APLICACIÓN DEL DESCUBRIMIENTO DE CONOCIMIENTO EN EL ANÁLISIS DE MORTALIDAD ACADEMICA EN LAS CIENCIAS BASICAS

OSWALDO RODRIGUEZ DIAZ

EDGAR FERNANDO VEGA MARIN

JORGE ERNESTO PEÑALOZA AFANADOR

MAESTRIA EN CIENCIAS COMPUTACIONALES

UNIVERSIDAD AUTONOMA DE BUCARAMANGA

INSTITUTO TECNOLOGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY

CORPORACION UNIVERSITARIA AUTONOMA DE OCCIDENTE

CALI, 2003

APLICACIÓN DEL DESCUBRIMIENTO DE CONOCIMIENTO EN EL ANÁLISIS DE MORTALIDAD ACADEMICA EN LAS CIENCIAS BASICAS

OSWALDO RODRIGUEZ DIAZ

EDGAR FERNANDO VEGA MARIN

JORGE ERNESTO PEÑALOZA AFANADOR

MAESTRIA EN CIENCIAS COMPUTACIONALES

Tesis

Director
FERNANDO MACHUCA
Doctor en Informática

UNIVERSIDAD AUTONOMA DE BUCARAMANGA

INSTITUTO TECNOLOGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY

CORPORACION UNIVERSITARIA AUTONOMA DE OCCIDENTE

CALI, 2003

CONTENIDO

pág.

LISTA DE FIGURAS

pág.

1. ANTECEDENTES

Uno de los grandes retos de las instituciones educativas a cualquier nivel es el de retener a sus estudiantes, sin perjuicio de la calidad académica. En las instituciones de educación superior las estadísticas muestran que los períodos académicos donde se presenta mayor índice de alumnos que reprueban asignaturas corresponden a los primeros semestres. Sobre este tema se han realizado varios estudios que entregan estadísticas de mortalidad académica y sobre los cuales se plantean posibles causas de la misma, tales como la deficiencia en conceptos básicos de educación media y que son necesarios para un buen desempeño en la educación superior, deficiencia en métodos de estudio, situación personal del alumno, diferencia entre los modelos pedagógicos de la educación media y la educación superior o falta de orientación profesional, entre otros.

Entre las soluciones sugeridas e implementadas por las instituciones de educación superior con relación a la mortalidad académica se encuentran la revisión

continúa de contenidos académicos, la implementación de períodos académicos preuniversitarios de formación básica y las ayudas de orientación profesional a aspirantes y estudiantes.

Las herramientas informáticas disponibles para almacenar, recuperar y analizar información evolucionan a un ritmo sorprendente. La tecnología actual en análisis de información ha posibilitado que las bases de datos donde se guarda información transaccional sobre cualquier tipo de actividad, incluyendo la educación, sean transformadas en bodegas de datos sobre las cuales se puedan realizar tareas de minería de datos tendientes a lograr el descubrimiento de patrones ocultos que marquen el comportamiento de dichas actividades.

Las instituciones de educación superior guardan registros computarizados del comportamiento académico de sus estudiantes, al igual que información básica sobre los mismos, incluyendo datos demográficos, familiares y de procedencia académica. Con esta información, la cual no es posible analizar directamente, se podría plantear un modelo de descubrimiento de conocimiento tendiente a encontrar patrones que describan las características propias de los estudiantes que reprueban asignaturas.

2. PLANTEAMIENTO DEL PROBLEMA Y JUSTIFICACION

La revolución del conocimiento en las últimas décadas del siglo XX se expresa no sólo en su caracterización como mercado sino, también, en la forma como los discursos especializados han sufrido un cambio en su naturaleza, orientación y organización. En la sociedad actual se presentan diferentes situaciones en torno a la Educación Superior; por un lado, se observa la creciente demanda del servicio de educación superior en una población que a pesar de la crisis económica y política del país, necesita formación profesional para conseguir empleo o mejorar sus condiciones laborales donde tienen que conjugar el estudio con el trabajo. Los espacios pedagógicos reales están cediendo terreno frente la virtualidad espacial, relativamente simple, creada por la complejidad del sistema de comunicaciones. En este escenario deben potenciarse un pensamiento y una relación creativa entre la ciencia, la cultura y la tecnología, esto es, una relación de interdependencia que permita analizar los obstáculos anteriores y aprovechar las posibilidades que surgen de sus influencias recíprocas.

Desde este punto de vista, las transformaciones se deben asumir como un pretexto para la reflexión sobre las relaciones estructurales entre la ciencia, la tecnología y la sociedad, las artes y las humanidades, sobre sus mutuas influencias, sobre sus usos en el progreso social y cultural, así como sobre los efectos de éstos.

La práctica social denominada "enseñanza y aprendizaje" y la relación particular que de ellas se deriva (maestro - alumno, enmarcados y sobredeterminados por un medio social específico) configuran un campo científico de teorización y experimentación que tiene como meta central la asimilación de la cultura universal y la nacional en particular con miras a impactar positivamente el desarrollo de los educandos y a transformar dinámicamente el medio social y natural.

Las nuevas prácticas pedagógicas presuponen la generación de diversas competencias para el desarrollo de novedosas formas de interacción y comunicación en la actividad docente con el fin de generar procesos educativos más abiertos al cambio y a la creación, por parte de todos los actores de la educación, de soluciones no convencionales a problemáticas de igual condición.

Durante las dos últimas décadas, la mayoría de las investigaciones en didáctica de las Ciencias han centrado su atención en el Aprendizaje ya que se ha cuestionado el paradigma de enseñanza – aprendizaje de las Ciencias basado en la transmisión verbal del conocimiento científico acabado surgiendo un nuevo paradigma con orientación constructivista (Gil 1983, Driver 1988, Novak 1988)[1]. Por otro lado asumir todos estos cambios, dentro de la práctica docente, hoy en día no son muy significativos además de no incorporar los avances de la sicología cognitiva y de las ciencias de la educación respecto al aprendizaje (Tiberghien 1985, Gutiérrez 1987, Aliberas et al. 1989), en particular desarrollos enfocados hacia el ámbito ideal de la enseñanza de las Ciencias

Por otra parte la matemática misma es una ciencia intensamente dinámica y cambiante y en los últimos treinta años se han generados cambios muy profundos en la enseñanza de las matemáticas donde la comunidad internacional de expertos en didáctica siguen realizando esfuerzos por encontrar el molde adecuado, por lo tanto está claro que se vive aún actualmente una situación de experimentación y cambio.

[1] Tendencias actuales en la formación del profesorado de Ciencias, Furió Mas. Departamento de Didáctica de las Ciencias Experimentales y Sociales. Universidad de Valencia.

Una de las tendencias generales más difundidas hoy consiste en el hincapié en la transmisión de los procesos de pensamientos propios de la matemática más bien que en la mera transferencia de contenidos. La matemática es, sobre todo, saber hacer, es una ciencia en la que el método claramente predomina sobre el contenido.[2]

Con la aparición de herramientas tan poderosas como la calculadora y el computador ha comenzado a influir fuertemente en la orientación de la educación matemática, convirtiéndose en un reto importante en este momento implicando nuevas forma de enseñanza y reformas.

Para Miguel de Guzmán, en nuestro ambiente contemporáneo, con una fuerte tendencia hacia la deshumanización de la ciencia, a la despersonalización producida por la cultura computarizada, es cada vez más necesario un saber humanizado en el que el hombre y la máquina ocupen cada uno el lugar que le corresponde. La educación matemática adecuada puede contribuir eficazmente en esta importante tarea.

Los alumnos hoy en día, se encuentran frecuentemente bombardeados por diferentes técnicas computacionales y comunicativas muy poderosas y

[2] Enseñanza de las Ciencias y la Matemáticas. Miguel de Guzmán. Universidad de Zaragoza

atrayentes, que son importantes reconocer y además son una fuerte competencia con la que se debe enfrentar la enseñanza, por lo tanto se debe aprovechar a fondo tales herramientas como el vídeo, la televisión, la radio, el periódico, el Internet, la multimedia, etc.

La matemática en el siglo XIX y XX que ha predominado es la matemática del continuo en la que el análisis, por su potencia y repercusión en las aplicaciones técnicas, ha jugado un papel predominante. Con la aparición de las herramientas computacionales y comunicativas antes mencionadas, con su inmensa capacidad de cálculo, rapidez, versatilidad, su poder gráfico, se han abierto las posibilidades a múltiples campos diversos que en la educación matemática permite hacer economías operativas o instrumentales en las que se le dedicaba mucho tiempo y no permitía fortalecer los conceptos matemáticos propios.

En un mundo discontinuo, la innovación estratégica, es clave para la creación de riqueza. La estrategia innovadora no es un proceso totalmente ordenado y previsible ni totalmente aleatorio. La sociedad a través de sus formas de vida y sus instituciones genera ambientes y procesos que implican el uso de nuevas tecnologías para contactar nuevas fuentes de conocimiento (V Conferencia Iberoamericana de Educación, 1995). Esto quiere decir que se busca fortalecer la

formación científica en el estudiante, aunque necesariamente esto afecta a todas las demás personas implicadas en los desarrollos curriculares, a través de medios en sí mismos novedosos y, dependiendo de su uso, facilitadores de estrategias pedagógicas creativas.

Existen hoy en día diversas preocupaciones importantes al nivel de la sociedad colombiana: En primer lugar, el reconocimiento de la necesidad de la educación como elemento fundamental para poder abordar el siglo XXI, caracterizado este último por las sociedades del conocimiento, de la información, y el nuevo contexto de la "aldea global". Esto significa que el elemento esencial de transformación o de construcción de sociedad debe estar centrado en el conocimiento, y por ende en la educación. En segundo lugar, el reconocimiento de la educación como factor de convivencia, paz, tolerancia y participación ciudadana. En tercer lugar, como elemento para enfrentar los nuevos retos de la educación para la sociedad del conocimiento, lo que implica que el sistema educativo debe responder a una doble exigencia: por una parte, lograr de la escuela que esta sea efectivamente universal y educadora, y por otra parte, prepararse para la inserción en la "aldea global", sobre la base de insumos como la información y el desarrollo del talento creador. En síntesis, se hace necesario trascender el falso dilema tradicional de calidad versus cobertura, hacia la nueva visión de educación universal de calidad.

Se puede hablar de dos funciones básicas de la educación: las funciones sociales y las funciones individuales. Dentro de las primeras, es decir las funciones sociales, se pueden destacar las siguientes: Facilitar la integración nacional, el crecimiento económico, y la superación de la pobreza. Dentro de las funciones individuales de la educación: la socialización, la transmisión de cultura y el desarrollo de la personalidad, la formación para el trabajo y la formación para la ciencia y la tecnología.

Desde las anteriores ópticas puede reafirmarse la educación como una necesidad para el cambio, y la necesidad de una educación en los códigos de la modernidad. En la primera óptica, se refuerzan los conceptos de formación de alta inteligencia; el *aprender a aprender*; el resolver problemas; la autonomía y la libertad; la educación para la empleabilidad y no para el empleo; y la educación permanente. En la segunda óptica, la de los códigos de la modernidad, está implicada las destrezas, los saberes, las actitudes y los valores.

Este proyecto surge a partir de las sugerencias que desde diferentes ópticas se están generando en el país. Inicialmente se puede aludir a los informes presentados en el "Simposio Internacional. Ciencia, Innovación y Desarrollo Nacional. Bucaramanga. 1998", al presentado por COLCIENCIAS "Haciendo de

Colombia una sociedad del conocimiento. Conocimiento, innovación y construcción de sociedad: *Una agenda para la Colombia del siglo XXI*" o el ICFES en "La Educación Superior a Distancia en Colombia. Visión histórica y lineamientos para su gestión". De otro lado, también se tomó como base a nivel general los planteamientos presentados en "Enseñanza de las Ciencias y la Matemáticas" de Miguel de Guzmán de la Universidad de Zaragoza (España) y a las reflexiones realizadas por la División de Ciencias Básicas en torno a la enseñanza de las matemáticas de la institución que permitió visualizar un panorama de necesidades que, reunidas con las presentadas en los informes arriba mencionados, se puede plantear de la siguiente manera:

1. Establecer propuestas innovadoras, en las que participen **grupos interdisciplinarios**, con el fin de garantizar una mejor visión de temas propuestos y así ofrecer múltiples soluciones a las necesidades establecidas. Establecer una sinergia creativa entre los diferentes profesionales para que transformen e impulsen la competencia empresarial, con deseos cooperativos, en beneficio de la innovación tecnológica.

2. Competir fuertemente y al mismo tiempo aprender de las nuevas tecnologías utilizadas en la producción, estableciendo buenas relaciones sociales e institucionales con prácticas informales de comunicación y colaboración.

A todo lo anterior se incluye otra problemática que es la Mortalidad Académica, presentada por los estudiantes de los primeros años, en las diferentes carreras o programas de las universidades colombianas.

En diferentes estudios realizados por universidades de la región se han mostrado las deficiencias con que llegan los bachilleres a la universidad fruto de las diferentes reformas académicas, en algunos casos implementadas de otros países con un nivel cultural diferente al nuestro, lo cual ha sido uno de los tantos factores o elementos de tal situación. Uno de estos estudios fue el realizado por la Escuela Regional de Matemáticas (ERM) donde participaron las universidades Del Valle, Cauca, Quindío, Tecnológica de Pereira y la Corporación Universitaria Autónoma de Occidente en el año 1989[3] donde se mostró que los estudiantes de primer semestre tenían falencias graves para abordar los primeros cursos universitarios.

Fruto de este análisis, algunas universidades hicieron reformas académicas para tratar de resolver el problema, este fue el caso de la Universidad del Valle que mediante una prueba diagnóstica, clasificó a los estudiantes que podían cursar Cálculo y Álgebra Lineal (Asignaturas de segundo semestre) y los que debían

cursar Matemáticas Fundamental y Geometría Vectorial (Asignaturas de primer semestre). También se ha cambiado el contenido del primer curso de matemáticas, que en algunas universidades se llama Matemáticas 1 o Matemáticas fundamentales o precálculo, primer curso de matemáticas o las matemáticas de empalme entre el bachillerato y la universidad en las diferentes carreras, donde se debe identificas las deficiencias que traen los estudiantes del bachillerato para potenciar los diferentes tópicos básicos que hay que explotar y así evitar su posible fracaso académico y que puedan abordar con mejor soporte los cursos de cálculo lo cual permitiría disminuir la mortalidad académica en los primeros semestres. En otras universidades se ha implementado el semestre cero donde se fortalece al estudiante en álgebra, trigonometría, geometría y aritmética, tópicos que en los estudios mencionados mostraron niveles altos de deficiencia. En otro estudio se identificó que los estudiantes de bachillerato tenían problemas graves de lecto-escritura, por eso en algunas universidades se han implementado programas al respecto para superar dicho problema, este es el caso de la Corporación Universitaria Autónoma de Occidente que tiene el Programa de Orientación Académica (PROA) donde con asignaturas como Métodos y Hábitos de Estudios buscar resolver en parte este problema.

[3] Este estudio fue publicado en los Documentos de Trabajo del Seminario LEMA en Univalle. Septiembre 1990

Al interior de la Corporación Universitaria Autónoma de Occidente, la División de Ciencias Básicas ha identificado el problema mediante una prueba diagnóstica que ha realizado en los dos últimos años donde ha identificado las deficiencias planteadas anteriormente y con las estadísticas crudas del porcentaje de estudiantes que no aprueban las asignaturas, implementó una propuesta metodológica donde se potenciará más el trabajo del estudiante en el aula de clase; esto implicó un cambio en el sistema de evaluación donde no solo se considerarían los tres exámenes parciales como nota de una asignatura sino también el trabajo hecho en el aula de clase[4].

En la gran mayoría de las universidades se habla del problema de deserción o mortalidad académica, pero es poco lo que se sabe acerca de sus orígenes, que son de múltiple naturaleza. Además es una obligación de las entidades educativas, especialmente las universitarias, establecer mecanismos académicos y administrativos para que sus estudiantes puedan superar las dificultades de los programas académicos para que culmine con éxito su carrera profesional escogida.

La deserción es un problema educativo y social ya que una vez que el estudiante esta en un nivel educativo universitario, socialmente se ha realizado un gran

[4] Ver la propuesta en el Documento interno

esfuerzo económico, por lo tanto el fracaso es una catástrofe. A nivel personal, la deserción produce desarraigo, soledad, ausencia de ritos, carencias de rutinas y pérdidas de la capacidad de negociación con otros, soledad social (Richards: 1997).

Para Gabriel Jaime Páramo y otros, la deserción es por excelencia, un problema del sistema educativo, íntimamente ligado a los entornos, contornos y dintornos del mismo, tales como los ambientes educativos, situaciones familiares, exigencias ambientales y culturales que afectan directamente al desertor.

El sistema de acreditación universitaria en Colombia, contempla en la característica 13, los niveles máximos de deserción universitaria y el tiempo promedio de permanencia de los estudiantes en la universidad (CNA; Ibíd.), lo cual muestra que las autoridades de la educación colombiana también considera estos elementos como parte de la autoevaluación con fines de acreditación.

Se puede notar entonces que el análisis de los problemas de mortalidad académica y deserción estudiantil desde la óptica de las ciencias computacionales puede presentar resultados interesantes que ayuden a las instituciones de educación superior a plantear estrategias de mejoramiento.

Para poder desarrollar este proceso se requiere entonces identificar una metodología de descubrimiento de conocimiento, al igual que las herramientas computacionales apropiadas para la extracción y manipulación de información, minería de datos y por último identificación de patrones inherentes a la problemática en cuestión.

A la fecha las acciones que toman las instituciones de educación superior con respecto a la problemática de la mortalidad académica se basan casi enteramente en apreciaciones subjetivas, tomadas por las impresiones que transmiten, en primer lugar, los docentes, por su contacto primario con los estudiantes; en segundo lugar, el staff directivo académico; y como apoyo fundamental el de herramientas de cálculo meramente estadístico.

El uso de KDD en este contexto permitirá obtener parámetros de caracterización de los estudiantes que presentan problemas de mortalidad académica, lo cual posibilitará a la institución de educación superior el tener información con mas detalle y robustez sobre estos estudiantes, lo que permitirá afinar los instrumentos de apoyo al alumnado.

El impacto de la aplicación de esta propuesta se verá reflejado en las estadísticas comparativas de mortalidad académica tomadas en períodos en los

que se involucren actividades de mejoramiento académico como resultado de las evaluaciones realizadas a los parámetros de caracterización de la población estudiantil encontrados.

3. OBJETIVOS

3.1. OBJETIVO GENERAL

Estudiar el impacto que el Descubrimiento de Conocimiento en Bases de Datos aporta al análisis del problema de Mortalidad Académica en la enseñanza de las Ciencias Básicas.

3.2. OBJETIVOS ESPECIFICOS

- Estudiar metodologías de implementación de proyectos de descubrimiento de conocimiento en bases de datos.

- Desarrollar un modelo prototipo de factores que afectan la mortalidad académica en las ciencias básicas.

- Aplicar el modelo de factores propuesto utilizando la metodología seleccionada a la base de datos de información académica de la Corporación Universitaria Autónoma de Occidente teniendo como foco las asignaturas de la División de Ciencias Básicas.

- Analizar los resultados encontrados.

4. MARCO TEÓRICO

4.1. BASES DE DATOS

El uso de los computadores en las empresas y organizaciones en general para el control y automatización de procesos permitió la creación de grandes depósitos electrónicos de almacenamiento de información generalmente llamados bases de datos. En estas bases de datos se encuentran los registros de las transacciones relevantes ocurridas en cada organización en ciertos períodos de tiempo.

En cierta forma, se puede decir que una base de datos es un modelo del mundo real, o de aquella parte del mundo real que es de interés para una organización. En la base de datos se encuentra, organizado de alguna forma, o quizás oculto, el conocimiento que una empresa tiene sobre su negocio.

Asociado al término base de datos, se encuentra el de DBMS (Sistema Administrador de Bases de Datos, por las siglas en inglés de *Database Management System*). Un DBMS consiste en una colección de datos interrelacionados (base de datos) y un conjunto de programas para administrar y acceder esos datos. Un DBMS debe garantizar una serie de requerimientos de administración de los datos de una organización:

- Independencia física y lógica de los datos
- Reserva y seguridad
- Integridad
- Respaldo y recuperación
- Consistencia de los datos
- Capacidad de auditoría
- Control de concurrencia o simultaneidad
- Capacidad de búsqueda
- Desempeño
- Cumplimiento de estándares[5]

El modelo de datos que están almacenados en una base de datos sigue algún enfoque determinado. Aunque han existido varios de ellos, en la década de los

[5] RODRIGUEZ, Miguel A. Bases de Datos. Madrid: McGrawHill, 1992, p. 65-75.

ochenta surgió un el enfoque relacional de los datos, el cual se convirtió en un estándar para el diseño de las bases de datos. El enfoque relacional está basado en una concepción del mundo real, según el cual, éste está compuesto de entidades (abstractas o concretas), las cuales se pueden agrupar, dependiendo del cumplimiento o no de ciertas condiciones; estas entidades se pueden relacionar o asociar con otras entidades que pueden ser o no del mismo tipo. Una entidad se describe por medio de sus atributos, los cuales son una característica única e indivisible de cada entidad que pertenece a un conjunto de entidades. La asociación existente entre las entidades de dos conjuntos de entidades puede ser de cuatro tipos: de una a una, de una a muchas, de muchas a una, o de muchas a muchas, identificando el número máximo de entidades de un conjunto que puede estar relacionada con una entidad del otro conjunto.

El enfoque relacional es en cierta forma simple y efectivo. Se probó exitosamente en el desarrollo de un sinnúmero de aplicaciones orientadas al manejo de datos relativamente estables sobre procesos más o menos fijos.

Con el advenimiento de nuevas tecnologías y el aumento en las capacidades de cómputo, así como también en la cultura informática de usuarios y organizaciones, surgieron requerimientos para nuevas aplicaciones que exigen modelar el mundo real de una forma más compleja que las simples entidades y

asociaciones del enfoque relacional. Como ejemplo de ellas están las aplicaciones que requieren el uso intensivo de tipos de datos "no tradicionales" (videos, sonidos, gráficas e imágenes) tales como enciclopedias, sistemas de información geográficos, y otros sistemas específicos en áreas como educación, publicidad, diseño gráfico, etc.

El enfoque de la orientación a objetos no solo se aplica al diseño de bases de datos. Es un cambio completo de paradigma en cuanto al análisis, diseño e implementación de sistemas de información. Ya no se analiza un sistema con base en los procesos y su descomposición, sino en términos de objetos y su comportamiento. El mundo se modela como un conjunto de objetos que tienen propiedades y comportamiento, y eventos que activan operaciones, las cuales modifican el estado de los objetos. Los objetos interactúan de manera formal con otros objetos. Un objeto puede estar compuesto a su vez, de otros objetos, y así sucesivamente, en forma similar al mundo real, es posible representar objetos complejos, con base en objetos más sencillos[6].

Una base de datos orientada a objetos almacena objetos, es decir, que los datos se almacenan junto con los métodos que los procesan. Surgieron inicialmente

[6] MARTIN, James y ODELL, James. Análisis y Diseño Orientado a Objetos. México: Prentice Hall Hispanoamericana, 1994, p.17-31.

para soportar la persistencia en la programación orientada a objetos. Posteriormente, se volvieron más importantes para aplicaciones con datos complejos como CAD (diseño asistido por computador), por ejemplo. Igualmente, se volvieron importantes para el manejo de BLOBs (objetos binarios de gran tamaño, *Binary Large Objects*), como imágenes, sonidos, videos y texto sin formato[7].

Se podría pensar que el modelo relacional es un caso particular dentro del modelo orientado a objetos, y efectivamente, algunas situaciones específicas serán mejor modeladas por herramientas del tipo relacional, sin embargo, la creciente necesidad de aplicaciones más complejas, que representen más fielmente la realidad, hace que las bases de datos orientadas a objetos sean la alternativa a escoger; como prueba de ello está el hecho de que las versiones más recientes de los DBMS relacionales más fuertes, como el caso de Oracle o DB2, están incorporando características del modelo orientado a objetos, como una forma de manejar el proceso de transición de su gran base instalada de clientes.

4.2. DESCUBRIMIENTO DE CONOCIMIENTO EN BASES DE DATOS (KDD)

[7] Ibid, p. 218-219.

Un análisis sobre los datos almacenados en una base de datos, desde diversas ópticas y con diferentes métodos puede identificar ciertas características o patrones en esos datos, que necesariamente, deben ser reflejo de lo que ocurre en el mundo real que esos datos están representando, y ese nuevo conocimiento puede hacer que las organizaciones reorienten en cierta forma sus procesos y actividades.

Se puede afirmar también, que la cantidad de datos que es administrado por las organizaciones sobrepasa ampliamente la capacidad de análisis de las mismas, por lo tanto, se requiere de herramientas automatizadas de análisis que realicen ese trabajo.

El principal problema para obtener conocimiento de los datos radica fundamentalmente en los mismos datos. Generalmente, solo se recolectan datos relevantes a determinados procesos de la organización, así que el conjunto de datos que se tiene está limitado tanto al tamaño y segmento de la población que representa, como a las necesidades específicas de ciertas áreas. Lo anterior hace que en el caso de tratar de buscar conocimiento sobre esos datos, se necesite hacer ciertos tipos de extrapolaciones para tratar que ese conjunto de datos represente lo mas fielmente posible el universo de interés. También se

puede dar el caso de que variables esenciales para el proceso pueden estar perdidas, por lo que es posible que un proceso de búsqueda de conocimiento genere resultados no exactamente acordes con la realidad[8].

