la Física y la química**
			1	Me gustó mucho
			2	Me gustó
			3	Me gustó poco
			4	No me gustó
gus_fis	Carácter	1		**Física**
			1	Me gustó mucho
			2	Me gustó
			3	Me gustó poco
			4	No me gustó
gus_qui	Carácter	1		**Química**
			1	Me gustó mucho
			2	Me gustó
			3	Me gustó poco
			4	No me gustó
gus_bio	Carácter	1		**Biología**
			1	Me gustó mucho
			2	Me gustó
			3	Me gustó poco
			4	No me gustó

Pregunta	Tipo	Longitud	Valor	Descripción
gus_esp	Carácter	1		**Español**
			1	Me gustó mucho
			2	Me gustó
			3	Me gustó poco
			4	No me gustó
gus_his	Carácter	1		**Historia**
			1	Me gustó mucho
			2	Me gustó
			3	Me gustó poco
			4	No me gustó
gus_geo	Carácter	1		**Geografía**
			1	Me gustó mucho
			2	Me gustó
			3	Me gustó poco
			4	No me gustó
gus_civ	Carácter	1		**Civismo**
			1	Me gustó mucho
			2	Me gustó
			3	Me gustó poco
			4	No me gustó
gus_lex	Carácter	1		**Lengua extranjera**
			1	Me gustó mucho
			2	Me gustó
			3	Me gustó poco
			4	No me gustó
gus_art	Carácter	1		**Expresión y apreciación artística**
			1	Me gustó mucho
			2	Me gustó
			3	Me gustó poco
			4	No me gustó
gus_efi	Carácter	1		**Educación física**
			1	Me gustó mucho
			2	Me gustó
			3	Me gustó poco
			4	No me gustó
gus_edt	Carácter	1		**Educación tecnológica**
			1	Me gustó mucho
			2	Me gustó
			3	Me gustó poco
			4	No me gustó

Pregunta	Tipo	Longitud	Valor	Descripción
gus_oed	Carácter	1		**Orientación educativa**
			1	Me gustó mucho
			2	Me gustó
			3	Me gustó poco
			4	No me gustó
gus_opt	Carácter	1		**Optativa**
			1	Me gustó mucho
			2	Me gustó
			3	Me gustó poco
			4	No me gustó
***** Pregunta 5 *****				**Señala que tan bién preparado te siente en cada una de las asignaturas**
pre_mat	Carácter	1		**Matemáticas**
			1	Muy mala
			2	Mala
			3	Buena
			4	Muy Buena
pre_ifq	Carácter	1		**Introducción a la Física y la química**
			1	Muy mala
			2	Mala
			3	Buena
			4	Muy Buena
pre_fis	Carácter	1		**Física**
			1	Muy mala
			2	Mala
			3	Buena
			4	Muy Buena
pre_qui	Carácter	1		**Química**
			1	Muy mala
			2	Mala
			3	Buena
			4	Muy Buena
pre_bio	Carácter	1		**Biología**
			1	Muy mala
			2	Mala
			3	Buena
			4	Muy Buena

Pregunta	Tipo	Longitud	Valor	Descripción
pre_esp	Carácter	1		**Español**
			1	Muy mala
			2	Mala
			3	Buena
			4	Muy Buena
pre_his	Carácter	1		**Historia**
			1	Muy mala
			2	Mala
			3	Buena
			4	Muy Buena
pre_geo	Carácter	1		**Geografía**
			1	Muy mala
			2	Mala
			3	Buena
			4	Muy Buena
pre_civ	Carácter	1		**Civismo**
			1	Muy mala
			2	Mala
			3	Buena
			4	Muy Buena
pre_lex	Carácter	1		**Lengua extranjera**
			1	Muy mala
			2	Mala
			3	Buena
			4	Muy Buena
pre_art	Carácter	1		**Expresión y apreciación artística**
			1	Muy mala
			2	Mala
			3	Buena
			4	Muy Buena
pre_efi	Carácter	1		**Educación física**
			1	Muy mala
			2	Mala
			3	Buena
			4	Muy Buena