Según Fayyad[9], el descubrimiento de conocimiento en bases de datos (KDD, por las siglas en inglés de *Knowledge Discovery in Databases*) se define como el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y fundamentalmente entendibles en los datos.

Los términos utilizados en esta definición se revisan a continuación:

- **Datos**: Son un conjunto de hechos. Esos datos están almacenados en una base de datos y representan el conocimiento que se tiene del universo donde se enmarca el problema.

- **Patrones**: Es una expresión en algún lenguaje que describe la realidad en un subconjunto del conjunto de hechos o datos mencionado anteriormente. Estos patrones se pueden presentar como la enumeración de los elementos de

[8] WIEDERHOLD, Gio. On the Barriers of Knowledge Discovery. En: Knowlege Discovery in Databases. AAAI/MIT Press, 1991.
[9] FAYYAD, Usama et al. From Data Mining to Knowledge Discovery: An Overview. En: Knowledge Discovery in Databases. AAAI/MIT Press, 1991, p. 6-8.

dicho subconjunto, o como una expresión que represente estos elementos en función de lo ya conocido.

- **Proceso**: En KDD se refiere a un proceso de varias etapas que involucra la preparación de los datos, la búsqueda de patrones, la evaluación del conocimiento obtenido y su refinamiento, y que puede ser en cierta forma iterativo. Este proceso se califica como no trivial, ya que tiene varios grados de autonomía en cuanto a la investigación que realiza.

- **Patrones válidos**: Los patrones descubiertos deben ser probados con algunos niveles de certeza en nuevos datos. Esto implica, ya dentro del proceso de KDD, la separación de los datos conocidos en dos grupos igualmente representativos: uno para realizar los procesos de minería de datos, y otro para validar los resultados obtenidos.

- **Patrones novedosos**: Los patrones descubiertos deben corresponder a algo no conocido anteriormente, por lo menos para la organización.

- **Patrones potencialmente útiles**: Los patrones descubiertos deberían llevar a la toma de acciones útiles para la organización.

- **Patrones fundamentalmente entendibles**: Una meta de KDD es hacer que los patrones descubiertos puedan ser entendibles por los humanos, para así lograr un mejor entendimiento de los datos de los cuales fueron obtenidos[10].

También se puede definir una función para determinar el grado de interés en una medición general de los patrones obtenidos como una combinación de validez, utilidad, simplicidad y novedad. Esta función puede ser tan arbitraria como lo requieran las necesidades de la organización, pero puede ser útil para determinar los costos del proceso. Se puede decir que un patrón encontrado es conocimiento si el valor de su función de medición de interés es mayor que un valor de umbral requerido por el usuario[11].

El proceso de descubrimiento de conocimiento es interactivo e iterativo, involucrando varios pasos con muchas decisiones realizadas por los usuarios del proceso. Los pasos básicos que se realizan son:

- Se debe desarrollar un entendimiento del dominio de la aplicación, del conocimiento previo relevante y de los objetivos y requerimientos del usuario final.

[10] Ibid.
[11] Ibid.

- Posteriormente, se crea un conjunto de datos objetivo, seleccionando un conjunto de datos o enfocándose en un subconjunto de variables o muestra de datos, en los cuales se realizará el proceso de descubrimiento.

- Limpieza y preprocesamiento de los datos: operaciones básicas como la remoción de ruido, decisión de las estrategias para manejar campos de datos perdidos, contabilización de información de secuencias de tiempo y cambios conocidos.

- Reducción y proyección de los datos: encontrar rasgos útiles para representar los datos dependiendo de las metas de la tarea. Utilizar reducción de la dimensionalidad o métodos de transformación para reducir el número efectivo de variables a considerar, o encontrar representaciones invariantes para los datos.

- Selección de las tareas de minería de datos, dependiendo de sí la meta del proceso de KDD es la clasificación, regresión, agrupamiento, etc.

- Selección de los algoritmos de minería de datos a utilizar para la búsqueda de patrones en los datos. Esto incluye la decisión de cuáles modelos y

parámetros serán apropiados con base en los criterios generales del proceso de KDD.

- Minería de datos: búsqueda de patrones de interés en los conjuntos de datos seleccionados para la tarea de acuerdo con alguna forma de representación seleccionada: reglas, árboles, regresión, agrupamiento, etc.

- Interpretación de los patrones obtenidos; posible retorno a cualquiera de los pasos anterior para iterar el proceso.

- Consolidación del conocimiento descubierto: incorporación de este conocimiento a los sistemas de rendimiento, o simplemente, documentarlo y reportarlo a las partes interesadas. Esto incluye también el chequeo y resolución de posibles conflictos con el conocimiento previamente aceptado[12].

[12] Ibid, p. 9-11.

4.3. MINERÍA DE DATOS

Es un paso en el proceso de KDD consistente en algoritmos particulares de búsqueda en los datos que, bajo ciertas limitaciones aceptables de eficiencia computacional, producen una enumeración particular de patrones sobre los datos[13]. La minería de datos involucra modelos para ajustar o patrones a determinar sobre los datos observados.

Las dos metas primarias de la minería de datos son la **predicción** y la **descripción**. La predicción involucra el uso de algunas variables o campos de la base de datos para predecir los valores futuros de otras variables de interés. La descripción se enfoca en encontrar patrones interpretables por humanos para la descripción de los datos. En términos de KDD suele ser más importante la descripción que la predicción. Para obtener estas metas se realizan las siguientes tareas primarias:

- Clasificación:
- Regresión
- Agrupamiento (*clustering*)

[13] Ibid, p. 9

- Sumarización
- Modelamiento de dependencias
- Detección de cambios y desviaciones[14]

Los algoritmos de minería de datos tienen tres componentes primarios:

- Representación del modelo: específicamente, trata de la identificación del lenguaje de representación para la descripción de los patrones descubribles.

- Evaluación del modelo: estima que tan bien un patrón particular satisface los criterios del proceso de KDD.

- Método de búsqueda, compuesto por búsqueda de parámetro y búsqueda de modelo. En la búsqueda de parámetros, el algoritmo debe buscar los parámetros que optimizan los criterios de evaluación del modelo, dados los datos observados y una representación fija del modelo. La búsqueda del modelo ocurre un ciclo por encima de la búsqueda de parámetros, es decir, se realiza el proceso de búsqueda de parámetros para toda una familia de modelos[15].

---

[14] Ibid, p. 12-16.
[15] Ibid, p. 16-17.

Los métodos más utilizados para el proceso de minería de datos son:

- Reglas y árboles de decisión
- Métodos de regresión no lineal y clasificación
- Métodos basados en ejemplos
- Modelos de dependencia gráfica probabilística
- Modelos de aprendizaje relacional[16]

Es común encontrar en ciertos autores y publicaciones la interpretación equivalente de los términos KDD y minería de datos. Como se ha dicho anteriormente, minería de datos es un paso dentro del proceso de KDD. El componente de minería de datos del proceso de KDD está relacionado con los medios por los cuales los patrones son extraídos y enumerados de los datos. El descubrimiento de conocimiento involucra la evaluación y posible interpretación de los patrones para determinar cuáles constituyen conocimiento y cuáles no;

[16] Ibid, p. 17-22.

adicionalmente, involucra también la selección de esquemas de codificación, preprocesamiento, muestreo y proyección de los datos en forma previa a la etapa de minería de datos.

4.4 DESERCION ESTUDIANTIL Y MORTALIDAD ACADEMICA

Aunque a primera vista se puede tomar la deserción académica como la referencia a aquellos estudiantes que por cualquier razón no continúan sus estudios en una institución y la mortalidad académica como la referencia a los estudiantes que reprueban asignaturas durante un período académico dado, se han encontrado diferentes posiciones al respecto que de alguna manera reflejan la falta de consenso sobre estos tópicos. A continuación se presenta una serie de definiciones para las expresiones *deserción estudiantil* y *mortalidad académica* extraídas de diferentes textos de referencia.

Por *deserción estudiantil* los autores consultados conceptúan:

- Fenómeno consistente en iniciar y no terminar un programa académico a nivel de la enseñanza, sea cual fuera la causa de la no finalización.

- Diferencia entre la matricula inicial y la final en un mismo año considerado

- Conjunto de individuos que abandonan las actividades de un tipo de educación, la temporalidad del abandono puede ser transitoria o definitiva.

- Abandono que el alumno hace de las actividades escolares, comprendiendo estas, como momentos pedagógicos en el desarrollo del acto escolar, durante cualquier período del año lectivo y dentro de los diferentes programas curriculares y cuya causa sea generada al interior o exterior de la institución

- Abandono del aula por razones ajena a las académicas

- El abandono definitivo de las aulas de clase por diferentes razones, la no-continuidad en la formación académica, que la sociedad requiere y desea en y para cada persona que inicia sus estudios de primaria, esperanzados en que termine felizmente los estudios universitarios

Por *mortalidad académica* los autores consultados conceptúan:

- Abandono del aula por razones estrictamente de índole académica (Arboleda y Picón, 1997).

- Batista y otros (1994) identifican algunos factores que determinan la mortalidad:
  - Respecto al alumno: bases inadecuadas de formación, irresponsabilidad, baja motivación, desinformación, métodos inadecuados de estudios, trabajos y estudios simultáneos.

  - Respecto al profesor: inadecuada formación ética y profesional, deficientes mecanismos de comunicación, posición intransigente.

  - Para Gabriel Jaime Páramo y Carlos Arturo Correa[17], el estudio de la deserción se ha enfocado desde diversos ángulos o perceptivas, según sean los intereses y necesidades de quienes lo emprenden. Muchos autores no formulan una clara definición de este concepto, y más aún, la confunden conceptualmente con otros fenómenos inherentes al sistema educativo, como incluirla dentro de los parámetros de mortalidad estudiantil, ausentismo y retiro forzoso.

[17] Profesores de la Escuela de Ingeniería. Universidad EAFIT

Los fenómenos de la mortalidad y la deserción estudiantil están presentes en cualquier sistema educativo, independientes del nivel de desarrollo o de las características cualitativas que hubiere alcanzado el estudiante (Arboleda y Picón: 1977). Puede añadirse como fenómenos simultáneos a la deserción estudiantil en el sistema educativo, el ausentismo a clases y el retiro forzoso.

Hay varias clases de deserción en educación:

- Deserción total: Abandono definitivo de la formación académica individual
- Deserción discriminada por causas: Según la causa de deserción
- Deserción por facultad (escuela o departamento): Cambio facultad - facultad
- Deserción por programa: cambio de programa en una misma facultad
- Deserción a primer semestre de carrera: por inadecuada adaptación a la vida universitaria.
- Deserción acumulada: sumatoria de deserciones en una institución

Entre las características que identifican a los estudiantes que desertan de una institución de educación superior se encuentran:

- Bajo aprovechamiento de oportunidades educativas
- Problemas de disciplina
- Hijos de padres que no les interesa la educación

- Problemas con la justicia
- Adolecen de motivación e interés para realizar su labor educativa
- Nivel socio - económico bajo o sin opción económica
- Ausentismo a clases
- Problemas de salud sicosomática
- Problemas inherentes a la edad
- Inadecuadas relaciones interpersonales
- Provienen de ambientes familiares y sociales violentos
- Baja empatía por el trabajo de sus pares
- Resistencia a desarrollar actividades formativas
- Inapetencia por el conocimiento
- Desmotivación hacia la carrera y a la universidad

Entre las variables asociadas a los factores de deserción estudiantil y mortalidad académica están las siguientes:

- Ambientes educativos universitarios en los cuales está inmerso el estudiante
- Ambientes familiares
- Procesos educativos y acompañamiento al estudiante en su formación
- Edad. La mayoría de los estudiantes universitarios son muy jóvenes
- Adaptación social del estudiante desertor con sus pares u homólogos

- Bajos niveles de compresión unidos a la falta de interés y apatía por programas curriculares
- Modelos pedagógicos universitarios diferentes a los modelos de bachillerato, que imprime un alto nivel de exigencia
- Programas micro - curriculares universitarios rígidos con respecto a los de su formación secundaria, de alta intensidad temática, dispuestos en corto tiempo
- Evaluaciones extenuantes y avasalladoras.
- Cursos no asociados ni aplicables con su ejercicio profesional
- Factores económicos que impiden la continuidad del desertor en la Universidad
- Cantidad de oferentes
- Orientación profesional
- Masificación de la educación

5. ESTADO DEL ARTE

5.1 MINERÍA DE DATOS

La minería de datos, consistente en la extracción de información oculta y predecible de grandes bases de datos, es una poderosa tecnología con gran potencial para ayudar a las compañías a concentrarse en la información más importante de sus bases de información.

Estas herramientas exploran las bases de datos en busca de patrones ocultos, encontrando información que un experto humano difícilmente encontraría, estableciendo relaciones y patrones de las cuales las empresas pueden obtener grandes beneficios.

El Data Mining surgió como una integración de múltiples tecnologías tales como la estadística, el soporte a la toma de decisiones, el aprendizaje automático, la

gestión y almacenamiento de bases de datos y el procesamiento en paralelo. Para la realización de estos procesos se aplican técnicas procedentes de muy diversas áreas, como pueden ser los algoritmos genéticos, las redes neuronales, los árboles de decisión, etc.

Aunque los componentes clave del Data Mining existen desde hace décadas en la investigación en áreas como la inteligencia artificial, la estadística o el aprendizaje automático, se puede afirmar que ahora estamos asistiendo al reconocimiento de la madurez de estas técnicas, lo que, junto al espectacular desarrollo de los motores de bases de datos y las herramientas para integración de información justifican su introducción en la esfera empresarial[18].

La idea de *data mining* no es nueva. Ya desde los años sesenta los estadísticos manejaban términos como *data fishing*, *data mining* o *data archaeology* con la idea de encontrar correlaciones sin una hipótesis previa en bases de datos con ruido. A principios de los años ochenta, Rakesh Agrawal, Gio Wiederhold, Robert Blum y Gregory Piatetsky-Shapiro, entre otros, empezaron a consolidar los términos de *data mining* y KDD.[19] A finales de los años ochenta sólo existían un par de empresas dedicadas a esta tecnología; hoy existen más de 100 empresas

[18] Minería Visualización y Descubrimiento de Conocimiento en Bases de Datos. Fernando Martín Sánchez; Nieves Ibarrola de Andrés; Guillermo López Campos. Instituto de Salud Carlos III. España

en el mundo que ofrecen alrededor de 300 soluciones. Las listas de discusión sobre este tema las forman investigadores de más de ochenta países. Esta tecnología ha sido un buen punto de encuentro entre personas pertenecientes al ámbito académico y al de los negocios[20].

Sin embargo, actualmente el proceso de minería de datos es dirigido, es decir que requiere de la interacción con el usuario de forma constante. Sería de gran utilidad e incrementaría la eficiencia del sistema si éste poseyera cierta autonomía para tomar decisiones, modificar ciertos parámetros, y redirigir las búsquedas de forma que los resultados obtenidos sean útiles.

Ciertamente el proceso no puede ser totalmente automatizado ya que las computadoras o los algoritmos en sí carecen de la experiencia e intuición humana para reconocer la diferencia entre un patrón relevante y uno que no lo es, en este punto la interacción con el especialista es fundamental; este definirá qué patrones de comportamiento son los que se van a buscar y en el momento en que se presente la información será quien de acuerdo a su criterio la descarte o haga uso de ella.

---

[19] http://www.kdnuggets.com/

[20] Data Mining: Torturando los datos hasta que confiesen. Luis Carlos Molina Félix. UOC

Durante el proceso de minería de datos el usuario plantea ciertas preguntas que indican qué patrón de comportamiento de los datos dentro de la base de datos es el que desea encontrar y además de esto, se deben fijar ciertas restricciones para dirigir la búsqueda. El problema reside en que si la pregunta no fue formulada de una forma adecuada los resultados devueltos por el minero pueden no ser los deseados o en el peor de los casos el resultado puede ser nulo. En ambos casos se ha perdido tiempo, se han procesado datos en vano y la pregunta debe de volver a ser planteada por el usuario, hecho que puede volver a llevarnos nuevamente a obtener los mismos resultados inútiles.

Una vez que el minero entrega los resultados deseados, estos no se pueden considerar útiles hasta que el usuario pueda interpretarlos y hacer uso de ellos; por ello es importante que estos sean presentados de forma clara y entendible para el usuario[21].

[21] Construcción del módulo de Minería. Carlos Emilio Castillo Hernández. Centro de Investigación en Computación de México.

5.2 MINERÍA DE DATOS EN EL ANÁLISIS DE MORTALIDAD ACADÉMICA

En las Universidades se ha encontrado una cantidad considerable de estudiantes que desertan tanto de la institución como de la carrera que han seleccionado, pero no hay estudios relativos que muestre las verdaderas causas, solo hay, en algunas, estadísticas de deserción y mortalidad.

A nivel nacional la Universidad de los Andes, al igual que otras universidades del país, ha estudiado el fenómeno de deserción y la cataloga en dos categorías: académica y no académica, y ha publicado estudios sobre el número de estudiantes que salen de la universidad por bajo rendimiento académico, sin analizar las causas del mismo, para el periodo 1990 a 1996 los resultados fueron altos en cada cohorte que va desde el 6.2% (máximo) al 2.53% (mínimo) (Uniandes: 1998)

En la Universidad EAFIT[22] analizaron este fenómeno para el periodo 1995 - 1998, con el argumento que la eficiencia del sistema educativo se mide a través de su capacidad para conservar o retener a los estudiantes en la institución y consideran que deserción es un indicador de situaciones de crisis en el ámbito

educativo. Sus hipótesis fueron: la deserción académica o voluntaria es una manifestación tanto de procesos de cambio al interior de la universidad, como también de situaciones de crisis en el ámbito psicológico y socioeconómico de los estudiantes, los estudiantes con menor rendimiento académico pueden considerarse como potenciales desertores, la deserción académica y no académica son inversamente proporcionales al desarrollo de los planes de estudios, y el grado de movilidad interna está asociado con variables tales como: prestigio, grado de profesionalización y afinidad temática de las carreras. Otro estudio realizado en esta Universidad fue el de Dr. Guillermo Vélez (1986) cuyo objetivo era "Tipificar al estudiante de EAFIT afectado por la deserción forzosa en los tres primeros niveles de los programas académicos de formación universitaria entre 1984 - 1985", su enfoque fue el seguimiento del bajo rendimiento académico como causa de deserción. También Sarmiento y Giraldo (1989) estudiaron los motivos que originaron la deserción forzosa, voluntaria o por transferencia y encontraron las siguientes causas: la no identificación con la carrera, problemas familiares (separación de padres), profesores con poca pedagogía, malas técnicas de estudio, costo de matricula, proceso de selección utilizado por la universidad.

[22] Documento de "Deserción Estudiantil en los programas de pre grados" por Rocío Osorio y Catalina M. Jaramillo. Julio 1999

Otros estudios coinciden en que la deserción es causada por factores internos y externos referentes al profesor, al alumno y a la institución. Salazar y Castillo (1968) realizaron un estudio en la Universidad Nacional sobre los retiros del programa de Sociología el cual dejo entrever que a medida que se avanzaba en la carrera la deserción disminuía y consideraron variables independientes como: nivel socioeconómico, edad al ingresar, sexo, interés motivacional al ingresar y el sistema de selección. También determinaron que la deserción hace subir el costo anual por alumno.

Vélez y Ramírez (1974) analizaron el fenómeno de deserción como la principal causa del bajo rendimiento académico en los programas académicos en la universidad en general. Las causas más sobresalientes las dividieron en dos tipos:

- **Problemas externos a la universidad**. El sistema educativo no permite un paso integral y armónico entre los distintos niveles de enseñanza. Condiciones socioeconómicas que obligan a los estudiantes a trabajar.

- **Problemas internos a la universidad.** Falta de planificación y programación; sobrepoblación escolar y deficiencias docente; falta de ayuda eficaz a los estudiantes.

Los autores concluyeron que la deserción impide un crecimiento armónico de la universidad, de sus planes y programas, y en consecuencia, delimita el logro de los objetivos mismo del ser universitario. Recomendaron:

- Levantar un censo universitario que permita la creación de bancos de datos con fines de planeación y programación general de la universidad.
- Investigar y conocer índices de mortalidad y deserción estudiantil en cada universidad.

- La deserción estudiantil debe prevenirse y curase desde la escuela secundaria.

Para Collazos y Gensini (1973) la deserción es un problema que afecta la eficiencia de todo el sistema educativo, medida en término de cobertura con indicadores como el rendimiento interno y externo.

Según Graciarena (1970), habría dos tipos principales de deserción:

- Deserción Académica o por fracaso en los estudios. La decisión de desertar sería en este caso una consecuencia natural del funcionamiento de los mecanismos de selección de un sistema universitario.

- Deserción por desmoralización. Cuando el estudiante llega poco motivado y con el tiempo se desmotiva más por el estudio, puede ser causada por: los estudios preuniversitarios, influencia familiar, otros intereses, desajustes vocacionales, etc.

Otro estudio realizado por Batista (1994) tuvo como objetivo general "establecer en series históricas, por cohortes, la mortalidad y la deserción de los estudiantes admitidos a la universidad desde 1980". Estableció las diferentes tasas de deserción y mortalidad académica, y los factores sociodemográficos asociados a ellos (sexo, edad, procedencia geográfica, tipo de colegio, estrato socio - económico)

A nivel internacional el problema también es preocupante. En México por ejemplo, entre los problemas que se presentan hoy al respecto de la mortalidad académica y la deserción estudiantil están: la escasa eficiencia terminal de los estudiantes universitarios, la falta de una metodología adecuada para el aprendizaje, el estancamiento de la oferta de lugares en la universidad y la dificultad para que los egresados se incorporen al medio laboral (Bojalil: 1998). La deserción estudiantil en México es grave, de 100 niños que ingresan a la primaria, 11 alcanza bachillerato y tan sólo 4 terminan una carrera. En la

Benemérita Universidad de Puebla, realizaron un proyecto para "Mejorar los índices de aprobación y deserción en Matemáticas"[23] analizaron las posibles causas respecto al alumno y al profesor y su justificación básica era que si bien los índices de reprobación en las materias de matemáticas no son el único factor que determina la deserción de los estudiantes, afecta de manera significativa.
En Puerto Rico, al igual que en otros países, se han hechos grandes esfuerzos para explicar el mismo fenómeno y no han precisado la magnitud, debido a la ausencia de métodos sistemáticos en la recopilación de datos. (Revista Educación: 1993).

En Argentina, en la Universidad Nacional de Cuyo, han realizado una investigación al respecto con el objeto de conocer las causas de deserción en esta universidad, encara el estudio por los métodos cuantitativos y cualitativos y tienen una metodología especial para realizar entrevistas a estudiantes que abandonaron sus estudios (CONICMEN: 1998).

En la Universidad Católica de Chile, se ha analizado el cambio de carrera como una deserción interna, evaluando la aplicación del "sistema curricular flexible" implementado por la institución en 1967. Se emplearon índices de recepción, participación y restricción de las carreras. La conclusión fue que el régimen

[23] Proyecto a cargo de la Dra. Leticia Gómez Esparza y otros. Facultad de

curricular flexible contribuyó a producir en la universidad, una situación de deserción interna debido a que muchos estudiantes usaban ciertas carreras para ingresar a la universidad y luego se cambiaban a otra de su interés.

No sólo las autoridades educativas se han preocupado por su sector e indirectamente por el fenómeno de deserción, también lo hacen las autoridades económicas del país. Es el caso del Banco de la República, en cuanto la política de reducción de la inflación que debe evitar distorsiones en los precios relativos que restrinjan el acceso a la educación de los jóvenes colombianos, e incide como factor de deserción. (Urrutia, 1995, nota editorial).

Por otro lado, Panambi Abadie en "Estudio sobre indicadores y costos en la Educación Superior" muestra que varias universidades del mundo usan diferentes indicadores para hacer su evaluación institucional. Dentro de los indicadores utilizados a partir de los ochenta se observó que muchos consideraban la tasa de deserción o mortalidad académica, este es el caso del el Informe Jarrat (Reino Unido. 1985) que incluyó indicadores de rendimiento interno, de rendimiento externo y de rendimiento operativo; y dentro de los indicadores de rendimiento interno consideraron la tasa de graduación y tipo de titulaciones, tasa de éxito de las titulaciones superiores. En los indicadores de enseñanza que se utiliza en

---

Ciencias Electrónicas. 2000

Gran Bretaña en la actualidad, se incluye la tasa de deserción (Ruiz, 1999 extraído de Kells, 1997).

Cave y otros (1998) construyeron un conjunto más pequeño de indicadores de rendimiento donde en enseñanza aparece abandonos y tasa de deserción.

En los indicadores de rendimiento en las universidades argentinas, según Pérez Lindo (1990), en la parte de rendimiento académico tienen considerado como uno de los indicadores, la tasa de deserción y lo consideran como un elemento en la evaluación institucional.

Otro elemento a considerar es el costo en la educación superior, donde en Estados Unidos y el Reino Unido, consideran dentro de este análisis, el costo por estudiante que consiste en sumar todos los gastos institucionales y dividirlos entre el número de estudiantes (Lewis, opcit), el costo por hora crédito (student credit hour) que se obtiene al dividir el promedio anual de horas créditos generadas durante el año académico por un número x de estudiantes de pregrado o un número x de estudiantes de post - grado (Syverson, 1997) y el costo por producto. Estos costos ofrecen, sin embargo, el inconveniente de no reflejar acabadamente el costo de la enseñanza dado que no considera elementos que

inciden en el costo final, como la deserción y la repitencia estudiantil (Petrei, 1989) o la duración de los programas de estudio (HEFCE, opcit).

La unidad egresado (o "graduado") es utilizada en los sistemas de financiamiento de Dinamarca, Finlandia y Holanda con el objetivo de reducir los costos de la Educación Superior, incentivando la reducción de la permanencia del estudiante en los programas de estudio (Albrecht, 1993).

En Argentina, desde los ochenta se ha trabajado para determinar los costos universitarios. En 1989, Petrei y otros, calcularon los costos por carreras en 26 universidades y casi en forma simultánea en la Universidad Nacional de Cuyo se elaboró una metodología para el cálculo de los costos en las carreras universitarias (Ginestar, 1990). En dicha metodología de los conceptos a tener en cuenta en la estimación de costos se incluyeron los elementos de repitencia y de deserción estudiantil.

De todo lo anterior se observa que el problema de deserción y mortalidad académica ha sido analizado por muchas universidades de diferentes países del mundo donde se han involucrado variables clásicas como sexo, edad, clase social, etc., pero no se han encontrado patrones de los desertores o los estudiantes que están en mortalidad académica.

De igual manera se puede notar también que la utilización de técnicas de descubrimiento de conocimiento y minería de datos sobre bases de datos de información académicas para el análisis del problema de mortalidad académica es un campo completamente inexplorado, el cual sólo ha sido analizado desde el punto de vista estadístico.

5.3 METODOLOGÍAS PARA LA APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS

A través de la investigación de metodologías estructuradas para los procesos de Minería de Datos, se puede notar que no existe un planteamiento específico para cada caso de análisis. Al respecto se propone la siguiente metodología siempre adaptable a la situación de negocio particular a la que se aplique:

- **Definición del Problema y Formulación.** En el punto inicial del proceso de solución del problema, el analista identifica el problema a ser resuelto y/o define los objetivos del sistema. En general, el analista debe conocer que es

lo que desea que los datos procesados revelen, y las variables relevantes posibles que el modelo recibirá como entrada.

- **Colección de Datos.** En el segundo paso el analista identifica los datos que son necesarios para cumplir la meta especificada. Los datos deberían representar el alcance del programa o área de interés. El modelo es solo válido para el rango de datos históricos.

- **Preparación de los Datos.** Un punto crítico en el proceso de minería es la preparación de los datos. Del 50% a 90% del tiempo del proceso de minería es consumido en la preparación de los datos. La mayoría de los datos son almacenados en un formato diferente que el requerido para ser usado por el modelo, estos deben ser limpiados para eliminar los datos incorrectos, incompletos (campos vacíos) e inconsistentes.

- **Exploración de los Datos.** Esto involucra la exploración de los datos preparados para tener un panorama general antes del descubrimiento de patrones y también para validar los resultados de la preparación de los datos. Típicamente esto involucra examinar las estadísticas (mínimo, máximo, promedio, etc.) y la distribución de la frecuencia de campos individuales.

También involucra el análisis de gráficas de campos contra campos para entender la dependencia entre campos.

- **Modelación.** La modelación es el corazón del proceso de minería de datos. Este paso impone la selección de los algoritmos a ser usados y la parametrización apropiada. Todos los algoritmos de minería de datos tienen parámetros de "afinación". Estos parámetros son usados para controlar cosas, tales como qué tan rápido los clasificadores aprenden o cuál es el soporte y confianza que deben cumplir las reglas de asociación. Dependiendo del problema que está siendo investigado y del tipo de conocimiento que se quiera obtener se debe seleccionar cuales algoritmos se usarán.[24]

5.4 HERRAMIENTAS DE SOFTWARE PARA MINERÍA DE DATOS

Debido al acelerado crecimiento de los negocios electrónicos, las compañías cada vez se llenan más de grandes cantidades de información operacional usualmente

[24] Generación de Reglas de Asociación en Bases de Datos Distribuidas mediante la implementación del algoritmo DMA en la SP2. Sergio Cruz Jiménez. Instituto Politécnico Nacional de México

almacenada en un datawarehouse o simples bases de datos. Las herramientas de Minería de Datos son una forma útil de explorar esa información corporativa, en busca de relaciones y patrones ocultos dentro de los datos de las empresas.

Actualmente existen Dentro de KDD varias herramientas tanto comerciales como académicas para realizar procesos de Minería de Datos, la diferencia entre una y otras radica en sus objetivos, la automatización, cantidad de datos que puede utilizar y aún mas importante, su integración con la base de datos. De estas herramientas se pueden citar algunas como:

- **Cognos Scenario (Cognos).** Scenario es una herramienta que posibilita descubrir tendencias y patrones ocultos y detectar con anterioridad insospechadas correlaciones en la información, sin requerir expertos en técnicas estadísticas. Scenario ayuda a entender mejor los factores que conducen a los negocios. Revela los patrones y relaciones existentes en los datos dándole la comprensión que necesita para tomar informadas y oportunas decisiones. Básicamente, Cognos Scenario permite:
    - Identificar y ranquear los factores que impactan en las metas de la empresa o entidad.
    - Ingresar al detalle para identificar cuáles son los factores que afectan los resultados.

- Crear un perfil de qué es lo que está funcionando y qué no.
- Detectar excepciones en los datos corporativos[25].

- **WEKA (Universidad de Waikato, Nueva Zelanda)**. Es un sistema desarrollado en el lenguaje de programación Java, implementa algoritmos de minería de datos que pueden aplicarse a bases de datos desde una línea de comando a través de su interfaz gráfica. Incluye herramientas para transformar datos en un esquema de aprendizaje, a fin de que los resultados puedan ser analizados y extraer de estos información. Trabaja con algoritmos de reglas de asociación y agrupamiento de datos(*clustering*)[26].

- **MineSet (Silicon Graphics).** MineSet utiliza mecanismos de análisis y generación de reglas de asociación y modelos de clasificación, usados para predicción, clasificación y segmentación. Combina estos modelos con animación y visualización interactiva. MineSet también provee herramientas para realzar y acelerar el análisis de los datos. Utiliza tres modelos de clasificación: Árboles de decisión, árboles de opción y clasificadores de evidencia (Simple-Bayes)[27].

[25] http://www.cognos.com/index.html
[26] http://www.cs.waikato.ac.nz/~ml/weka/
[27] http://techpubs.sgi.com/library/tpl/cgi-bin/search.cgi

- **DBMiner (Universidad de Simon Fraser, Canadá).** Inicialmente llamada DBLearn, se caracteriza por manejar grandes cantidades de datos, f grafica, posee habilidades para realizar asociaciones, clasificaciones y agrupamientos de datos[28].

- **Clementine (SPSS INC.).** Es la herramienta para Data mining de SPSS, que le permite desarrollar y distribuir modelos predictivos. Maneja con facilidad grandes volúmenes de información para encontrar patrones en los datos y permite tomar decisiones confiables a partir de estos. Maneja técnicas de segmentación, scoring y predicción. Esta herramienta posee una internase gráfica amigable, que hace de la Minería de Datos (Data Mining) un proceso interactivo, de alta productividad, con el que obtiene modelos de gran escalabilidad, fácilmente distribuibles a lo largo de la organización[29].

- **Enterprise Miner (SAS).** Solución de minería de datos que permite incorporar patrones inteligentes a los procesos de marketing, tanto operativos como estratégicos[30].

[28] http://www.dbminer.com/
[29] http://www.spss.com/spssbi/clementine/
[30] http://www.sas.com/technologies/analytics/datamining/miner/

- **Intelligent Miner for Data (IBM)**. Esta herramienta, brinda soporte superior de extracción de datos, incluyendo capacidades de procesamiento de datos, análisis estadístico y visualización de resultados. Dentro de Intelligent Miner for Data están sus algoritmos de extracción dirigidos a la toma de decisiones más intuitiva y estratégica. Estos algoritmos pueden utilizarse individualmente o en combinación para tratar una gran variedad de problemas de negocios y, aún más importante, ofrecer resultados medibles. Proporciona patrones ocultos y predice tendencias futuras, permitiendo obtener una ventaja competitiva. Permite extraer datos de bases DB2, archivos planos y otras bases de datos relacionales a través de DataJoiner[31].

La Corporación Universitaria Autónoma de Occidente tiene, desde hace algunos años, un convenio para la utilización del software de la compañía IBM en actividades y proyectos de tipo académico lo que permite tener acceso a diferentes herramientas de ellas entre ellas *Intelligent Miner*, por lo tanto se profundizará un poco más en sus características fundamentales.

Básicamente el trabajo de Intelligent Miner en relación con la base de datos, es de soportar, además de las sofisticadas técnicas mining encapsuladas en una interfaz visual fácil de manejar, las funciones de preparación de los datos para

[31] Data mining Minería de datos. Catálogo de Software

extraer información desde bases de datos Oracle o Sybase y cargarlos en DB2 para mining.

IM soporta seis categorías de minería de datos diferentes: Asociaciones, Clasificación, Agrupamiento, Predicción, Patrones Secuenciales y Análisis de secuencia de tiempo. Cada categoría soporta uno o más algoritmos y provee interfaces de configuración que permiten definir los datos y parámetros a ser analizados. Estas técnicas pueden ser usadas para las siguientes finalidades, entre otras:

- Detección de desviaciones y fraudes.
- Segmentación de clientes.
- Prospectos de mercados.
- Búsqueda textual abstracta.

IBM ha comparado el desempeño de su herramienta de Minería de Datos con rivales tales como la herramienta SPSS de la firma SPSS Inc. y su combinación de herramientas de análisis, soporte de datos e interoperabilidad con otras herramientas de inteligencia, lo posicionan como un gran competidor[32].

Como ya se mencionó anteriormente, además de las bondades y ventajas de la herramienta Intelligent Miner consideradas en el presente documento, la Corporación Universitaria Autónoma de Occidente tiene acceso a los productos de IBM para uso académico, esto junto con las demás virtudes propias del software permiten que sea considerado la mejor alternativa para realizar el proceso de minería de datos de este proyecto.

[32] Intelligent Miner revela joyas corporativas. Maggie Biggs. InfoWorld Test Center

6. DEFINICIÓN Y DESARROLLO DEL PROTOTIPO

6.1 METODOLOGÍA DE KDD APLICADA AL ANÁLISIS DE MORTALIDAD ACADÉMICA EN LAS CIENCIAS BÁSICAS DE LA CUAO:

Para llevar a cabo el procedimiento de descubrimiento de Conocimiento en Bases de Datos aplicado al proyecto actual se tomarán en cuenta los pasos de la metodología planteada por F. J, Cantú[33].

[33] Una Metodología de KDD por Francisco J. Cantú del Centro de Inteligencia Tecnológica del TEC de Monterrey en http://www-cia.mty.itesm.mx/~fcantu/orgint/metoedologia-KDD.PDF

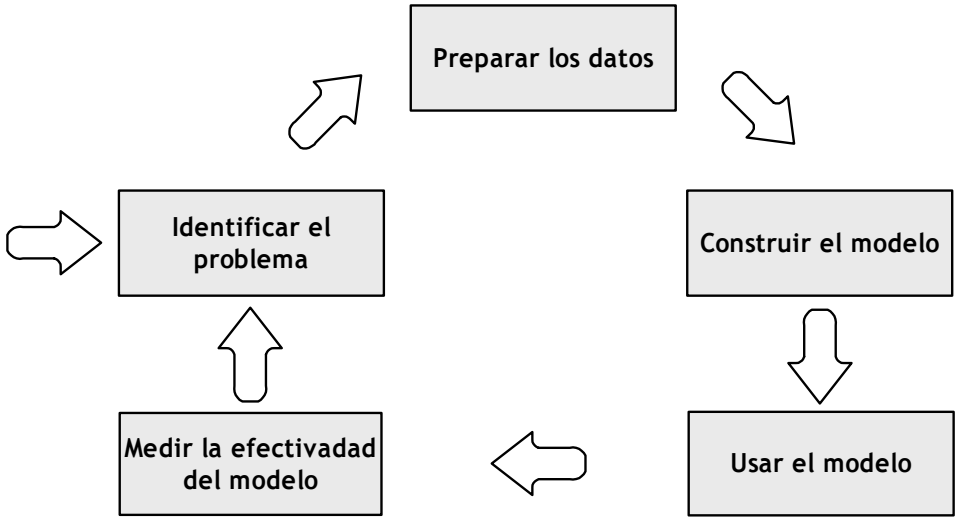


Figura 1. Metodología de KDD

6.1.1. IDENTIFICAR EL PROBLEMA

En la gran mayoría de las universidades se habla del problema de deserción o mortalidad académica, pero es poco lo que se sabe acerca de sus orígenes, que son de múltiple naturaleza. Además es una obligación de las entidades educativas, especialmente las universitarias, establecer mecanismos académicos y administrativos para que sus estudiantes puedan superar las dificultades de los programas académicos para que culmine con éxito su carrera profesional escogida.

La deserción es un problema educativo y social ya que una vez que el estudiante esta en un nivel educativo universitario, socialmente se ha realizado un gran esfuerzo económico, por lo tanto el fracaso es una catástrofe. A nivel personal, la deserción produce desarraigo, soledad, ausencia de ritos, carencias de rutinas y pérdidas de la capacidad de negociación con otros, soledad social (Richards: 1997).

Para Gabriel Jaime Páramo y otros, la deserción es por excelencia, un problema del sistema educativo, íntimamente ligado a los entornos, contornos y dintornos del mismo, tales como los ambientes educativos, situaciones familiares, exigencias ambientales y culturales que afectan directamente al desertor.

El sistema de acreditación universitaria en Colombia, contempla en la característica 13, los niveles máximos de deserción universitaria y el tiempo promedio de permanencia de los estudiantes en la universidad (CNA; Ibíd.), lo cual muestra que las autoridades de la educación colombiana también consideran estos elementos como parte de la autoevaluación con fines de acreditación.

Se puede notar entonces que el análisis de los problemas de mortalidad académica y deserción estudiantil desde la óptica de las ciencias

computacionales puede presentar resultados interesantes que ayuden a las instituciones de educación superior a plantear estrategias de mejoramiento.

El uso de KDD en este contexto permitirá obtener parámetros de caracterización de los estudiantes que presentan problemas de mortalidad académica, lo cual posibilitará a la institución de educación superior el tener información con más detalle y robustez sobre estos estudiantes, lo que permitirá afinar los instrumentos de apoyo al alumnado.

El impacto de la aplicación de ésta propuesta se verá reflejado en las estadísticas comparativas de mortalidad académica tomadas en períodos en los que se involucren actividades de mejoramiento académico como resultado de las evaluaciones realizadas a los parámetros de caracterización de la población estudiantil encontrados.

En diferentes estudios realizados por universidades de la región se han mostrado las deficiencias con que llegan los bachilleres a la universidad fruto de las diferentes reformas académicas, en algunos casos implementadas de otros países con un nivel cultural diferente al nuestro, lo cual ha sido uno de los tantos factores o elementos de tal situación, fruto de este análisis, algunas universidades hicieron reformas académicas para tratar de resolver el problema.

Al interior de la Corporación Universitaria Autónoma de Occidente, la División de Ciencias Básicas ha identificado el problema mediante una prueba diagnóstica que ha realizado en los dos últimos años donde ha identificado las deficiencias planteadas anteriormente

En el caso concreto del presente análisis, se estudiarán los posibles patrones de comportamiento que puedan mostrar los estudiantes de los primeros semestres de las diferentes carreras de la división de Ingeniería, Ciencias Básicas y Ciencias económicas y administrativas de la CUAO, para ello se considerarán los datos con que cuenta la institución:

- Base de datos de historial académico de estudiantes.

- Base de datos de prueba diagnóstica de la División de Ciencias Básicas

6.1.2. PREPARAR LOS DATOS

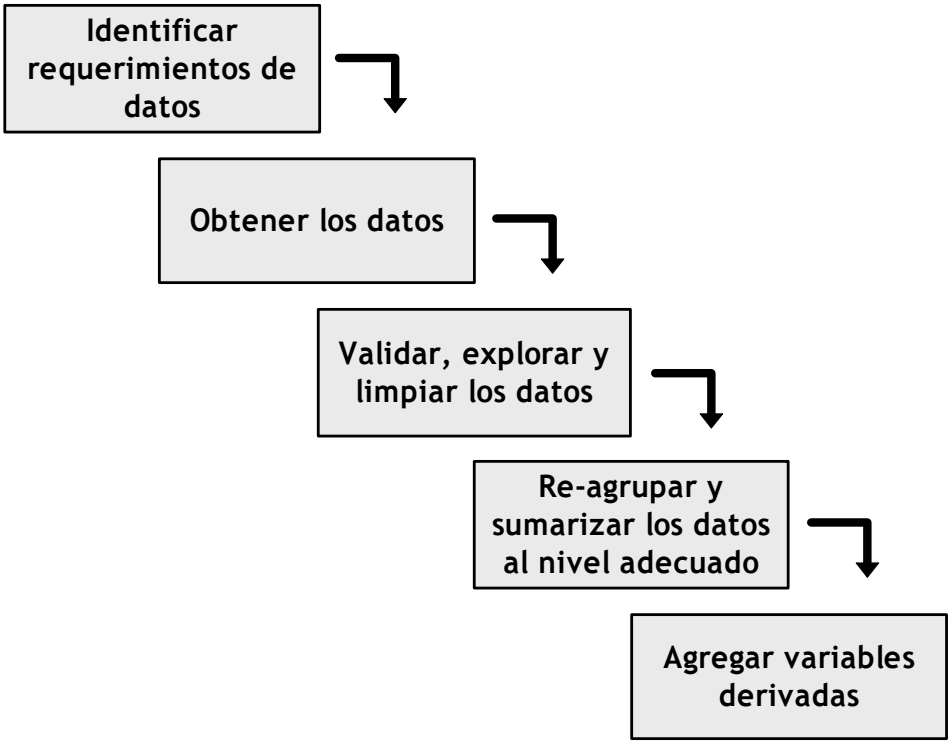


Figura 2. Preparación de datos

- **Identificar requerimientos de datos**: de los datos que se tomarán como modelo de entrada, extraídos de las bases de datos de la División de Ciencias Básicas y del historial académico de estudiantes de la CUAO, se identificarán los siguientes parámetros o variables requeridos para el análisis:

a. Base de datos de historial académico de estudiantes: Código del estudiante, código asignatura, nota matemáticas 1, año y período en que cursó la materia, y nota Icfes en matemática y en lenguaje.

b. Base de datos de prueba diagnóstica de la División de Ciencias Básicas: Número de prueba diagnóstica, código estudiante, código de programa, jornada, nota prueba diagnóstica, edad, sexo, ciudad de residencia, estrato, colegio, año de grado de bachillerato, ciudad de grado, tipo y jornada del colegio, y modalidad de bachillerato.

- **Obtener los datos**: la extracción de los datos requeridos para el modelo de evaluación se realizará de la siguiente manera:

  Los datos de la Base de datos *Oracle* del historial académico de estudiantes, se extraerán mediante consultas SQL hechas a la base de datos, y los datos de las pruebas diagnósticas hechas a los alumnos de los cursos iniciales de matemáticas de la División de Ciencias Básicas se obtendrán de una hoja electrónica en Excel proporcionada por dicha división.

- **Validar, explorar y limpiar los datos**: con el fin de evitar situaciones de incongruencia en los resultados del análisis y teniendo en cuenta que los

datos no fueron recogidos inicialmente para tareas de Data Mining, ésta fase tiene como objetivo garantizar la calidad de los datos y para ello se escogerá la información de los estudiantes en cuyos registros existan la totalidad de los campos relacionados con los parámetros o variables escogidas para el caso de evaluación.

- **Re-agrupar y sumarizar los datos al nivel adecuado**: después de haber seleccionado de la base de datos *Oracle* del historial académico y del archivo plano suministrado por la División de Ciencias Básicas los datos para el análisis de mortalidad académica, se creará con estos la Bodega de datos implementando el manejador de Bases de Datos DB2, dicha bodega permitirá reposar en ella los datos en el nivel adecuado para realizar el proceso de minería de datos.

- **Agregar variables derivadas**: una vez creada la Bodega de Datos en DB2, de acuerdo con las hipótesis o supuestos de análisis se armarán los perfiles de consultas, los cuales permitirán determinar patrones de comportamiento en el análisis de mortalidad académica.

6.1.3. CONSTRUIR EL MODELO

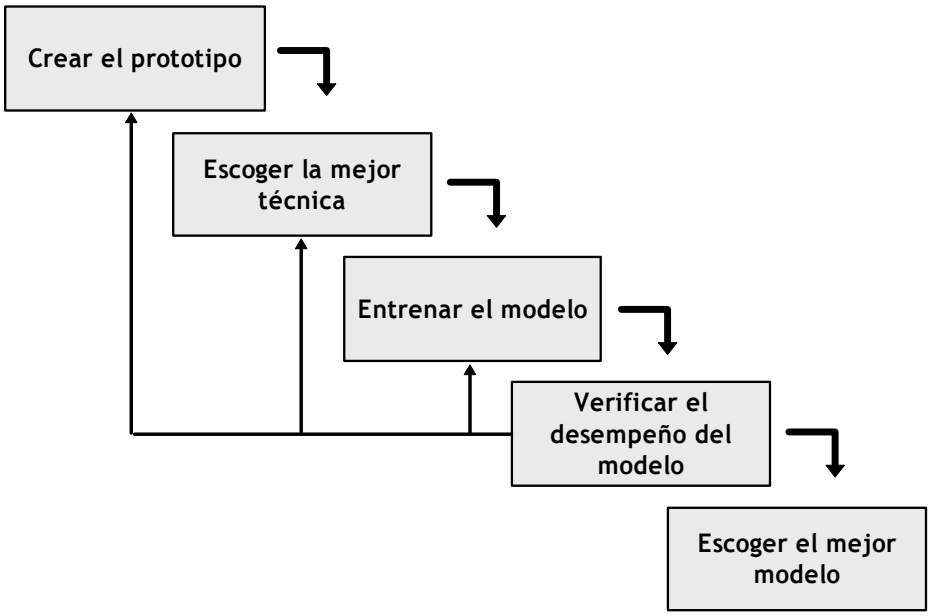


Figura 3. Construcción del modelo.

- **Crear el prototipo**: concluido el proceso de transformación de los datos el cual los modela de manera que los algoritmos de data mining puedan operar con ellos ya será posible acceder a la base de datos de la bodega, haciendo uso de la herramienta Intelligent Miner la cual permitirá hacer las respectivas consultas con base en las hipótesis o supuestos de análisis planteados inicialmente.

- **Escoger la mejor técnica**: para llevar a cabo el proceso de descubrimiento de patrones es necesario hacer uso de las técnicas o categorías de minería de datos. Los patrones obtenidos tienen niveles de efectividad, exactitud y generalidad de acuerdo con la técnica utilizada, y la interpretación de estos es un excelente complemento al conocimiento de los expertos y reduce de manera significativa la incertidumbre en la toma de decisiones. Lo anterior quiere decir que es indispensable analizar dichas técnicas y determinar cual es la que más se adecua a las hipótesis de trabajo.

  La herramienta Intelligent Miner, soporta seis categorías de minería de datos:

  - **Asociación**: Las asociaciones permiten descubrir relaciones entre los registros en un análisis, es decir, en una colección de datos existentes en un conjunto de registros, una función de asociación es una operación sobre el conjunto de registros, la cual regresará las afinidades existentes entre la colección de datos. Las asociaciones pueden involucrar cualquier numero de datos.

  - **Clasificación:** Este método agrupa los datos de acuerdo a similitudes o clases, es decir, dado un conjunto de registros, cada uno comprendido por un numero de atributos, un conjunto de etiquetas (representando clases

de registros) y una asignación de una etiqueta a cada registro, una función de clasificación examina el conjunto de registros etiquetados y produce descripciones de las características de los registros para cada una de las clases. Estas descripciones de clases pueden ser usadas para etiquetar nuevos registros determinando a que clase pertenecen.

- **Agrupación**: Esta técnica permite la identificación de grupos en los cuales los elementos guardan similitud entre sí y se diferencia de los otros grupos. Utiliza una técnica de aprendizaje no supervisado, es decir, no se le proporciona ninguna información al sistema. No se parte de un conjunto prefijado de categorías sino que a través del análisis de los datos y de su naturaleza la técnica agrupa dichos datos en las distintas categorías.

  Una vez realizada la agrupación se podrán realizar estudios sobre ellos mediante técnicas estadísticas, árboles de decisión, redes neuronales, etc.

- **Predicción**: Se enfoca en predecir eventos o comportamientos específicos, basado en información histórica. Permite clasificar la información por factores, estos factores se pueden ajustar a la necesidad del usuario o la empresa.

- **Descubrimiento de patrones secuenciales**: Los patrones secuenciales revelan hábitos en el comportamiento en determinadas actividades en relación a un grupo de actividades a lo largo del tiempo.

- **Descubrimiento de secuencias temporales similares**: Análisis de una secuencia de medidas hechas a intervalos específicos. El tiempo es usualmente la dimensión dominante de los datos. Pueden revelar las variaciones existentes en un periodo de tiempo dado.

Y cada una de estas a su vez soporta uno o más algoritmos formales que ayudan a analizar los datos de acuerdo con los parámetros o supuestos planteados inicialmente.