Pregunta	Tipo	Longitud	Valor	Descripción
pre_edt	Carácter	1		**Educación tecnológica**
			1	Muy mala
			2	Mala
			3	Buena
			4	Muy Buena
pre_oed	Carácter	1		**Orientación educativa**
			1	Muy mala
			2	Mala
			3	Buena
			4	Muy Buena
pre_opt	Carácter	1		**Optativa**
			1	Muy mala
			2	Mala
			3	Buena
			4	Muy Buena
sec_est	Carácter	1		**¿Tienes la intención de seguir estudios superiores, después de la educación media superior?**
			1	Si
			2	No
			3	No sé
ins_ingr	Carácter	2		**¿A qué institución de educación superior te gustaría ingresar al terminar tus estudios de educación media superior?**
			01	Escuela de Educación Normal
			02	Instituto Politécnico Nacional
			03	Instituto Tecnológico (SEP.)
			04	Universidad Autónoma Metropolitana
			05	Universidad Nacional Autónoma de México
			06	Universidad Pedagógica Nacional
			07	Universidad privada
			08	Universidad pública estatal
			09	Universidad Tecnológica (SEP.)
			10	Otra
act_rea	Carácter	1		**En tu tiempo libre, ¿cuál de las siguientes actividades te gustaría realizar más?**
			1	Hacer deporte
			2	Ir al cine
			3	Pasear y platicar con los amigos
			4	Jugar juegos de mesa
			5	Jugar en la calle
			6	Leer
			7	Ver televisión
			8	Otra

Pregunta	Tipo	Longitud	Valor	Descripción
tpo_lec	Carácter	2		**A la semana ¿cuántas horas de tu tiempo libre inviertes en leer?**
			01	1
			02	2
			03	3
			04	4
			05	5
			06	6
			07	7
			08	8
			09	9
			10	10
			11	Más de 10
			12	Ninguna
tip_lec	Carácter	2		**¿Que tipo de lectura te gusta más?**
			01	Novela
			02	Poesía
			03	Ciencia y Tecnología
			04	Política, sociedad y economía
			05	Deportes
			06	Artes
			07	Espectáculos
			08	Noticias
			09	Historietas
			10	Otra
			11	Ninguna
tpo_tv	Carácter	2		**¿Cuántas horas de tu tiempo libre inviertes en ver televisión AL DÍA?**
			01	1
			02	2
			03	3
			04	4
			05	5
			06	6
			07	7
			08	8
			09	9
			10	10 ó más
			11	Ninguna
prog_tv	Carácter	2		**¿Que tipo de programas prefieres ver en la televisión?**
			01	Caricaturas
			02	Comedia juvenil

Pregunta	Tipo	Longitud	Valor	Descripción
			03	Culturales y/o educativos
			04	De concurso
			05	Deportes
			06	Noticieros
			07	Películas
			08	Series de aventuras o policiacas
			09	Telenovelas
			10	Videoclips y musicales
			11	Otros
***** Pregunta 13 *****				**En cada enunciado, elige la opción que mejor describe la forma de actuar en clase, del grupo de maestros que tuviste durante el *TERCER GRADO* de secundaria**
act_mt01	Carácter	1		**Asisten regularmente a clase**
			1	Casi todos
			2	Muchos
			3	Pocos
			4	Casi ninguno
act_mt02	Carácter	1		**Se dedican la mayor parte del tiempo de la clase a trabajar con los alumnos**
			1	Casi todos
			2	Muchos
			3	Pocos
			4	Casi ninguno
act_mt03	Carácter	1		**Califican injustamente a sus alumnos**
			1	Casi todos
			2	Muchos
			3	Pocos
			4	Casi ninguno
act_mt04	Carácter	1		**Se esfuerzan para que los alumnos comprendan lo tratado en clase**
			1	Casi todos
			2	Muchos
			3	Pocos
			4	Casi ninguno

Pregunta	Tipo	Longitud	Valor	Descripción
act_mt05	Carácter	1		**Ayudan a los alumnos en el desarrollo de sus trabajos en clase**
			1	Casi todos
			2	Muchos
			3	Pocos
			4	Casi ninguno
act_mt06	Carácter	1		**Realizan evaluaciones regularmente (al menos una vez al mes)**
			1	Casi todos
			2	Muchos
			3	Pocos
			4	Casi ninguno
act_mt07	Carácter	1		**Castigan injustamente a sus alumnos**
			1	Casi todos
			2	Muchos
			3	Pocos
			4	Casi ninguno
act_mt08	Carácter	1		**Promueven el trabajo en equipo entre los alumnos**
			1	Casi todos
			2	Muchos
			3	Pocos
			4	Casi ninguno
act_mt09	Carácter	1		**Organizan la clase tomando en cuenta la opinión e intereses de los alumnos**
			1	Casi todos
			2	Muchos
			3	Pocos
			4	Casi ninguno
act_mt10	Carácter	1		**Promueven la participación de todos los alumnos durante la clase**
			1	Casi todos
			2	Muchos
			3	Pocos
			4	Casi ninguno
act_mt11	Carácter	1		**Promueven en clase un ambiente amistoso y de confianza**
			1	Casi todos
			2	Muchos
			3	Pocos
			4	Casi ninguno