Dada la naturaleza del proyecto, la técnica más adecuada para realizar el proceso de Minería de Datos es la de Clasificación, ya que dentro del contexto de la búsqueda de las causas de la mortalidad académica en los estudiantes de Ciencias Básicas, precisamente lo que se pretende encontrar es un conjunto de características comunes que sirvan como patrones de identificación temprana de los estudiantes con riesgo de reprobar las asignaturas de Ciencias Básicas. La categoría de clasificación es soportada por la herramienta Intelligent Miner for Data.

La herramienta Intelligent Miner for Data maneja, dentro del método de clasificación, técnicas de minería como árboles de decisión y redes neuronales.

- **Árboles de decisión;** es un análisis que genera un árbol de decisión para predecir el comportamiento de una variable, a partir de una o más variables predictoras, de forma que los conjuntos de una misma rama y un mismo nivel son disjuntos.

  Es útil en aquellas situaciones en las que el objetivo es dividir una población en distintos segmentos basándose en algún criterio de decisión.

- **Redes neuronales**; genéricamente son métodos de proceso numérico en paralelo, en el que las variables interactúan mediante transformaciones lineales o no lineales, hasta obtener unas salidas. Estas salidas se contrastan con los que tenían que haber salido, basándose en unos datos de prueba, dando lugar a un proceso de retroalimentación mediante el cual la red se reconfigura, hasta obtener un modelo adecuado. Las redes neuronales podrán usarse como método del descubrimiento del conocimiento, particularmente útiles para el reconocimiento de patrones.

- **Entrenar el modelo:** una vez escogida la técnica más idónea para llevar a cabo el proceso de minería de datos con Intelligent Miner, el paso siguiente es determinar el conjunto de datos o variables con las que se desea trabajar, los cuales deberán estar basados en las hipótesis planteadas inicialmente, y después de haber definido dicha fuente de datos de minería se cargarán los datos en el modelo.

- **Verificar el desempeño del modelo:** ésta fase corresponde a la verificación de la consistencia de los resultados arrojados por la herramienta de acuerdo con la técnica y el conjunto de datos escogidos para el proceso de minería de datos, esto quiere decir que se probará el prototipo teniendo en cuenta varias de las hipótesis planteadas, con el fin de asegurar que los resultados que se vayan obteniendo en cada proceso sí estén determinando patrones lógicos y acertados.

  Como se puede notar en la figura 3, los cuatro pasos mencionados anteriormente integran un proceso de retroalimentación, lo que indica que el prototipo se deberá probar varias veces, incluso con diferentes conjuntos de datos y técnicas de minería, esto con el fin de determinar la validez de los resultados obtenidos.

- **Escoger el mejor modelo:** de acuerdo con los resultados obtenidos del proceso de minería de datos haciendo uso de las diferentes técnicas y conjuntos de datos, se determinará el modelo más acertado para el proceso de análisis de mortalidad académica en la CUAO.

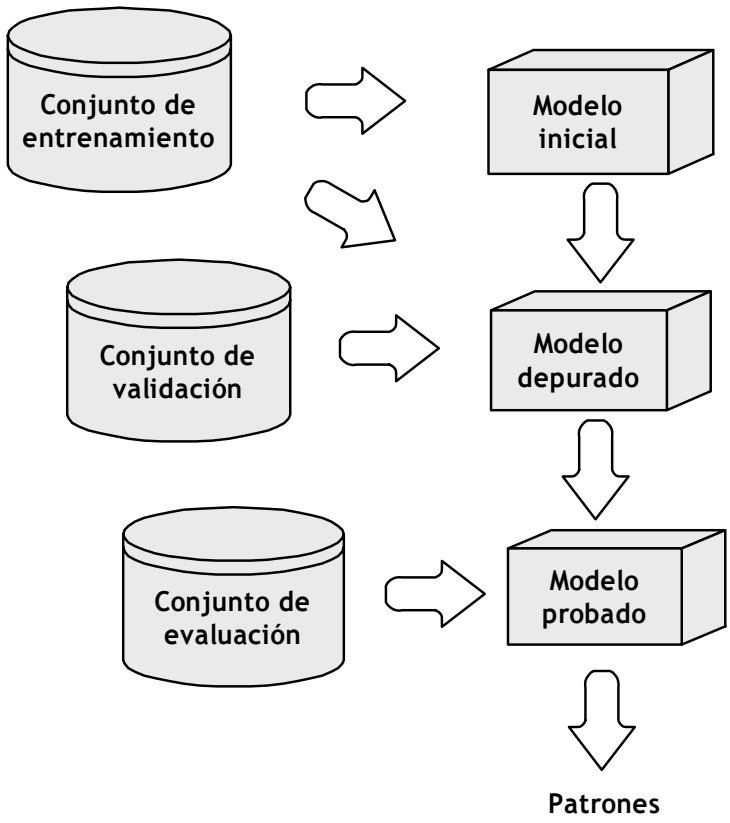


Figura 4. Entradas del modelo

De acuerdo con la figura No. 4 para el caso del análisis de la mortalidad académica en las Ciencias Básicas, el modelo inicial estará representado por

el estudio estadístico realizado en la fase inicial de esta investigación, donde fueron tomados como conjunto de entrenamiento los resultados de las Pruebas diagnósticas que reposan en la base de datos de la División de Ciencias Básicas.

De igual manera para pasar al modelo depurado, (el prototipo) donde se transformarán los datos para correr sobre estos los procesos de minería, se tomarán como entrada tanto el conjunto de entrenamiento como el de validación determinado por los datos de la Base de Datos de Registro Académico de la CUAO, esto permitirá, mediante un conjunto de evaluación representado por la integración de perfiles de consultas basados en los supuestos de evaluación, obtener un modelo probado que mediante análisis y asimilación de sus resultados permita determinar patrones de comportamiento útiles al análisis de mortalidad académica.

6.1.4. USAR EL MODELO

El uso del modelo se planteará con base en los supuestos planteados al inicio del proceso de descubrimiento de conocimiento en las bases de datos, donde se esperan obtener patrones que determinen índices de mortalidad académica, sugeridos por ejemplo por la edad y el sexo de los estudiantes que ingresan a la universidad, el estrato, colegio o ciudad de dónde provengan, la relación entre sus calificaciones obtenidas en diferentes períodos, entre otros.

6.1.5. MEDIR LA EFECTIVIDAD DEL MODELO

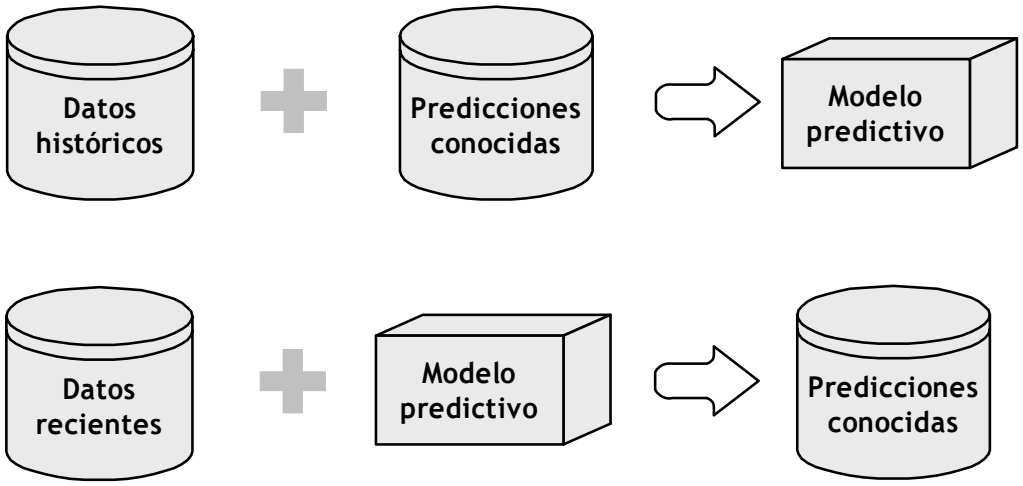


Figura No. 5 Medir la efectividad del modelo

De acuerdo con la figura No. 5, la efectividad del modelo comprende dos ciclos: Ciclo de vida del modelo y ciclo de vida de la predicción, en el primero "Ciclo de vida del modelo" mediante la entrada de datos históricos, que para el caso particular se tomarán los registros de la base de datos correspondientes a los períodos 2001A y 2001B dado que es un rango en el cual los datos se encuentran completos y la integración de predicciones conocidas como son los supuestos de análisis con los que se trabajará se obtendrá un modelo predictivo, el cual ayudará a determinar patrones de comportamiento que caractericen el problema de mortalidad académica en las Ciencias Básicas durante los períodos mencionados, y sugiere un modelo abierto ya que si es cierta la efectividad del modelo se podrá seguir aplicando con diversos conjuntos de datos y supuestos.

En el segundo ciclo: "Ciclo de vida de la predicción", mediante la entrada de datos recientes y junto con el modelo predictivo obtenido del ciclo de vida del modelo, se podrá determinar qué tan válidos y por cuánto tiempo se siguen presentando los patrones o comportamientos hallados en el proceso de descubrimiento de conocimiento en bases de datos en el análisis de la mortalidad académica. Lo anterior permitirá determinar por ejemplo qué tan efectivas han sido las decisiones tomadas con base en los resultados obtenidos a partir del modelo predictivo, y de igual manera comparar resultados en diferentes tiempos.

6.2. FACTORES QUE INCIDEN EN LA PROBLEMÁTICA DE LA MORTALIDAD ACADÉMICA

Con base en la información que posee la División de Ciencias Básicas de la CUAO, determinada por los datos obtenidos mediante el proceso de Prueba Diagnóstica (Ver anexo de formato de Prueba Diagnóstica) que se viene llevando a cabo desde el primer semestre del año 2000, se realizó un estudio Estadístico interno para identificar elementos que determinen las dificultades con que ingresan a la CUAO los estudiantes del primer semestre de los diferentes programas de Ingeniería, Ciencias Básicas y Ciencias Económicas y Administrativas.

Se proporcionaron además de los resultados de la prueba, datos adicionales del estudiante, los cuales fueron solicitados por medio de un formato (Ver anexo formato de solicitud de información – División de Ciencias Básicas) a cada uno de ellos en la División de Ciencias Básicas el día en que presentaron la prueba y los cuales representan información como: colegio de procedencia, sexo, estrato, tipo de colegio, año de graduación, jornada del colegio, edad, y ciudad. Y con el fin de obtener un Modelo Inicial que correlacionara las variables anteriores se le aplicaron técnicas multivariadas a los datos de la prueba diagnóstica (Ver anexo de Informe Estadístico).

Con base en este estudio se generaron las siguientes hipótesis

1. Es probable que las respuestas a las pruebas ICFES y diagnóstica obedezcan más a comportamientos aleatorios a la hora de responder. Se recomienda que en la aplicación de futuras pruebas diagnosticas, se realicen pruebas de aleatoriedad mediante test apareados.

2. La pérdida del curso de Matemáticas 1 o su aprobación, pueden estar ligadas a otras variables no medidas tales como situación económica, ambiente familiar, expectativas no satisfechas, tiempo de desplazamiento entre el sitio de trabajo o el hogar y la universidad, tiempo de dedicación al estudio, relación del contenido con el resto de asignaturas, entre otras.

3. Puede existir una influencia importante en la transición entre la secundaria y la universidad.

4. Las pruebas ICFES, diagnóstica y las evaluaciones del curso de Matemáticas 1, tienen objetivos diferentes y son realizadas en momentos diferentes; lo cual puede conllevar a una ausencia de relación entre ellas.

Por otro lado la universidad cuenta con información particular de los estudiantes como su historial académico consistente en el código de las asignaturas cursadas, la nota de cada asignatura cursada, jornada, programa, periodo, profesor que dictó la asignatura, grupo, etc.

Teniendo en cuenta lo mencionado y partiendo del hecho de que tanto en los datos suministrados por Ciencias Básicas, como en los que reposan en la base de datos de Registro Académico existe información oculta y útil, el presente proyecto pretende demostrar que mediante el uso de técnicas de Minería de Datos es posible extraer dicha información con el fin de determinar parámetros que caractericen la Mortalidad Académica o deserción en la CUAO, y para dicho proceso se considerarán los siguientes supuestos:

El índice de mortalidad académica en las asignaturas de Ciencias Básicas, puede estar ligado a las variables:

a) Sexo: Los estudiantes que más pierden o desertan de la asignatura x, son los hombres / mujeres.

b) Edad: Los que más pierden o desertan son de x años.

c) Estrato: El estrato que más aporta a estos índices es el x.

d) Colegio: Los estudiantes que más incurren en estos índices son de colegios públicos /privados.

e) Año de grado: En estos índices influye el tiempo de graduación del estudiante del bachillerato a su ingreso en la universidad. Tienen más dificultad los estudiantes que se graduaron hace más de x años.

f) Programa: Caen más en el índice de mortalidad académica o deserción los estudiantes de la carrera x / los estudiantes de la carrera y.

g) Jornada de la carrera: Se encuentran mas en estos índices los estudiantes de la jornada x / los estudiantes de la jornada y.

h) Tipo o modalidad del bachillerato: Presentan más problemas de mortalidad académica o deserción los estudiantes de colegios bachillerato: académico/ técnico / normalista.

i) Ciudad. Los estudiantes que más caen en estos índices son de la ciudad x.

j) Jornada del Bachillerato. Influye en estos índices, la jornada del bachillerato cursado por un estudiante.

Vale la pena aclarar que en este proyecto sólo se han considerado uno de los agentes que participa en el proceso Enseñanza – Aprendizaje, el estudiante, pero el prototipo planteado permitirá también hacer un análisis con respecto al docente y a la institución como tal, donde se podrían plantear hipótesis alrededor de estos elementos como: El profesor con el que más pierden o ganan una asignatura *x* los estudiantes, es de sexo *y*, con edad *z*, de planta o de hora cátedra, de w años de experiencia docente, entre otras.

6.3. DESARROLLO DEL PROTOTIPO

6.3.1. DEFINICIÓN DEL CASO DE EVALUACIÓN

El caso de evaluación de herramientas de minería de datos se desarrollará sobre la base de datos de información académica de la *Corporación Universitaria*

Autónoma de Occidente, la cual está montada sobre el manejador de base de datos *Oracle 8i*, y en particular se trabajará sobre el subconjunto que relaciona la información de *Estudiantes*, *Notas* y *Programas Académicos*. Existen datos importantes para el modelo, tales como el resultado de las pruebas diagnósticas que se hacen a los estudiantes de los cursos iniciales de matemáticas, que no se encuentran registrados en la base de datos por lo que se hace necesario incorporar dicha información al modelo utilizando herramientas de adquisición de datos.

6.3.2. DEFINICIÓN DEL SUBCONJUNTO DE DATOS PARA EL MODELO

El modelo de datos académicos de la *Corporación Universitaria Autónoma de Occidente* comprende estructuras de almacenamiento que guardan información sobre:

- Estudiantes
- Asignaturas
- Programas Académicos

- Planes de Estudio
- Prerrequisitos
- Horarios del semestre
- Asignaturas programadas para el semestre
- Calificaciones del semestre
- Histórico de calificaciones

Para efectos del modelo de evaluación se tomarán los datos básicos de las tablas de *Colegios*, *Estudiantes*, *Asignaturas*, *Programas Académicos* y *Calificaciones*, tal como se muestra en la figura número 5.

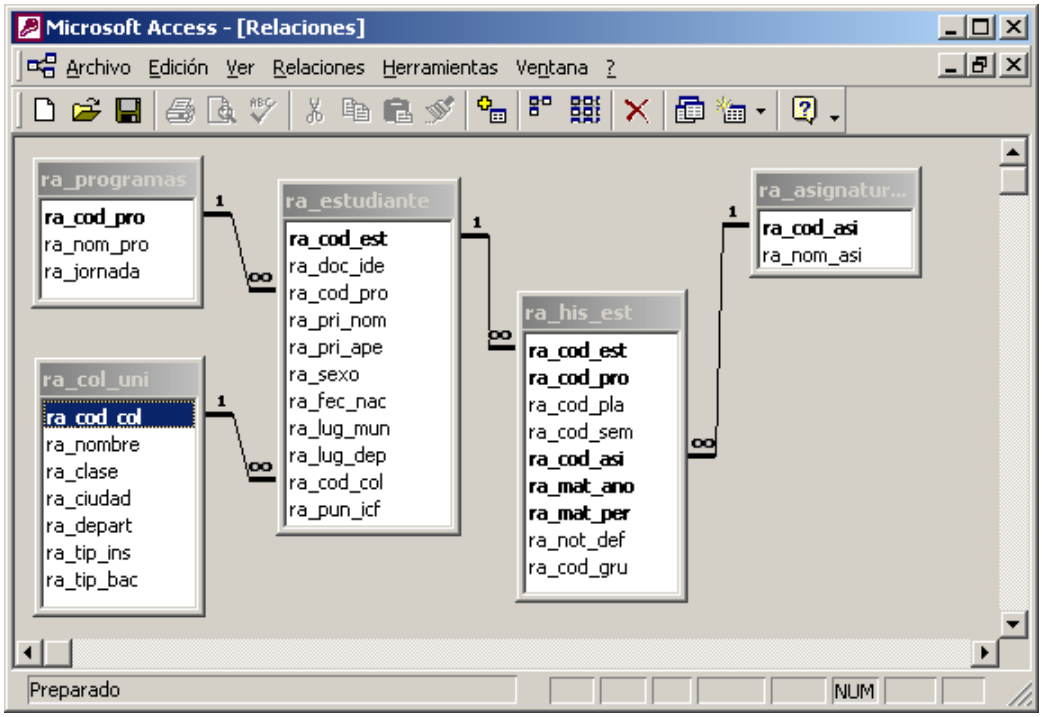


Figura 5. Modelo de datos simplificado para el proceso de evaluación

6.3.3. DEFINICIÓN DE TAREAS DE MINERIA A REALIZAR

El ciclo de vida de un proceso de minería de datos está compuesto por cuatro etapas:

- Identificar el problema de negocio.
- Usar técnicas de minería para la transformación de los datos en información útil.
- Actuar de acuerdo con la información obtenida.
- Medir los resultados[34]

El proceso de evaluación que se desarrollará en este proyecto está centrado en la problemática de la mortalidad académica en las asignaturas de ciencias básicas de la *Corporación Universitaria Autónoma de Occidente*. Para ello se tomará como base del análisis el comportamiento de los estudiantes de los programas académicos de las divisiones de Ingenierías, Ciencias Económicas y Ciencias Básicas que cursaron la asignatura *Matemáticas I* en el año 2001. Los datos disponibles para realizar el análisis son los siguientes:

[34] BERRY, Michael J. A. y LINOFF, Gordon. Data Mining Techniques for Marketing, Sales and Customer Support. Wiley, 1997, p. 17-45.

- Año de graduación
- Asignatura cursada (para Ingenierías o Ciencias Económicas)
- Ciudad de residencia
- Programa Académico
- Edad
- Estrato socioeconómico
- Jornada de estudio en la Universidad
- Jornada de estudio en el Colegio
- Sexo
- Nota Icfes en Matemáticas
- Nota Icfes en Lenguaje
- Tipo de Colegio
- Tipo de Grado de Bachillerato
- Resultado de la prueba diagnóstica aplicada.

Sobre esta base de información se generará un modelo de clasificación por árboles, utilizando la herramienta seleccionada *Intelligent Miner for Data*.

6.3.4. EVALUACIÓN DE PRODUCTOS COMERCIALES DE MINERIA DE DATOS

Para la evaluación de software comercial de minería de datos se utilizará un subconjunto de los datos académicos de la Corporación Universitaria Autónoma de Occidente; estos datos se encuentran alojados en una base de datos relacional *Oracle* versión 8i. El software de minería a evaluar es el *DB2 Intelligent Miner versión 8.1* de la firma *IBM*. Los datos se extraerán de la base de datos académica y de archivos planos complementarios y se alojarán en una base de datos temporal administrada con el manejador de bases de datos relacionales *DB2 Universal Database* también de la firma *IBM*. Esta base de datos temporal así como el software de minería residirán en la misma máquina para efectos de la prueba.

El proceso de evaluación tiene cinco fases definidas: instalación del software, creación de las bases y bodegas de datos, definición del proceso de importación de datos, aplicación del modelo e interpretación de los resultados. Estas fases se detallan a continuación:

6.3.4.1. INSTALACIÓN DEL SOFTWARE REQUERIDO PARA LA EVALUACIÓN

Para la realización de la evaluación se instalaron y configuraron correctamente, de acuerdo con las recomendaciones de los diferentes proveedores, las siguientes herramientas de software:

a. Sistema Operativo *Microsoft Windows 2000 Professional*
b. Manejador de bases de datos *IBM DB2 Universal Database Personal* versión *7.2*
c. Software para minería de datos *IBM Intelligent Miner for Data* versión *8.1*
d. Software de acceso a base datos *Oracle Client* versión *8.1.6*.

6.3.4.2. CREACIÓN DE BASES DE DATOS PARA EL MODELO DE DATOS

Los datos del modelo de evaluación serán extraídos de la información académica almacenada en una base de datos *Oracle,* así como de una hoja electrónica en Excel donde se encuentran los resultados de las pruebas diagnósticas hechas a los alumnos de los cursos iniciales de matemáticas, y serán almacenados en una base de datos *DB2*. Esta base de datos recibirá el nombre de *CUAODATA* y se crea desde el *Centro de Control* de *DB2* que es invocado a su vez por la opción

Programas -> IBM DB2 -> Centro de Control del menú de *Inicio* de *Windows 2000 Professional*.

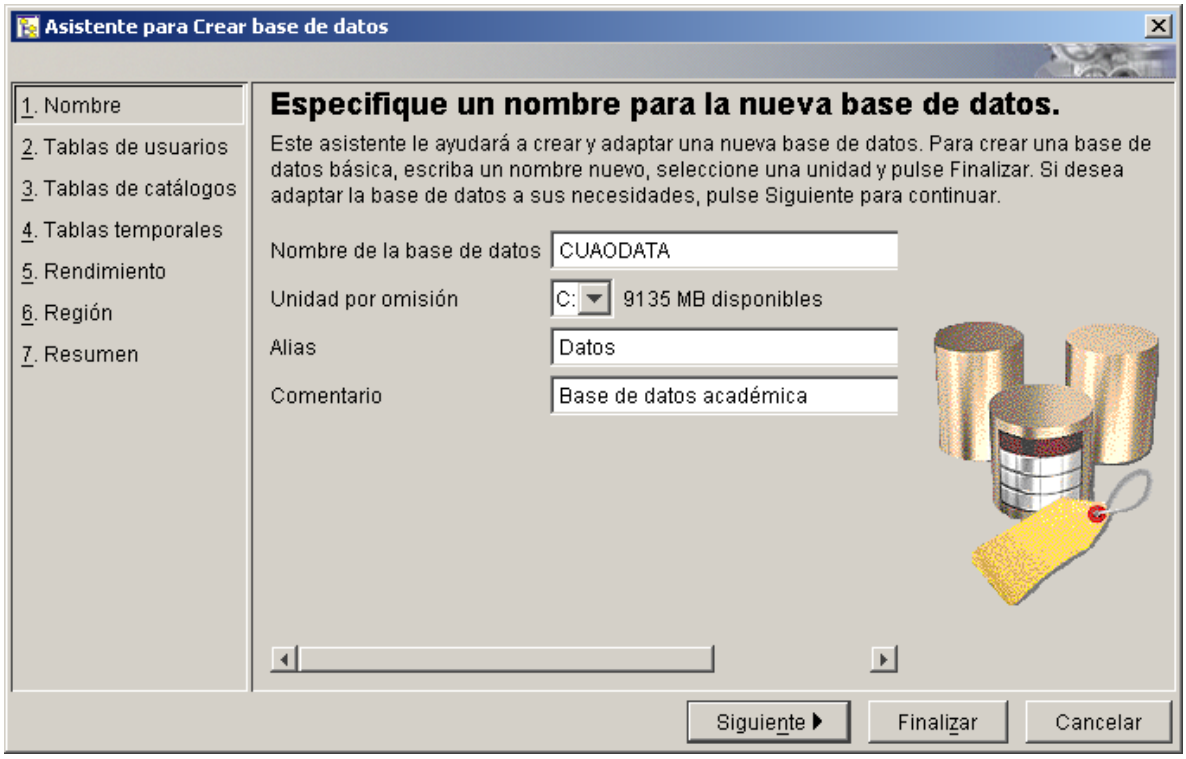


Figura 6. Creación de base de datos para información académica

Para poder tener acceso a esta base de datos se requiere registrarla en *ODBC*, para ello se invoca la opción *Programas -> IBM DB2 -> Asistente de configuración de cliente* del menú de *Inicio* de *Windows 2000 Professional*, se selecciona la base de datos *CUAODATA*, se oprime el botón *Propiedades* y se marca la base de datos como *fuente de datos del sistema*.

Luego de realizar este proceso, la lista de bases de datos *DB2* que pueden ser accesadas queda de la siguiente manera:

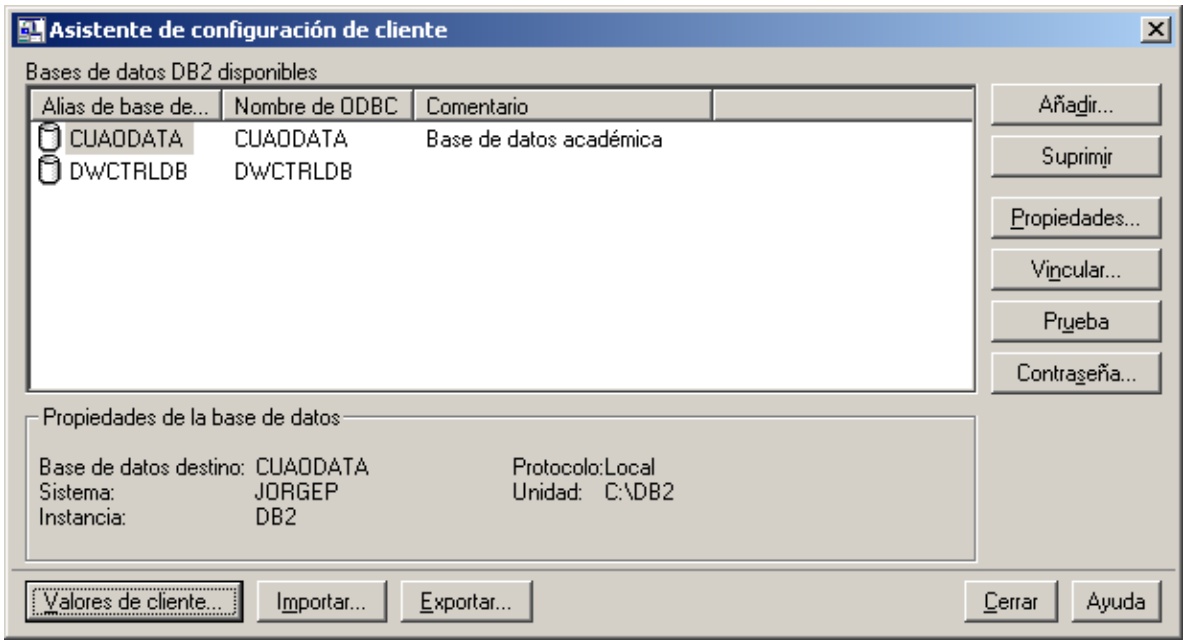


Figura 7. Configuración de base de datos para acceso por ODBC

También se requiere crear un acceso a datos desde *ODBC* para la información de la base de datos de registro académico. Para ello se invoca la herramienta de configuración de *ODBC* del sistema operativo y se crea un nuevo origen de datos de sistema para conectarse a bases de datos *Oracle*.

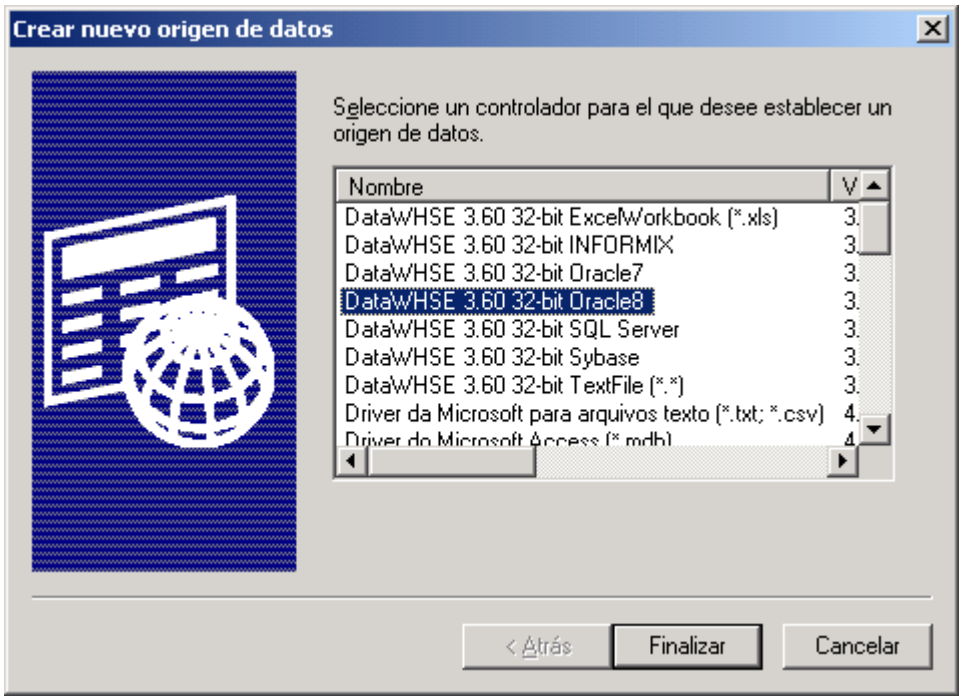


Figura 8. Creación de nuevo origen de datos ODBC

Este acceso de datos *ODBC* a *Oracle* se configura para tener acceso a la base de datos académica, de acuerdo con los parámetros de configuración de la herramienta *Oracle Client*.

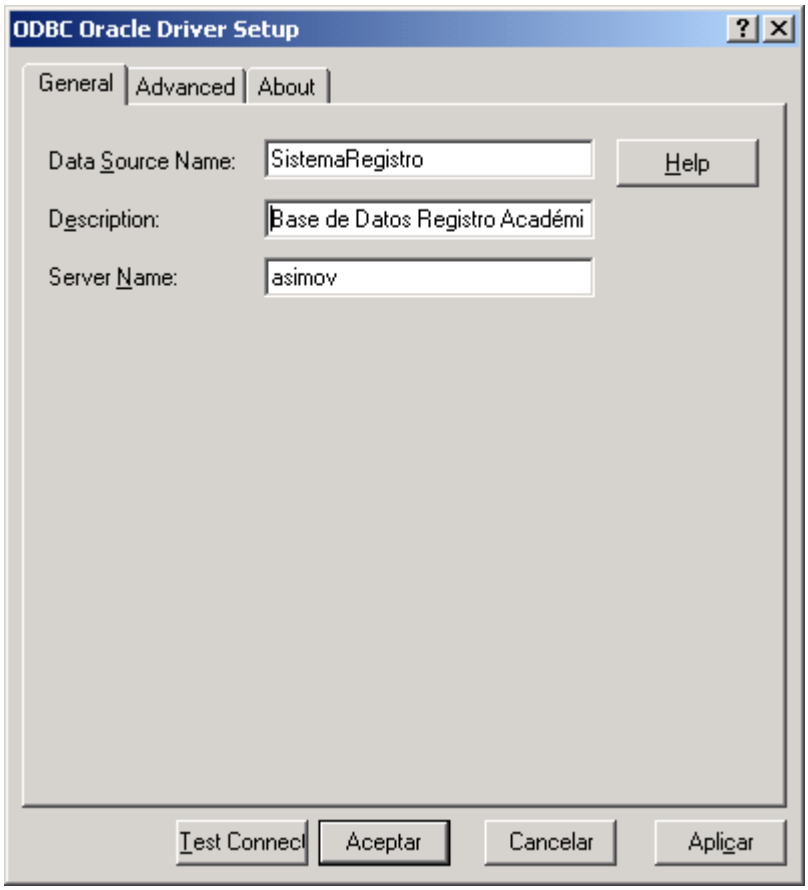


Figura 9. Parámetros de conexión a base de datos académica

Se requiere, de igual forma, crear una base de datos adicional en DB2, que servirá para almacenar los metadatos del proceso de transformación de la información almacenada en la base de datos de producción en *Oracle* a la base de datos de prueba en *DB2*. Esta base de datos se denominará *CTRLCUAO* y su proceso de creación se realiza desde la opción *Programas -> IBM DB2 -> Gestión de base de datos de control de depósito* dentro del menú de *Inicio* de *Microsoft*

Windows 2000. Este proceso define esta base de datos como la base de datos de control por omisión de *DB2*, para ello se invoca la opción

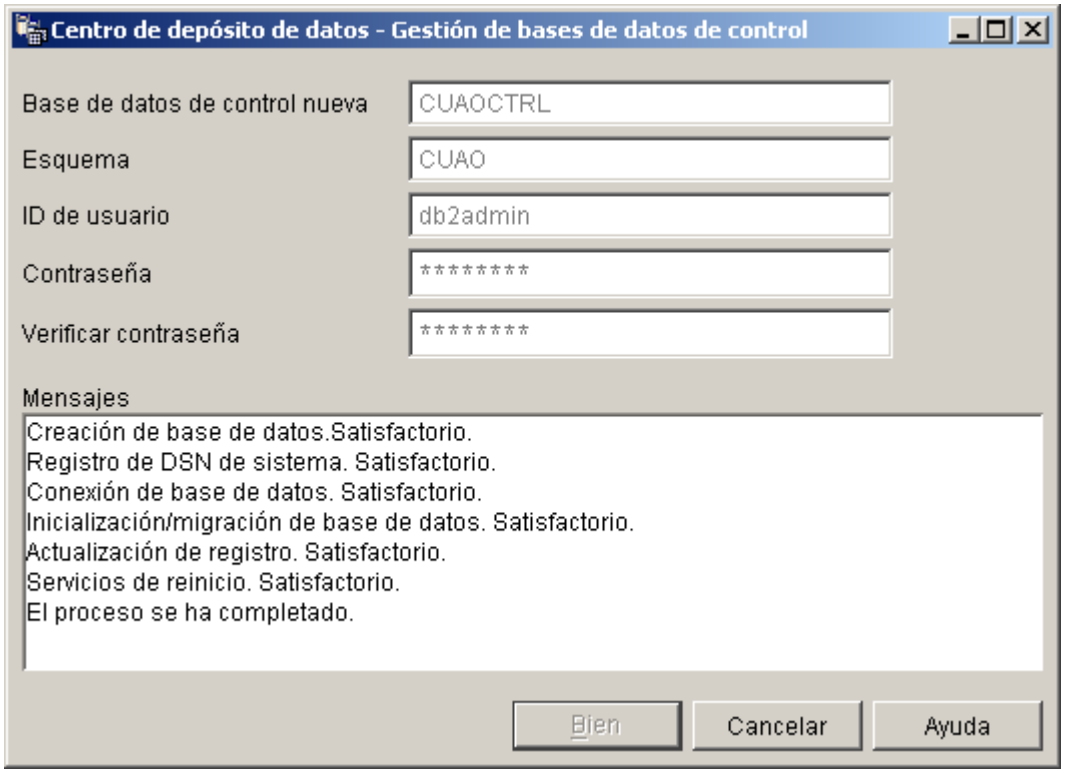


Figura 10. Creación de la base de datos de control de depósito

6.3.4.3. DEFINICIÓN DEL PROCESO DE IMPORTACIÓN DE DATOS

La definición de la importación y carga de datos se realiza desde el *Centro de depósito de datos*, el cual se invoca desde el menú de *Herramientas* del *Centro de Control* de *DB2*.

Para conectarse al *Centro de depósito de datos*, se deben ingresar los datos del usuario en la ventana de *Conexión al Centro de depósito de datos*. Se debe verificar, oprimiendo el botón *Avanzado...*, que la base de datos de control sea *CUAOCTRL*, en caso contrario se actualiza este valor, se oprime el botón *Bien* y se acepta en el botón *Bien...* de la venta de *Conexión al Centro de depósito de datos*.

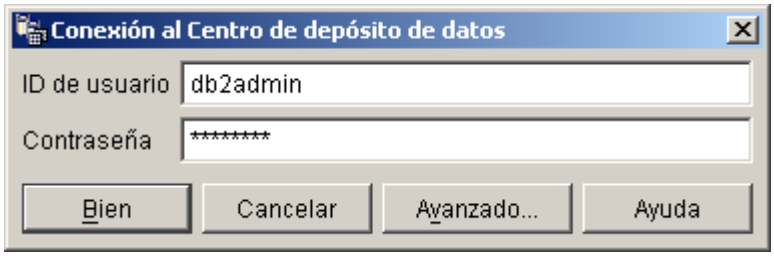


Figura 11. Ventana de Conexión al Centro de depósito de datos

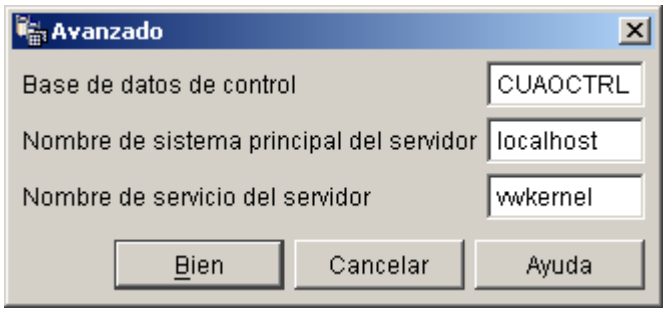


Figura 12. Ventana de opciones avanzadas de conexión al Centro de depósito de datos

Una vez conectado al *Centro de depósito de datos* se procede a definir las *fuentes de depósito*, es decir, la definición de las fuentes de información para el proceso de minería. Es acá donde se pueden realizar transformaciones sobre los datos, que permitan seleccionar los conjuntos de datos apropiados para realizar los procesos de minería y descubrimiento de conocimiento.

Se crea una nueva fuente de depósito, seleccionando la opción *Fuentes de depósito -> definir -> Oracle* desde la ventana de exploración del *Centro de depósito de datos*.

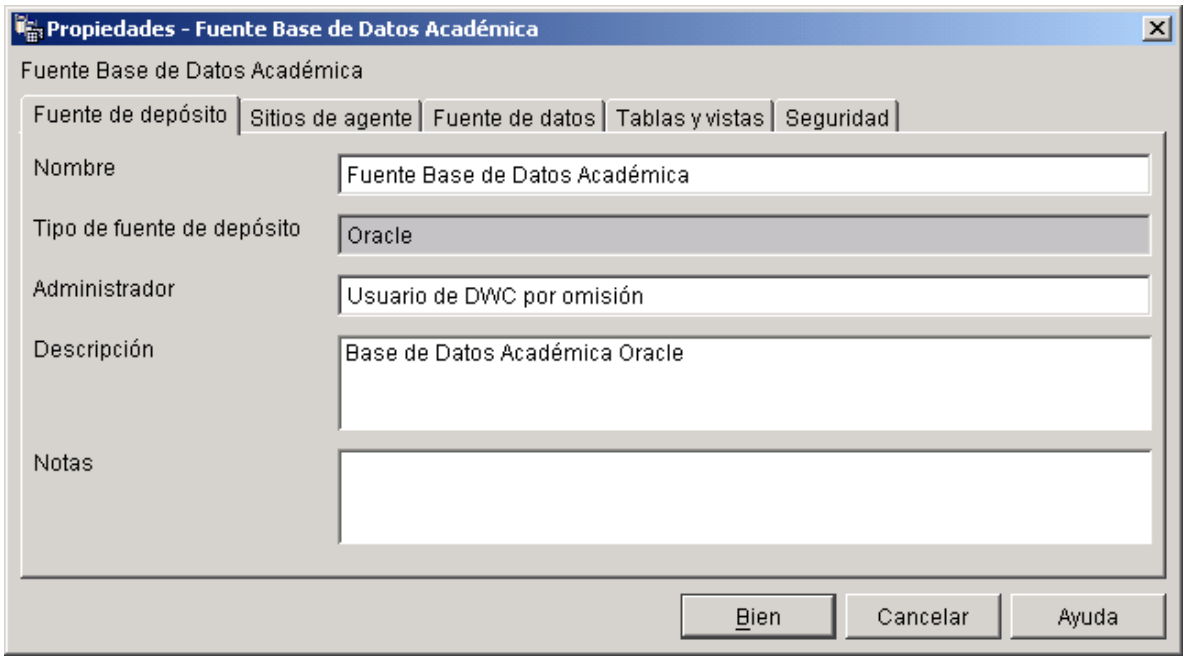


Figura 13. Definición de una nueva fuente de depósito basada en Oracle

En la pestaña *Fuente de datos* se selecciona la conexión *ODBC* a la base de datos *Oracle* creada anteriormente.

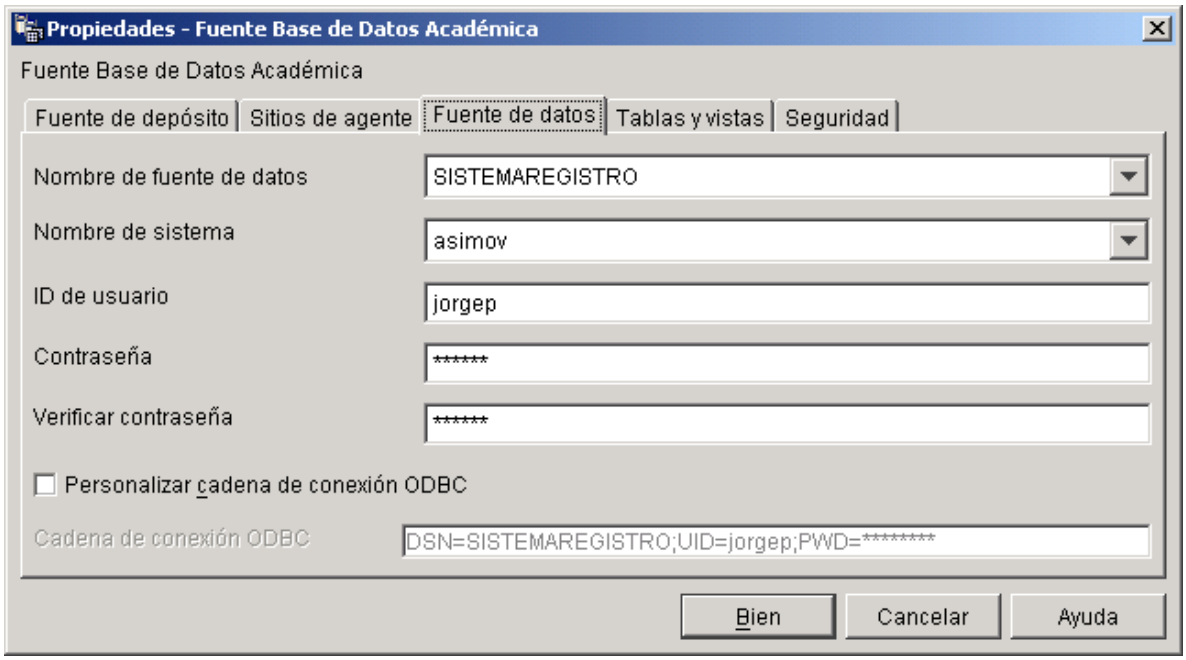


Figura 14. Definición de la fuente de datos de una fuente de depósito

Por último, se definen las tablas o vistas que se usaran como fuentes de datos del modelo.

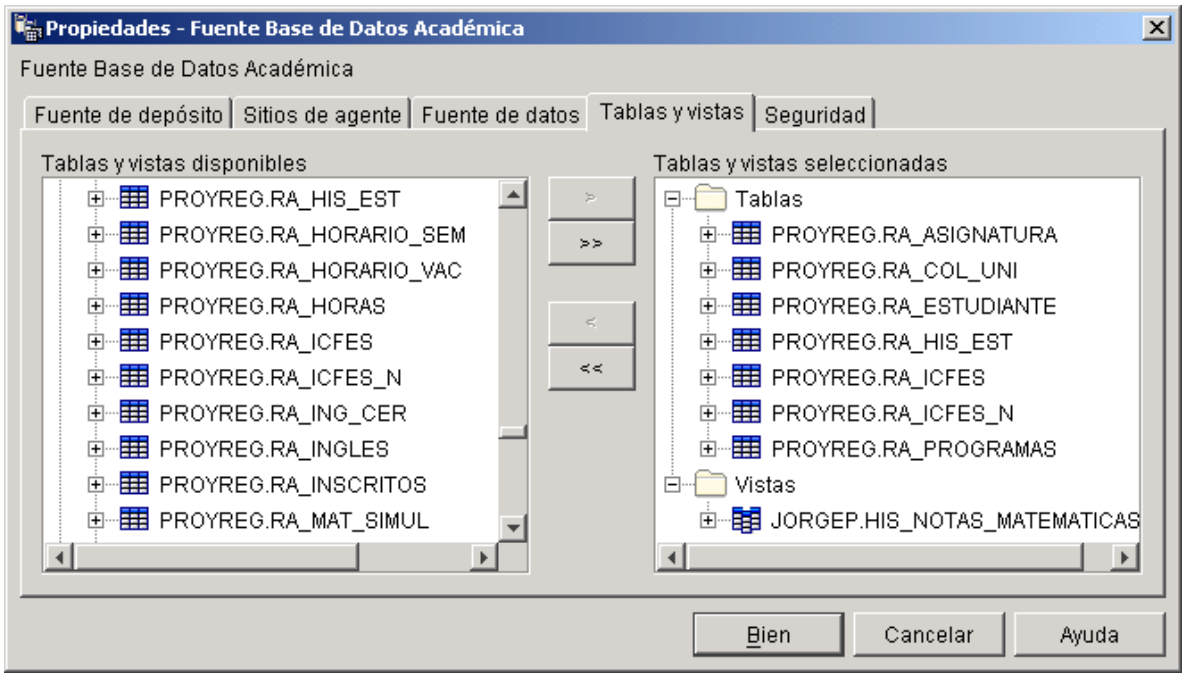


Figura 15. Selección de tablas para el modelo

Se debe crear ahora una nueva fuente de depósito la cual va a apuntar al archivo plano donde se encuentra la información de las pruebas diagnósticas aplicadas a los estudiantes de los primeros cursos de matemáticas en la universidad.

Para esta nueva fuente de datos se selecciona el tipo de fuente de depósito como *archivos locales* así:

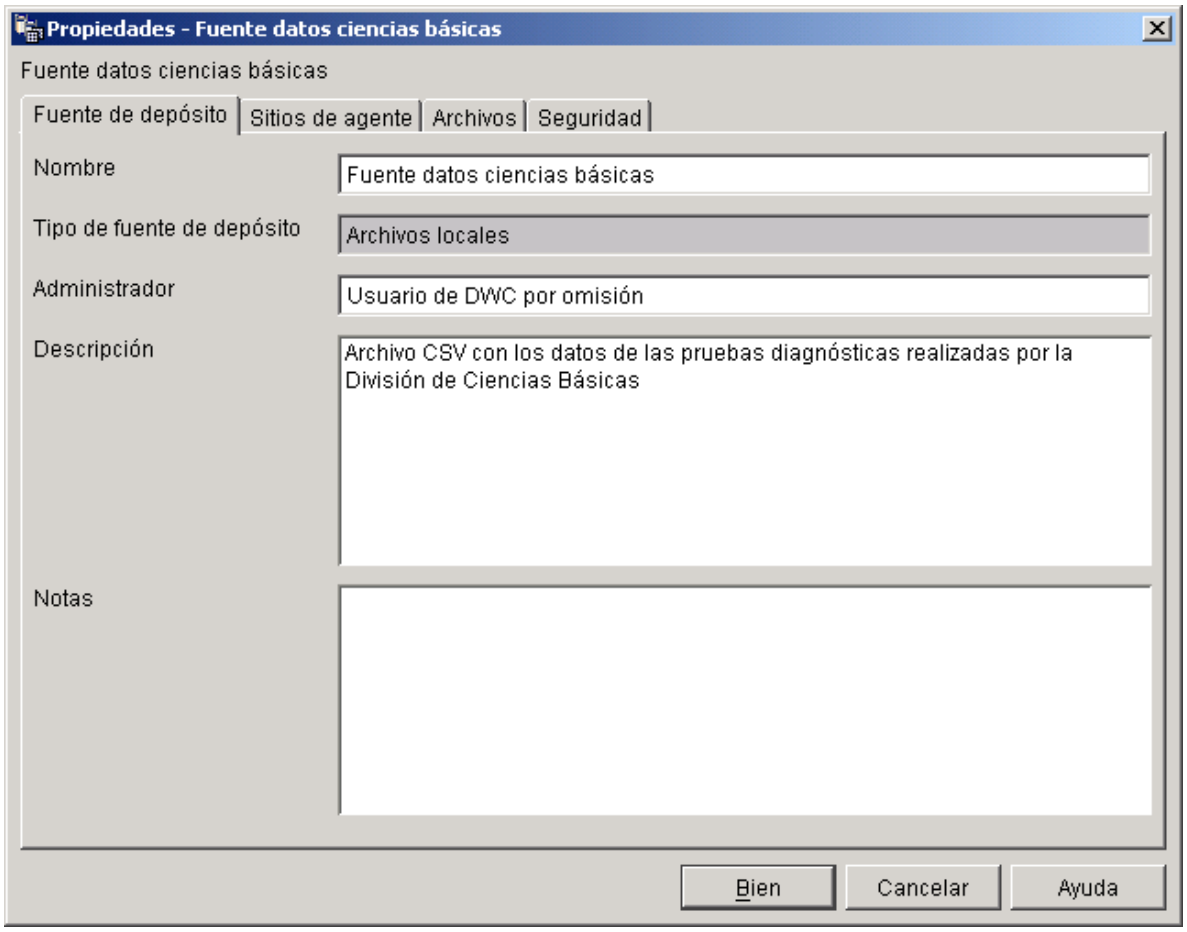


Figura 16. Definición de una nueva fuente de depósito basada en archivos planos

Se define ahora el nombre del archivo y los campos que se van a importar del archivo plano.