Pregunta	Tipo	Longitud	Valor	Descripción
cal_for	Carácter	1		**Califica la calidad de la formación que, en general, recibiste en tu secundaria**
			1	Excelente
			2	Buena
			3	Mala
			4	Deficiente
cua_hno	Carácter	1		**¿Cuántos hermanos tienes?**
			0	0
			1	1
			2	2
			3	3
			4	4
			5	5
			6	6
			7	7
			8	Más de 7
lug_ocu	Carácter	1		**¿Qué lugar ocupas entre tus hermanos (de mayor a menor)?**
			1	1º
			2	2º
			3	3º
			4	4º
			5	5º
			6	6º
			7	7º
			8	Posterior al 7º
edad_mad	Carácter	1		**¿Qué edad tienen tus padres?**
				Madre
			1	30 años o menos
			2	entre 31 y 40
			3	Entre 41 y 50
			4	Entre 51 y 60
			5	Más de 60 años
edad_pad	Carácter	1		**¿Qué edad tienen tus padres?**
				Padre
			1	30 años o menos
			2	entre 31 y 40
			3	Entre 41 y 50
			4	Entre 51 y 60
			5	Más de 60 años

Pregunta	Tipo	Longitud	Valor	Descripción
vive_con	Carácter	2		**¿Con quién vives actualmente?**
			01	Padre, madre y hermanos
			02	Padre y madre
			03	Sólo padre
			04	Sólo madre
			05	Sólo hermanos
			06	Padre y hermanos
			07	Madre y hermanos
			08	Otros familiares
			09	Cónyuge o pareja
			10	Solo
			11	Otra situación
***** Pregunta 19 *****				**En cada uno de los siguiente enunciados selecciona la opción que mejor describe la forma de actuar de tus padres respecto a tu actividad escolar**
as_esc01	Carácter	1		**Me ayudan a mis tareas escolares**
			1	Casi nunca
			2	Pocas veces
			3	Muchas veces
			4	Casi siempre
as_esc02	Carácter	1		**Me exigen mucho en el estudio de mis materias**
			1	Casi nunca
			2	Pocas veces
			3	Muchas veces
			4	Casi siempre
as_esc03	Carácter	1		**Me felicitan o premian cuando me va bien en la escuela**
			1	Casi nunca
			2	Pocas veces
			3	Muchas veces
			4	Casi siempre
as_esc04	Carácter	1		**Me regañan o castigan cuando me va mal en la escuela**
			1	Casi nunca
			2	Pocas veces
			3	Muchas veces
			4	Casi siempre
as_esc05	Carácter	1		**Me apoyan cuando tengo algún problema en la escuela**
			1	Casi nunca
			2	Pocas veces
			3	Muchas veces
			4	Casi siempre

Pregunta	Tipo	Longitud	Valor	Descripción
as_esc06	Carácter	1		**Respetan mis opiniones sobre lo que ocurre en la escuela**
			1	Casi nunca
			2	Pocas veces
			3	Muchas veces
			4	Casi siempre
as_esc07	Carácter	1		**Promueven que tome mis propias decisiones sobre lo que pasa en la escuela**
			1	Casi nunca
			2	Pocas veces
			3	Muchas veces
			4	Casi siempre
esc_mad	Carácter	2		**Indica el último nivel de estudios que concluyeron completamente tus padres**
				Madre
			01	No sabe leer y escribir
			02	Sabe leer y escribir (sin concluir la primaria)
			03	Primaria
			04	Capacitación técnica (posterior a la primaria)
			05	Secundaria
			06	Capacitación técnica (posterior a la secundaria)
			07	Profesional técnico
			08	Bachillerato o preparatoria o vocacional
			09	Normal (no licenciatura)
			10	Licenciatura
			11	Posgrado
esc_pad	Carácter	2		**Indica el último nivel de estudios que concluyeron completamente tus padres**
				Padre
			01	No sabe leer y escribir
			02	Sabe leer y escribir (sin concluir la primaria)
			03	Primaria
			04	Capacitación técnica (posterior a la primaria)
			05	Secundaria
			06	Capacitación técnica (posterior a la secundaria)
			07	Profesional técnico
			08	Bachillerato o preparatoria o vocacional
			09	Normal (no licenciatura)
			10	Licenciatura
			11	Posgrado