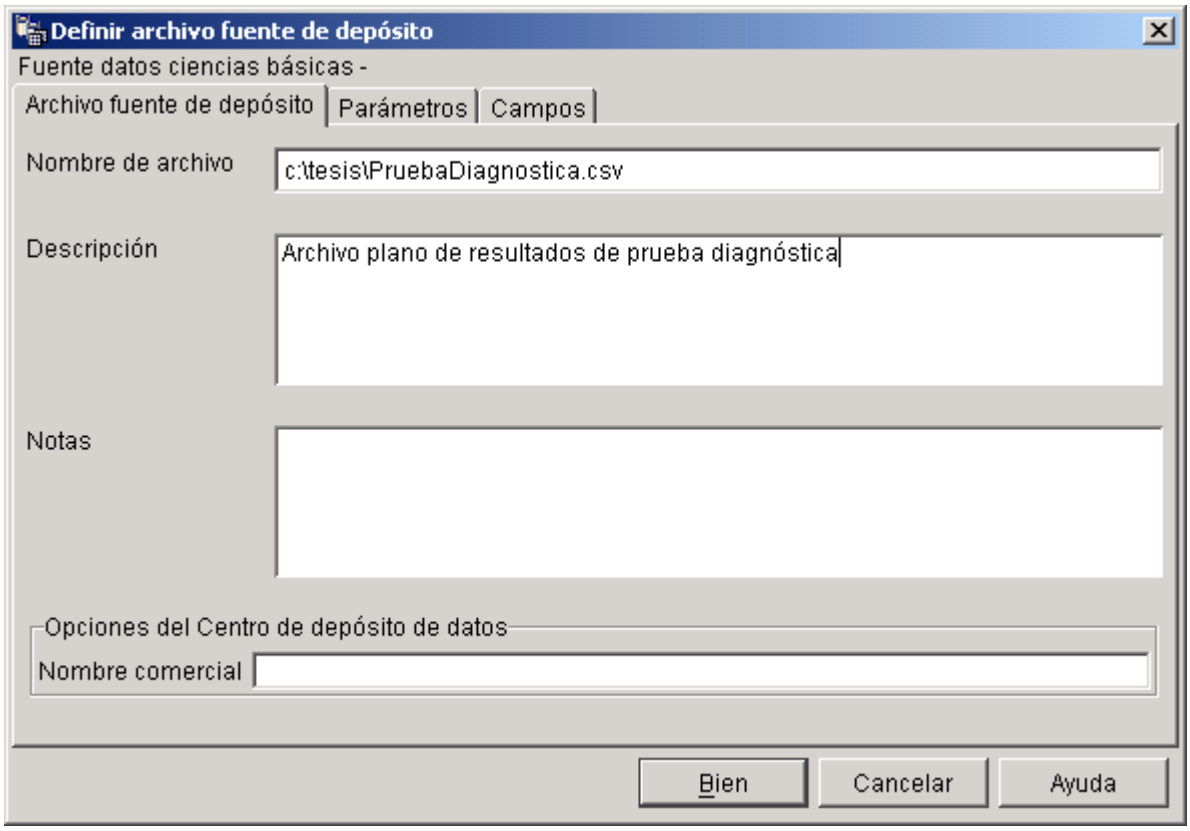


Figura 17. Definición de archivo plano como fuente de depósito

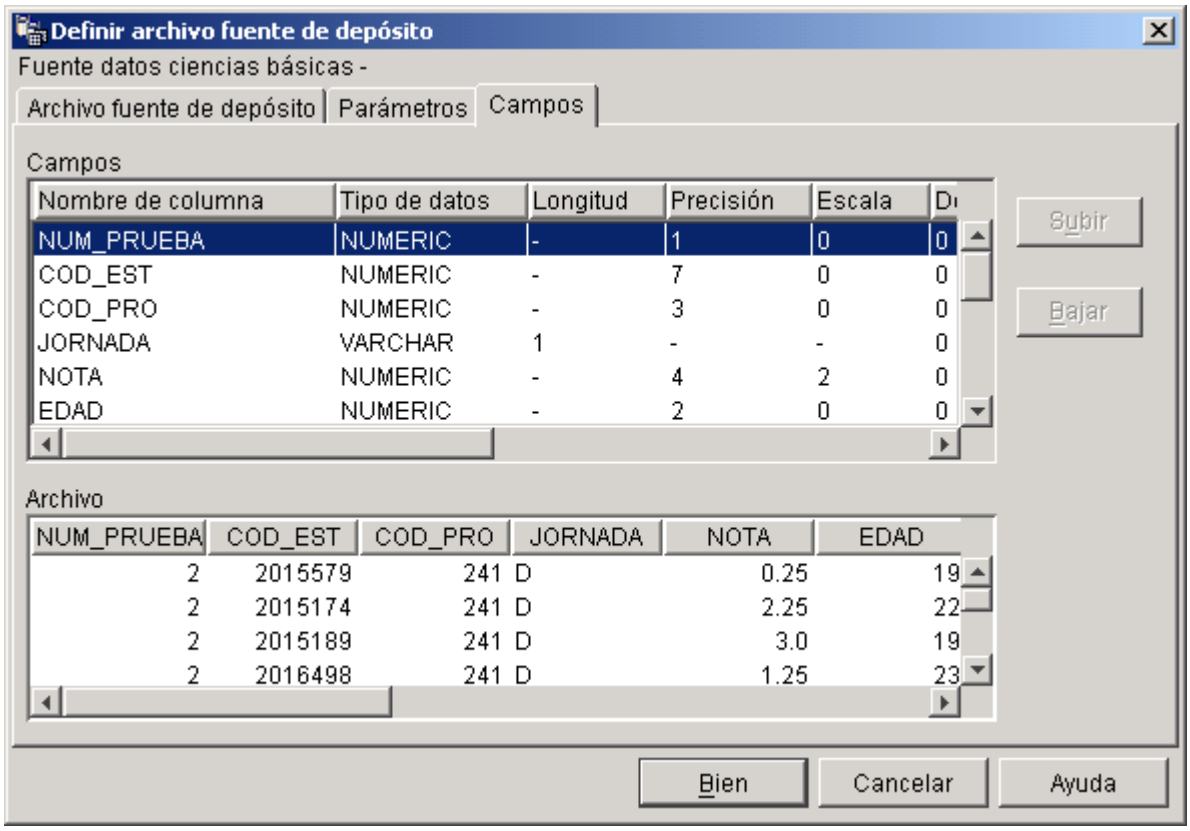


Figura 18. Definición de campos a importar desde archivo plano

El siguiente paso es definir un destino de depósito, el cual se denominará *Destino Registro Académico* y es donde se almacenarán las tablas de salida que genere el proceso de transformación. Se define en primera instancia el nombre del destino de depósito, y luego en la pestaña *Base de datos*, el equipo, base de datos y esquema de usuario donde se almacenarán las tablas de salida.

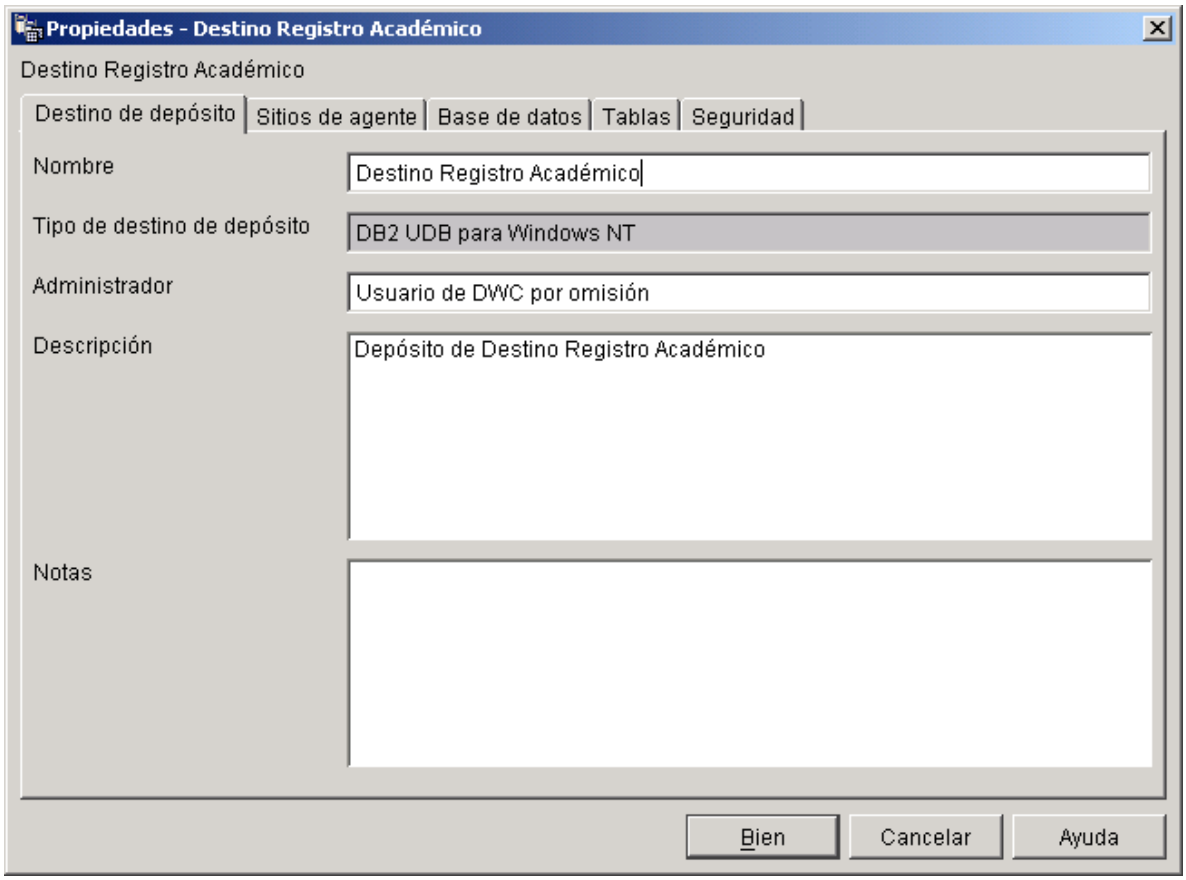


Figura 19. Definición Destino de Depósito

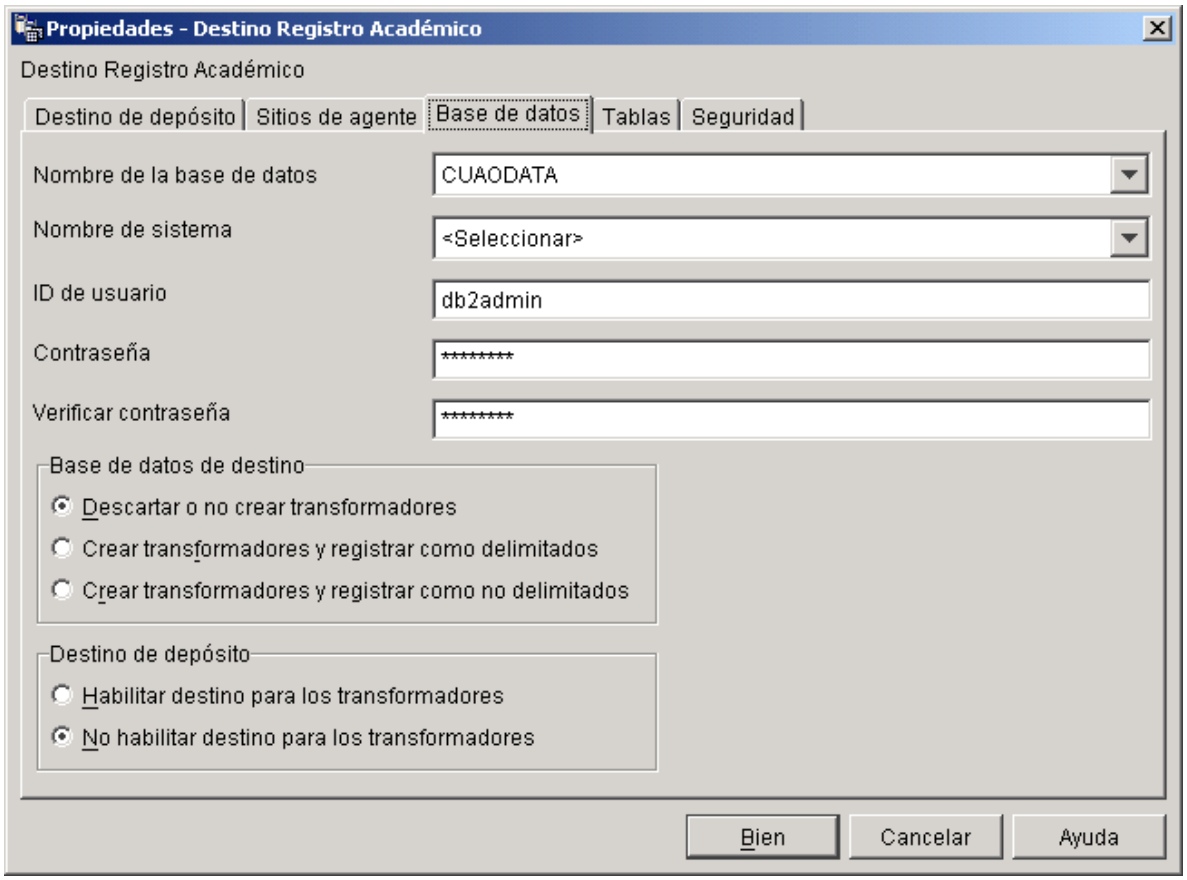


Figura 20. Definición de base de datos para la bodega de datos

El siguiente paso es definir una nueva *área temática*, denominada *Análisis Mortalidad Académica*, lo cual se hace desde la carpeta de *Áreas temáticas* del *Centro de depósito de datos*.

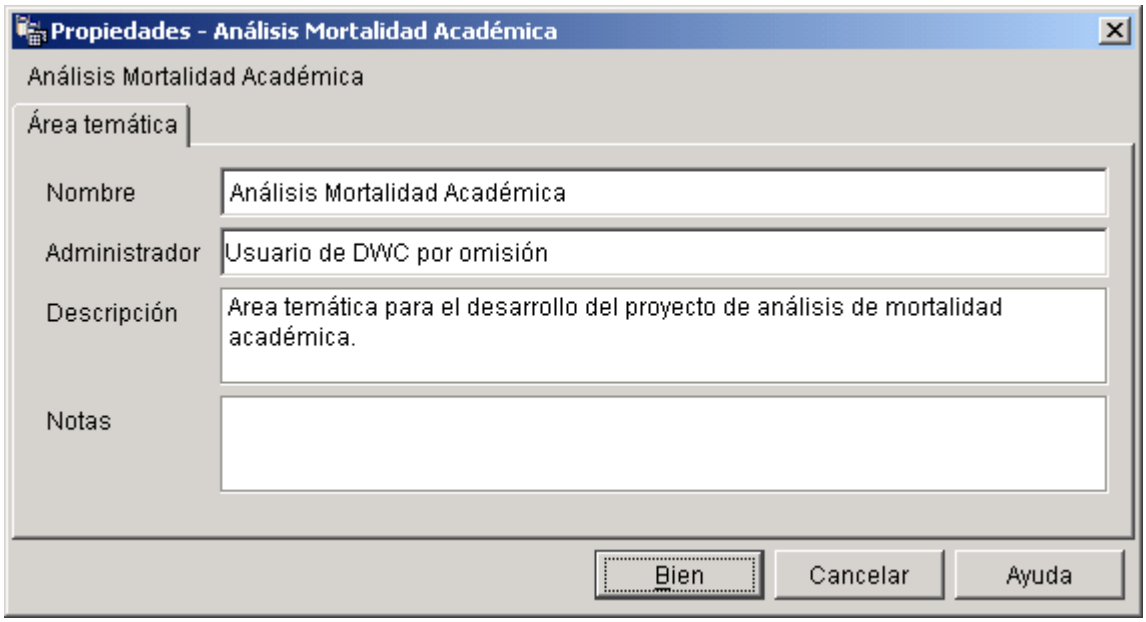


Figura 21. Definición de una nueva área temática

Una vez definida la nueva área temática, se procede a definir un proceso o conjunto de procesos dentro de dicha área temática. Se definirá inicialmente el proceso de carga de datos sobre mortalidad académica.

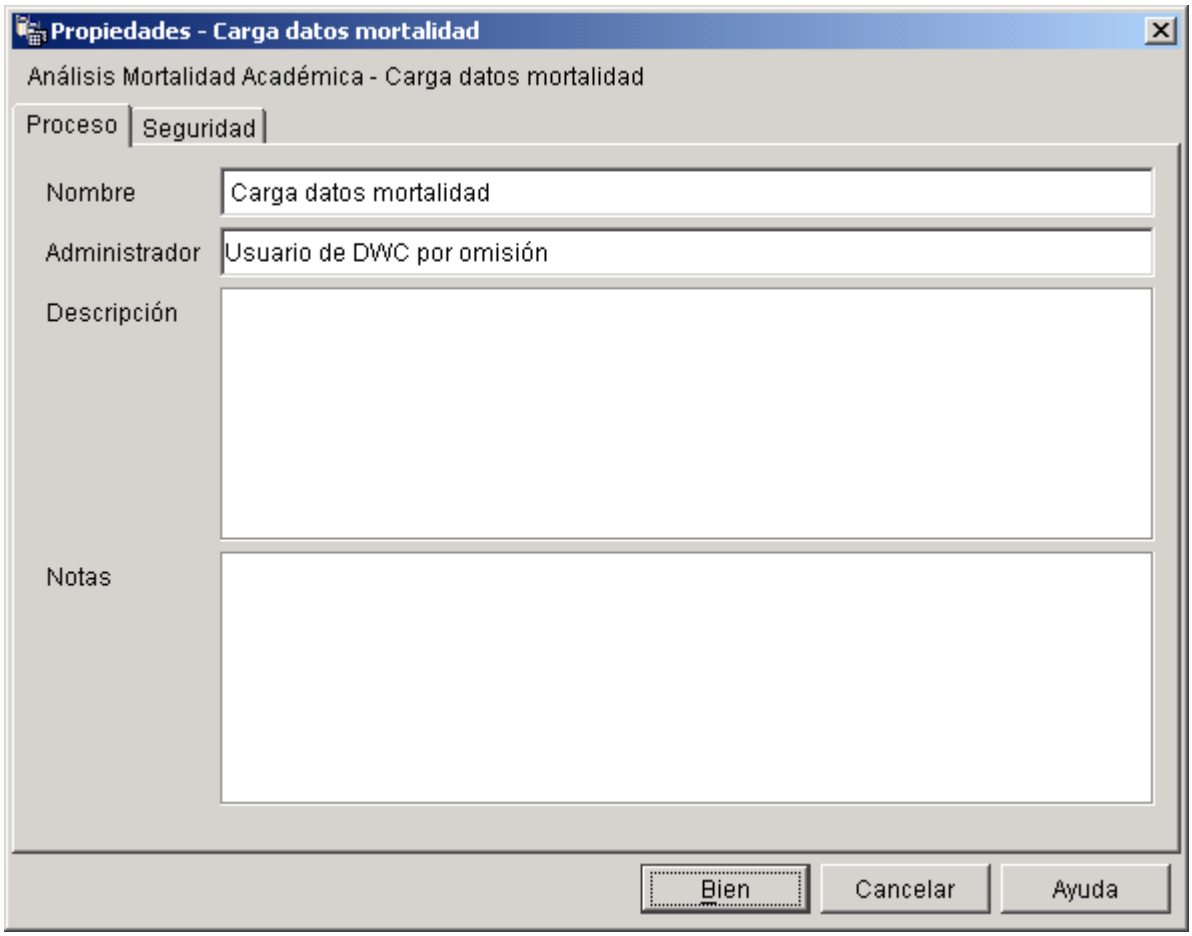


Figura 22. Definición de un nuevo proceso.

Una vez creado el proceso, se selecciona con el botón derecho y se presiona la opción *abrir*, con lo cual se abre la ventana de diseño del proceso.

El proceso de carga de datos de mortalidad académica consta de los siguientes tres pasos:

- Importar datos de estudiante desde la base de datos de Registro Académico

- Importar datos de pruebas diagnósticas desde el archivo plano

- Combinar los datos de estudiante y los de pruebas diagnósticas para tener una sola tabla para desarrollar sobre ella los procesos de minería de datos.

En primer lugar se deben ubicar sobre la ventana de diseño del proceso las dos fuentes de datos que se van a usar, para ello se selecciona la herramienta de añadir datos así:

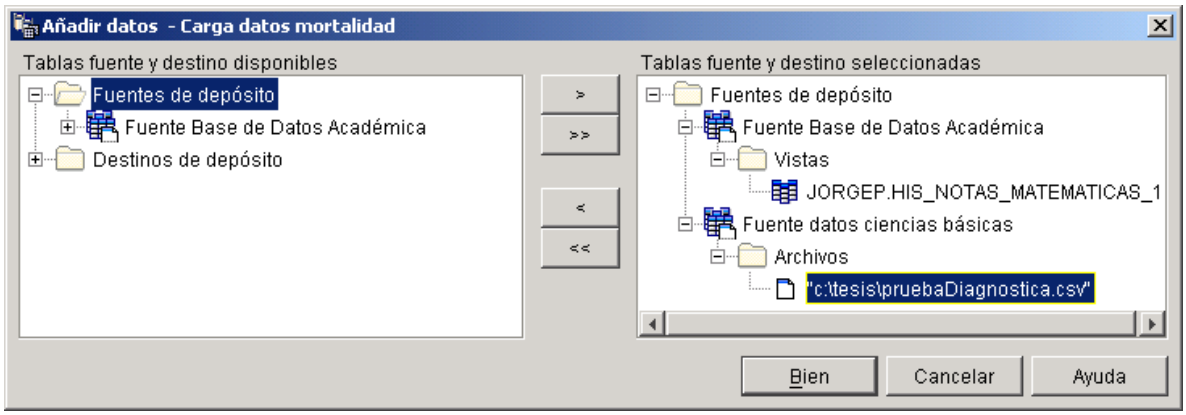


Figura 23. Selección de fuentes de depósito para el proceso de carga

A continuación se crea un paso para el cargue de los resultados de pruebas diagnósticas utilizando la herramienta de carga de archivos planos, se le asigna como fuente de este proceso a la fuente de archivos planas que se colocó sobre la ventana de diseño en el paso anterior.

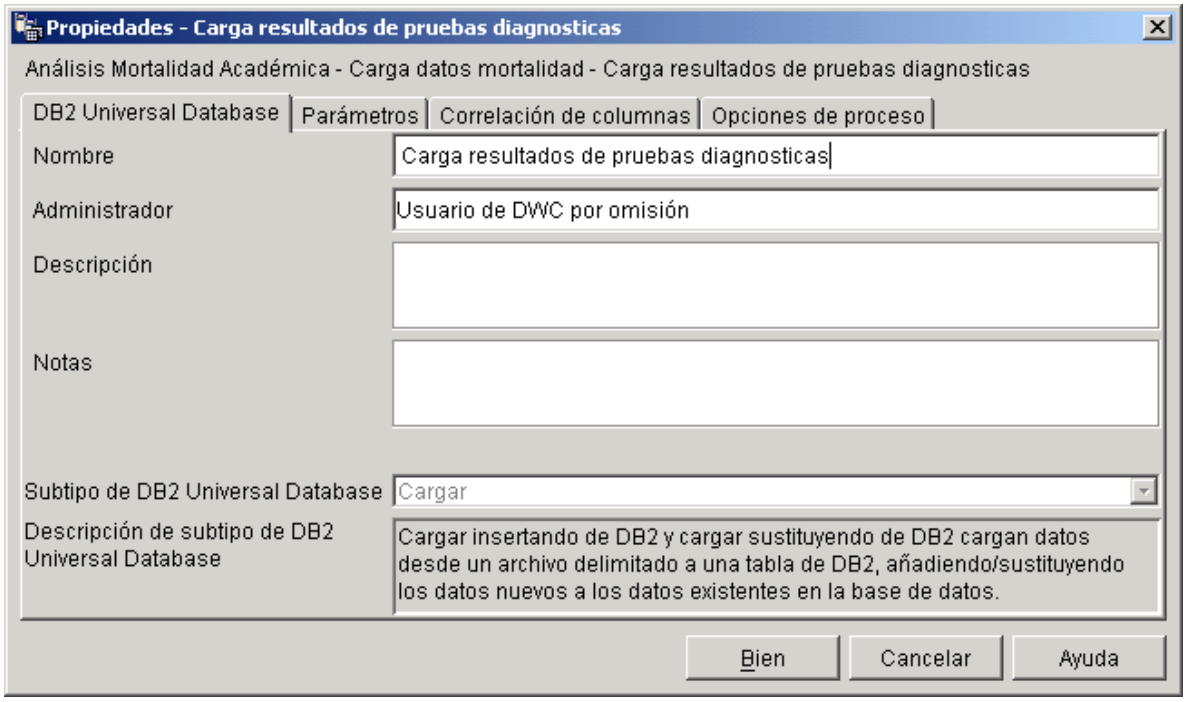


Figura 24. Definición del proceso de carga de archivos planos

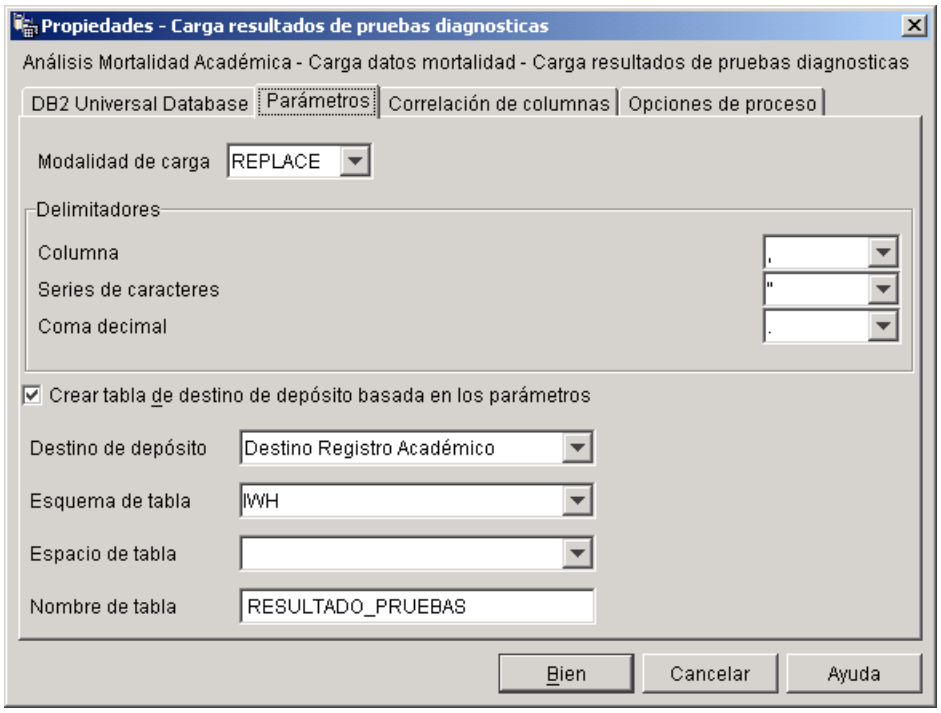


Figura 25. Definición de sitio de descarga del proceso de lectura de archivos planos

A continuación se crea un paso para el cargue de la tabla de datos de notas y estudiantes traídos desde la base de datos de Registro Académico, usando la herramienta de carga de datos vía SQL, se le asigna como fuente de este proceso a la fuente de datos de la vista traída de la base de datos Oracle que se colocó sobre la ventana de diseño en el paso anterior.

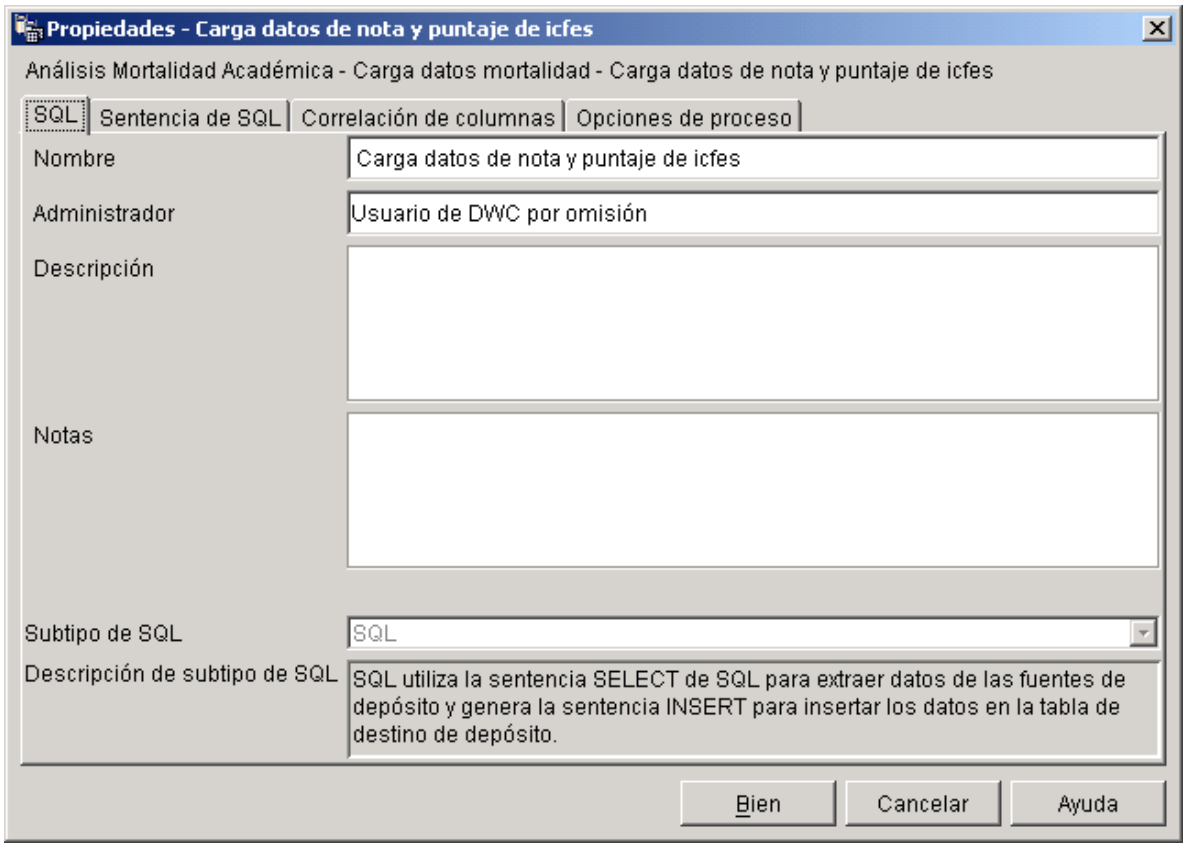


Figura 26. Definición proceso de carga de datos de la base de datos de registro

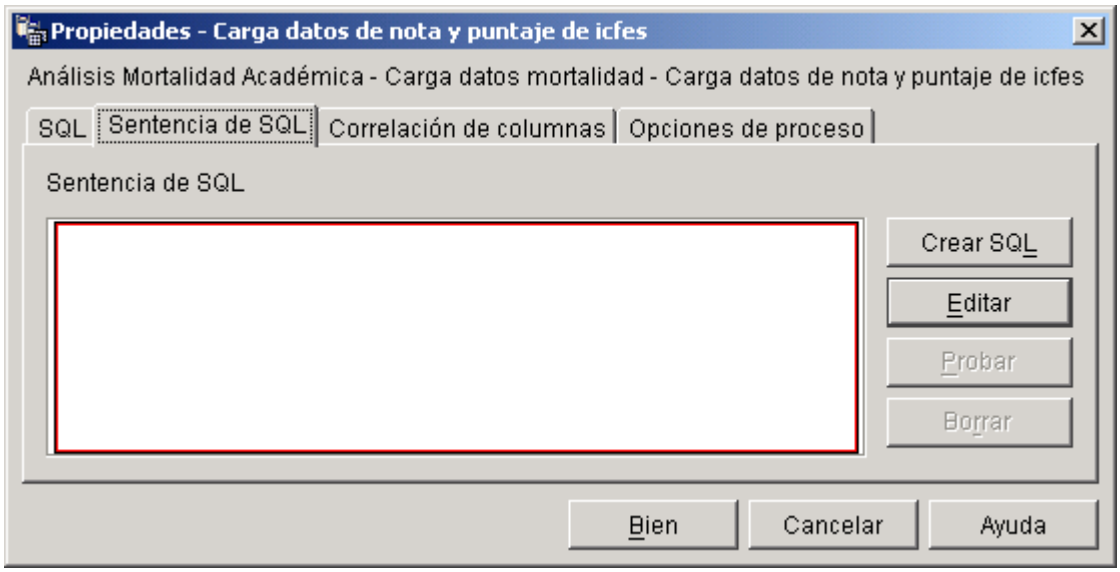


Figura 27. Creación sentencia SQL para lectura de datos de origen

Se selecciona el botón *Crear SQL* y se seleccionan las tablas, columnas

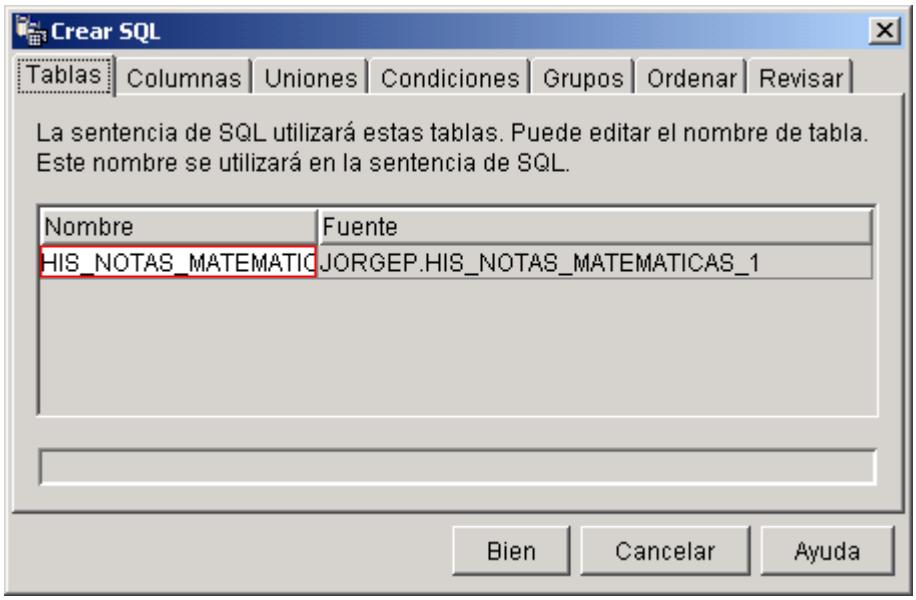


Figura 28. Selección de tablas para consulta

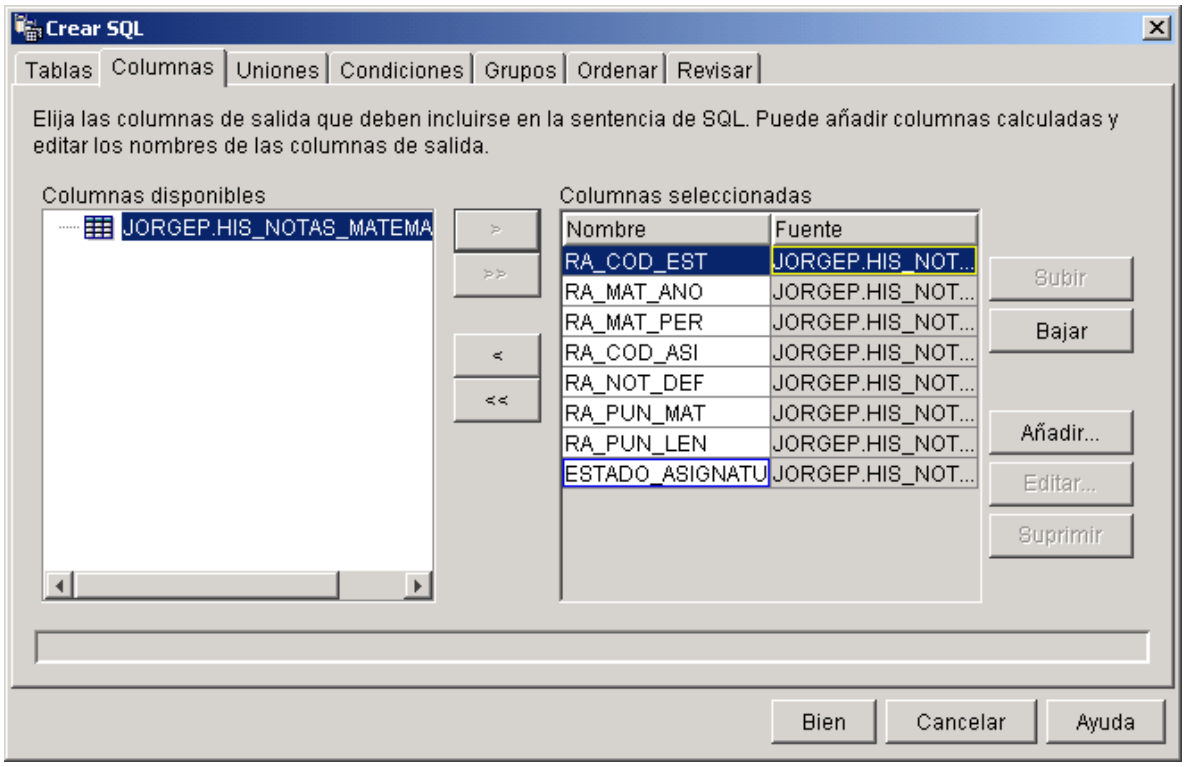


Figura 29. Selección de columnas a incluir en la recuperación

Finalizada esta parte se selecciona el botón *Bien* para regresar a la ventana de configuración del proceso y en la pestaña de *Correlación de Columnas* se selecciona el botón de *Generar Tabla por Omisión* para dejar que el proceso de transformación automáticamente cree una nueva tabla en el deposito de datos de destino definido.

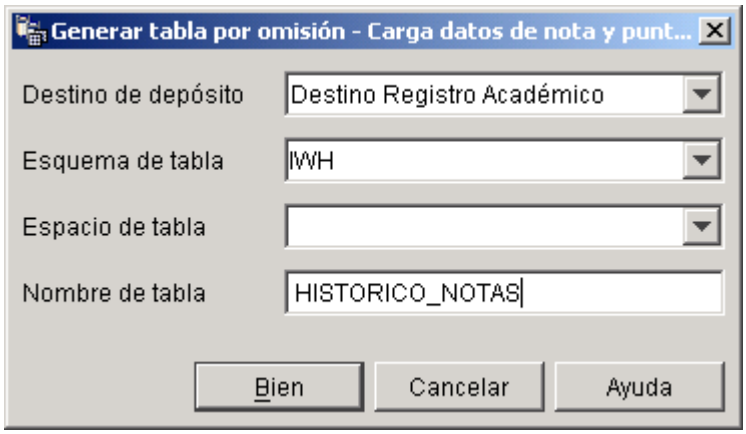


Figura 30. Generación de tabla por omisión

La ventana de diseño del proceso está ahora de la siguiente manera:

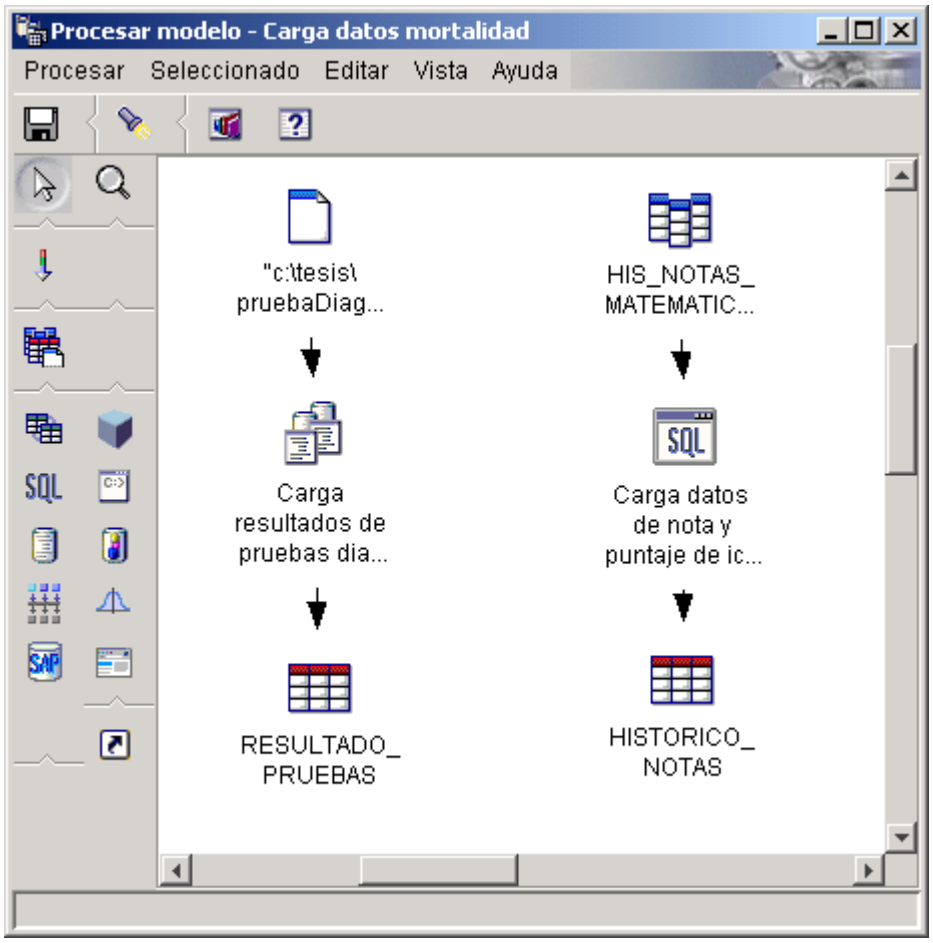


Figura 31. Ventana de diseño del modelo transformación de datos.

A continuación se crea un nuevo paso que relaciona las tablas RESULTADO_PRUEBAS e HISTORICO_NOTAS para generar la tabla ESTADO_ESTUDIANTES sobre la cual se correrán los procesos de minería de datos con la herramienta *Intelligent Miner*.

Posteriormente cado paso del proceso debe ser promocionado de la modalidad de desarrollo a la modalidad de prueba, la cual es la que permitirá su ejecución, la

que se realiza dando click con el botón derecho sobre cada paso y seleccionando la opción *probar*.

Una vez seleccionada la opción *probar* en cada uno de los pasos ya es posible acceder a la base de datos de la bodega desde la herramienta de minería de datos. Los pasos de transformación se pueden pasar a la modalidad de producción donde ya pueden ser planificados para ejecutarse periódicamente y pueden ser invocados por usuarios de la base de datos.

La ventana de diseño del proceso queda entonces así:

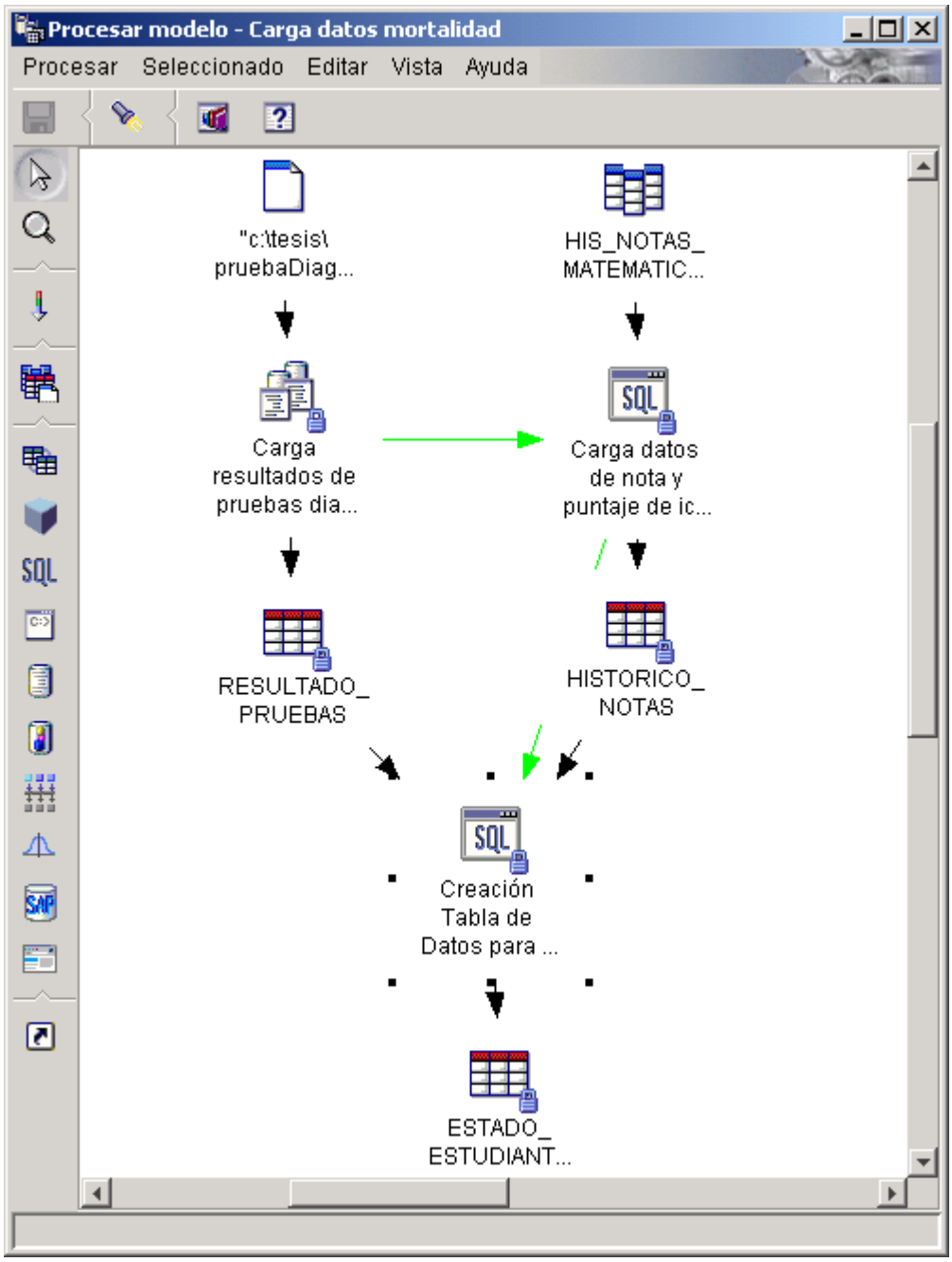


Figura 32. Ventana de diseño del proceso de transformación de datos.

6.3.4.4. IMPLEMENTACION PROCESO DE MINERIA DE DATOS

En primer lugar se debe iniciar el programa *Intelligent Miner*, el cual se invoca desde el menú *Inicio -> IBM DB2 Intelligent Miner for Data V8.1*. La ventana principal de *Intelligent Miner* luce así:

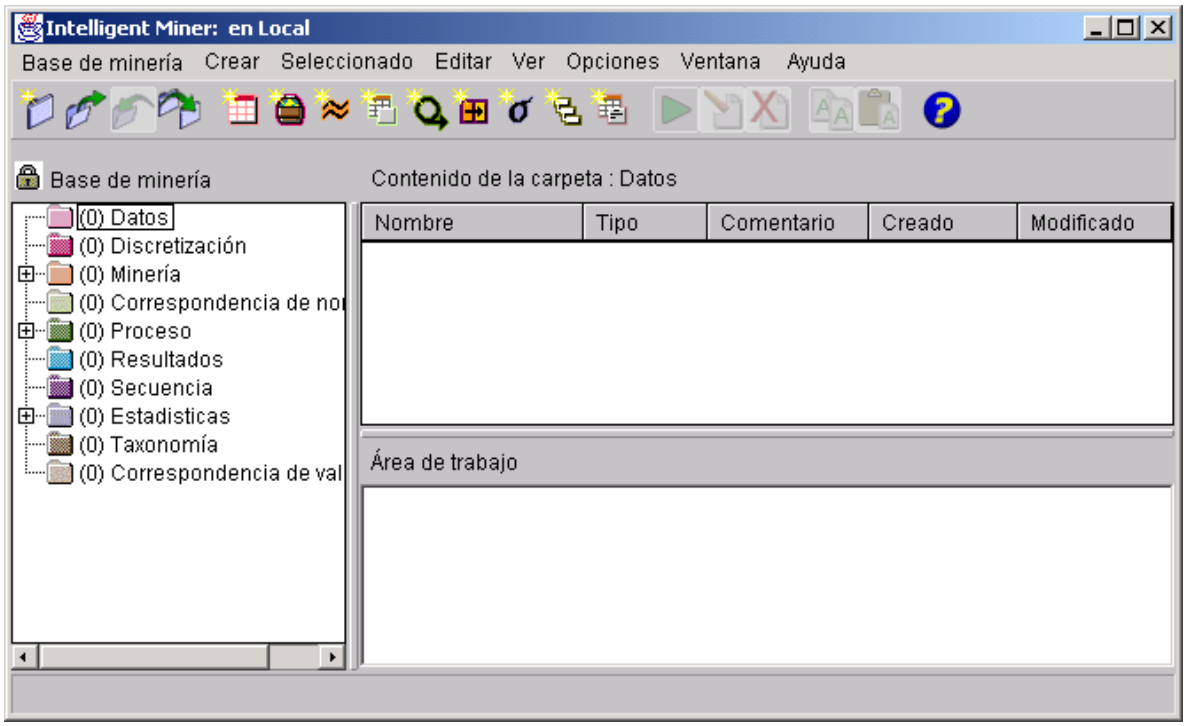


Figura 33. Ventana principal de Intelligent Miner

El primer paso es crear un nuevo conjunto de datos, para lo cual se selecciona en la columna izquierda la opción *Crear Datos* con el botón derecho sobre el elemento *Datos* del panel izquierdo. En primer lugar se define el nombre con

que se identificará el conjunto de datos y en segundo lugar el origen de dicho conjunto de datos.