Pregunta	Tipo	Longitud	Valor	Descripción
ocu_mad	Carácter	2		**Indica la ocupación de tus padres**
				Madre
			01	No trabaja actualmente
			02	Jubilado o pensionado
			03	Labores del hogar
			04	Labores relacionadas con el campo o la pesca
			05	Labores relacionadas con la construcción
			06	Obrero
			07	Comerciante o vendedor
			08	Trabajador en servicios personales
			09	Trabajador en oficios o por su cuenta
			10	Directivo o funcionario
			11	Empleado en el ámbito profesional
			12	Empleado en el ámbito técnico o administrativo
			13	Ejercicio de la profesión por cuenta propia
			14	Otra ocupación
			15	No lo sé
ocu_pad	Carácter	2		**Indica la ocupación de tus padres**
				Padre
			01	No trabaja actualmente
			02	Jubilado o pensionado
			03	Labores del hogar
			04	Labores relacionadas con el campo o la pesca
			05	Labores relacionadas con la construcción
			06	Obrero
			07	Comerciante o vendedor
			08	Trabajador en servicios personales
			09	Trabajador en oficios o por su cuenta
			10	Directivo o funcionario
			11	Empleado en el ámbito profesional
			12	Empleado en el ámbito técnico o administrativo
			13	Ejercicio de la profesión por cuenta propia
			14	Otra ocupación
			15	No lo sé
num_foc	Carácter	1		**¿Cuántos focos tiene tu casa, incluyendo lámparas?**
			1	De 1 a 5
			2	De 6 a 10
			3	De 11 a 15
			4	De 16 a 40
			5	Más de 40

Pregunta	Tipo	Longitud	Valor	Descripción
********** Pregunta 23 ********				**¿Con qué frecuencia consumes los siguientes alimentos?**
fre_ali1	Carácter	1		**Carne de res, cerdo, pollo o pescado**
			1	Casi nunca
			2	Una vez a la quincena
			3	Una vez a la semana
			4	entre 2 y 3 veces a la semana
			5	entre 4 y 6 veces a la semana
			6	Diario
fre_ali2	Carácter	1		**Huevos**
			1	Casi nunca
			2	Una vez a la quincena
			3	Una vez a la semana
			4	entre 2 y 3 veces a la semana
			5	entre 4 y 6 veces a la semana
			6	Diario
fre_ali3	Carácter	1		**Leche y derivados**
			1	Casi nunca
			2	Una vez a la quincena
			3	Una vez a la semana
			4	entre 2 y 3 veces a la semana
			5	entre 4 y 6 veces a la semana
			6	Diario
fre_ali4	Carácter	1		**Fruta y verduras frescas**
			1	Casi nunca
			2	Una vez a la quincena
			3	Una vez a la semana
			4	entre 2 y 3 veces a la semana
			5	entre 4 y 6 veces a la semana
			6	Diario
fre_ali5	Carácter	1		**Frijol, arroz, lentejas, habas, etc.**
			1	Casi nunca
			2	Una vez a la quincena
			3	Una vez a la semana
			4	entre 2 y 3 veces a la semana
			5	entre 4 y 6 veces a la semana
			6	Diario

Pregunta	Tipo	Longitud	Valor	Descripción
fre_ali6	Carácter	1		**Pan y pastas**
			1	Casi nunca
			2	Una vez a la quincena
			3	Una vez a la semana
			4	entre 2 y 3 veces a la semana
			5	entre 4 y 6 veces a la semana
			6	Diario
num_pers	Carácter	2		**Número de personas que viven en tu casa**
			01	1
			02	2
			03	3
			04	4
			05	5
			06	6
			07	7
			08	8
			09	9
			10	Más de 9
trabaja	Carácter	1		**Actualmente, ¿desarrollas algún trabajo por el cual recibes un sueldo?**
			1	Si
			2	No
ingr_fam	Carácter	2		**¿Cuál es el ingreso familiar mensual?**
			01	Menos de $ 1000
			02	De $ 1001 a $ 2000
			03	De $ 2001 a $ 3000
			04	De $ 3001 a $ 4000
			05	De $ 4001 a $ 5000
			06	De $ 5001 a $ 6000
			07	De $ 6001 a $ 7000
			08	De $ 7001 a $ 8000
			09	De $ 8001 a $ 9000
			10	De $ 9001 a $ 10000
			11	De $ 10001 a $ 12500
			12	De $ 12501 a $ 15000
			13	De $ 15001 a $ 17500
			14	De $ 17501 a $ 20000
			15	$ 20001 ó más

Pregunta	Tipo	Longitud	Valor	Descripción
prepara	Carácter	1		**¿Qué tan bién preparado te sientes para presentar con buen éxito tu examen de ingreso al nivel medio superior?**
			1	Muy bien
			2	Bien
			3	Mal
			4	Muy mal
for_llen	Carácter	1		**¿Cómo llenaste esta hoja?**
			1	Solo
			2	Con ayuda de mis padres
			3	Con ayuda del orientador
lug_llen	Carácter	1		**¿Dónde llenaste esta hoja?**
			1	En la escuela
			2	En mi casa
cal_orie	Carácter	1		**Califica el apoyo que has recibido por parte de tu orientador en este concurso**
			1	Muy bueno
			2	Bueno
			3	Malo
			4	Muy malo