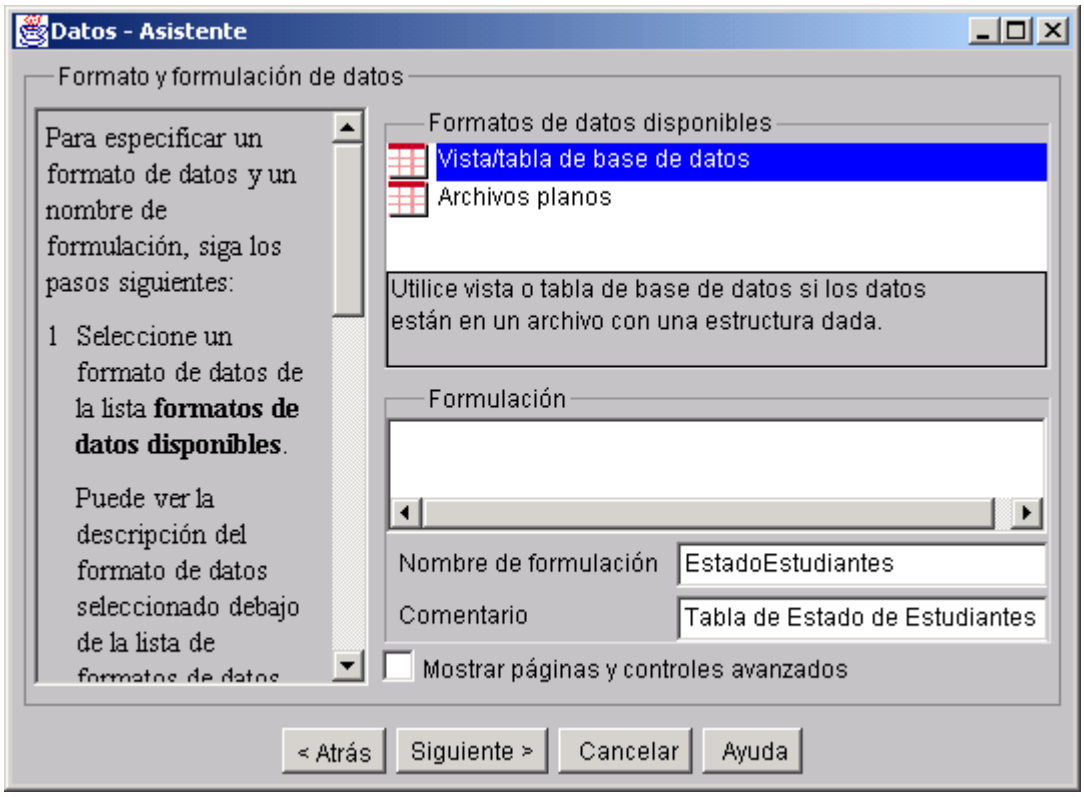


Figura 34. Definición de Datos para los procesos de minería

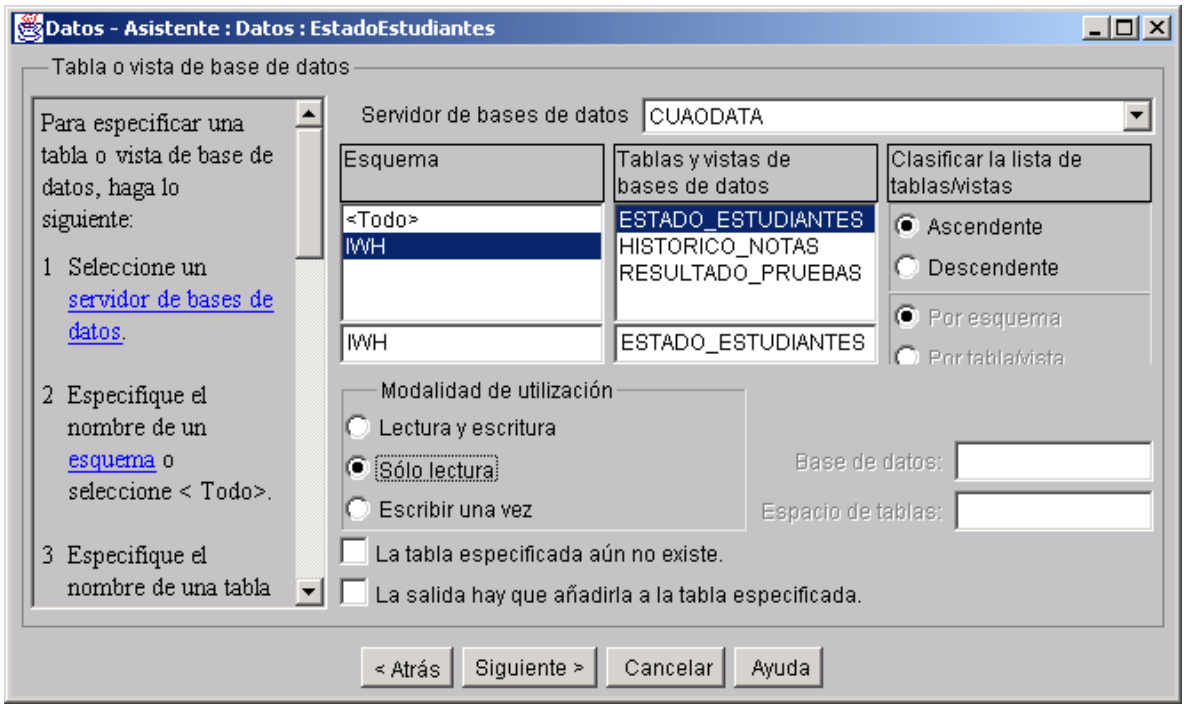


Figura 35. Definición de origen de datos

A continuación se define el tipo de datos de cada una de las columnas de la tabla, con lo que finaliza el proceso de definición de la fuente de datos de minería.

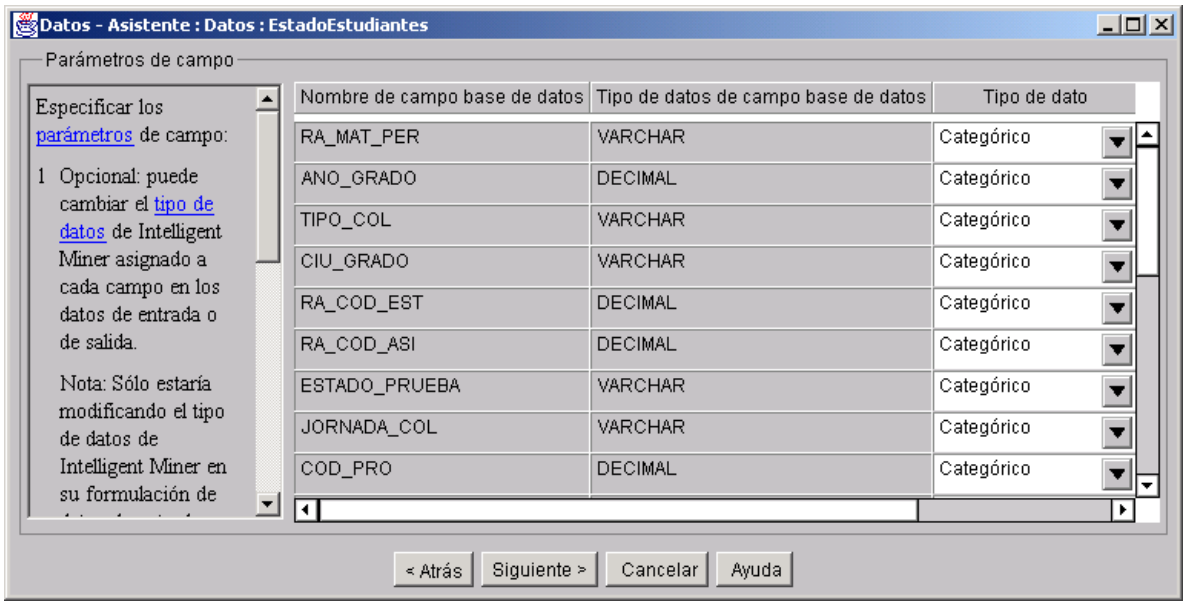


Figura 36. Definición de tipos de datos de columna

Para crear el proceso de minería se selecciona en el menú de tipos de minería a utilizar la categoría **Clasificación**.

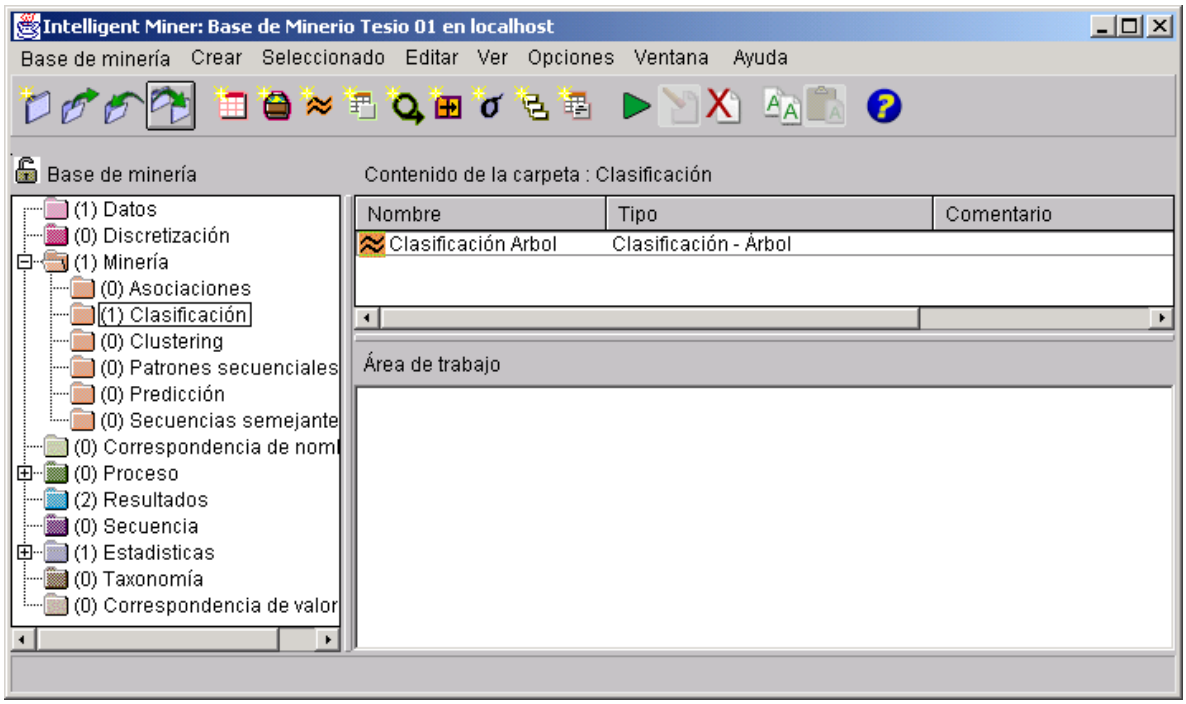


Figura 37. Selección de tipo de minería a utilizar

El proceso de clasificación puede realizar por árboles o por redes neuronales, se optará por el tipo de clasificación por árboles debido a que los datos son en su mayoría de carácter cualitativo.

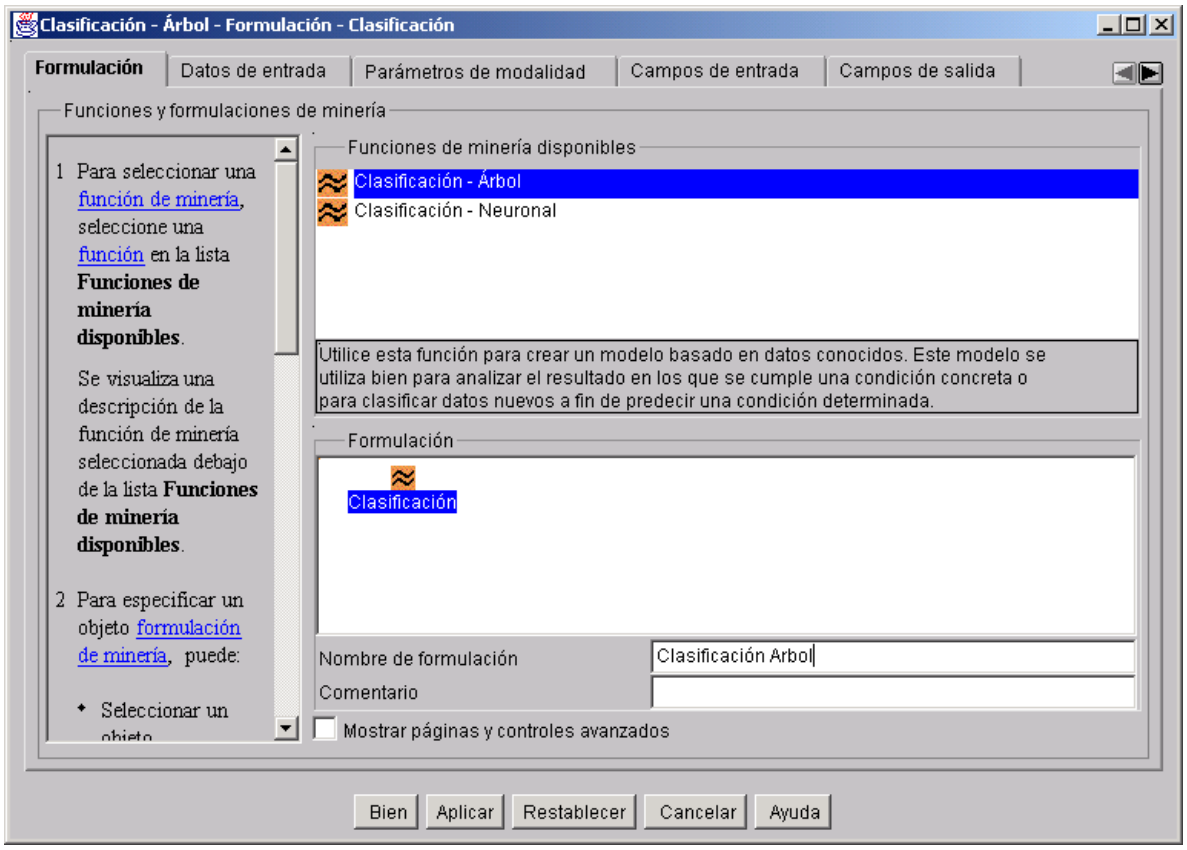


Figura 38. Selección de tipo de clasificación.

Los datos de entrada del modelo se toman del modelo de datos definido previamente:

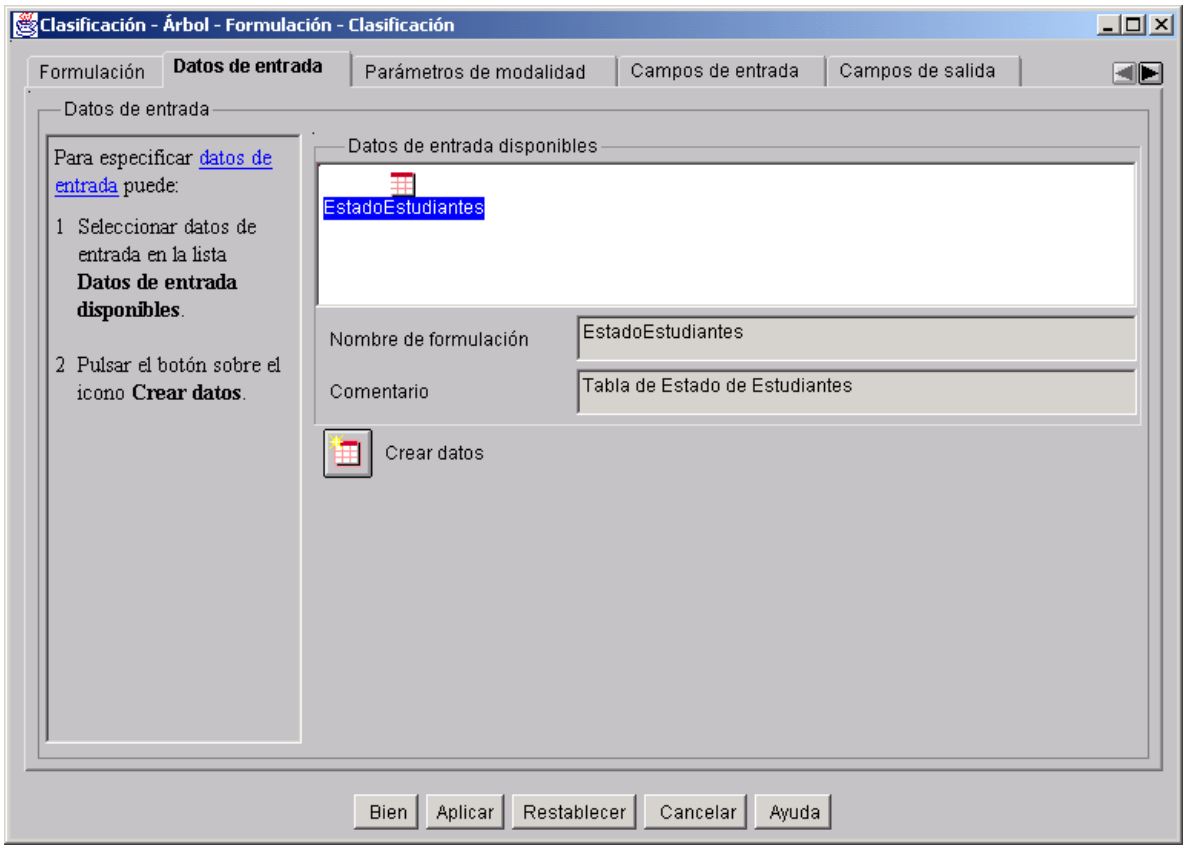


Figura 39. Selección de modelo de datos para ejecutar minería

Se selecciona como etiqueta de clase la representada por el estado final de la asignatura (G para ganadas y P para perdidas), e igualmente se seleccionan las variables que participarán en la clasificación:

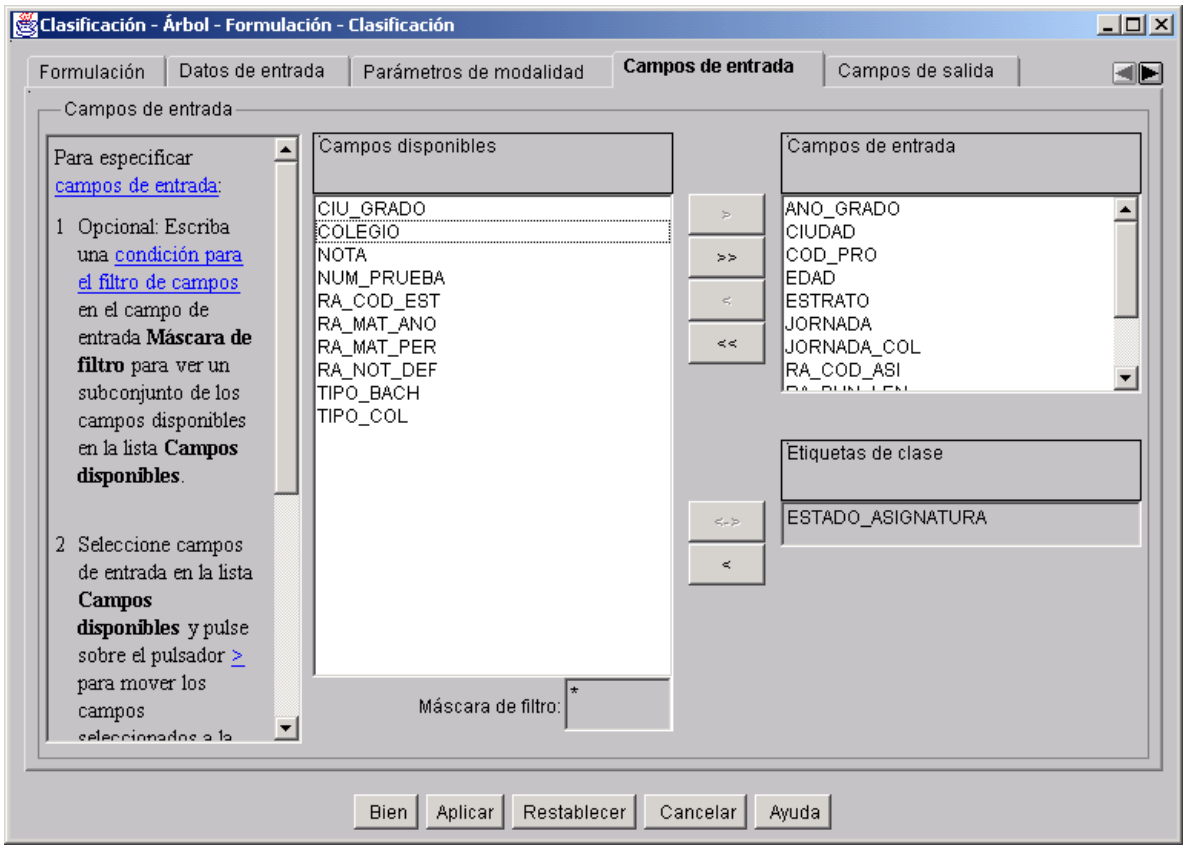


Figura 40. Selección de variables de clasificación

Por último se ejecuta el modelo desde la ventana principal, seleccionando el modelo de minería definido y pulsando el icono de ejecución.

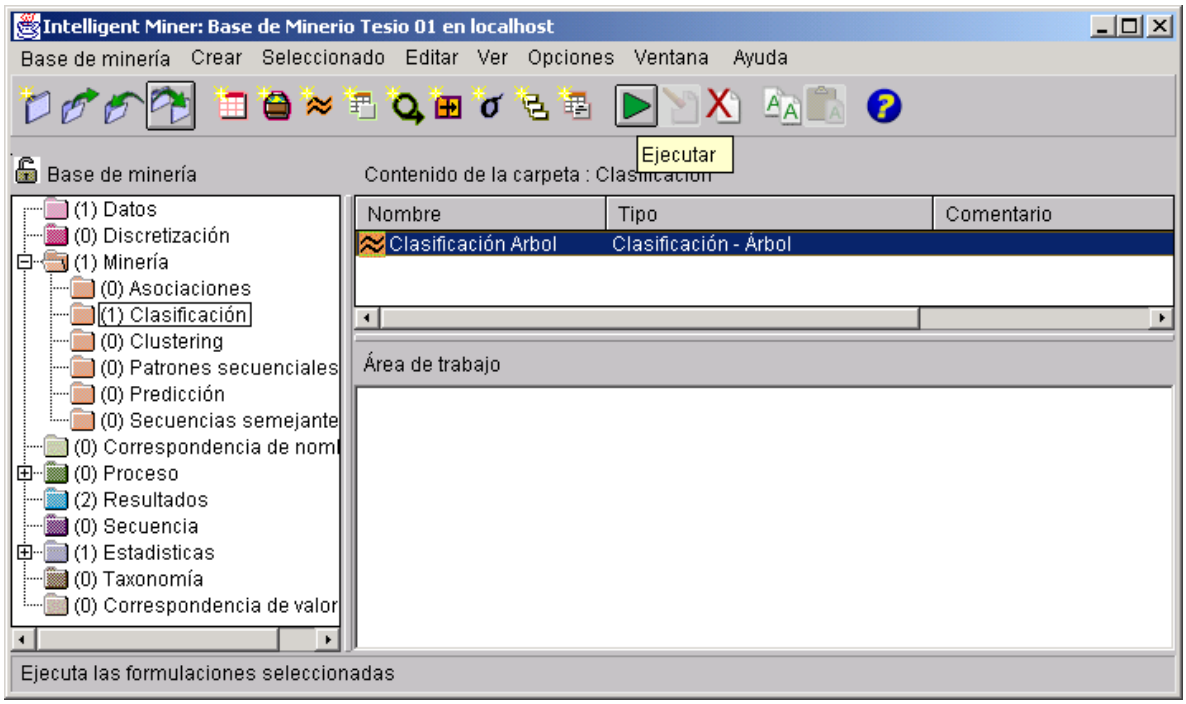


Figura 41. Ejecución del modelo de minería.

El proceso de ejecución del modelo de minería se presenta en el siguiente cuadro de avance:

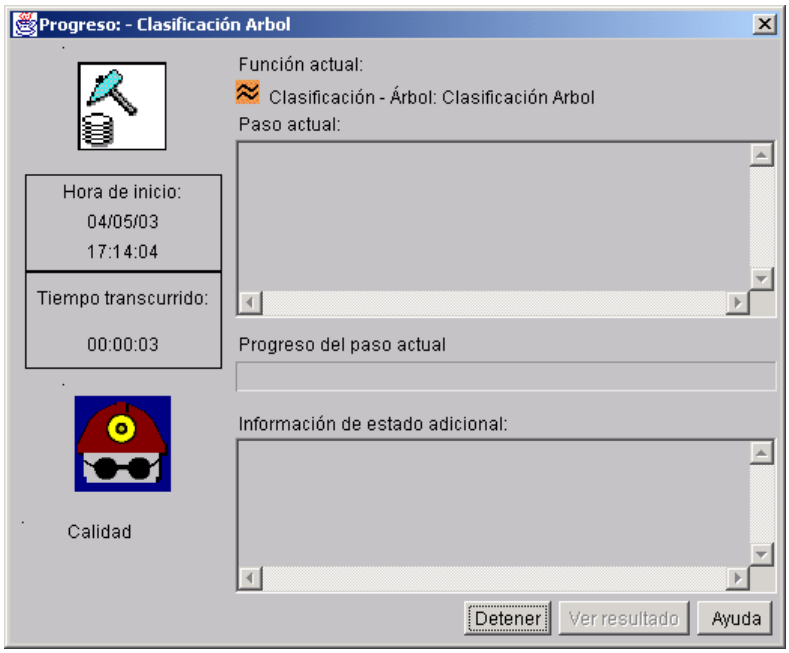


Figura 42. Avance del proceso de minería.

Los resultados del proceso son desplegados a continuación.

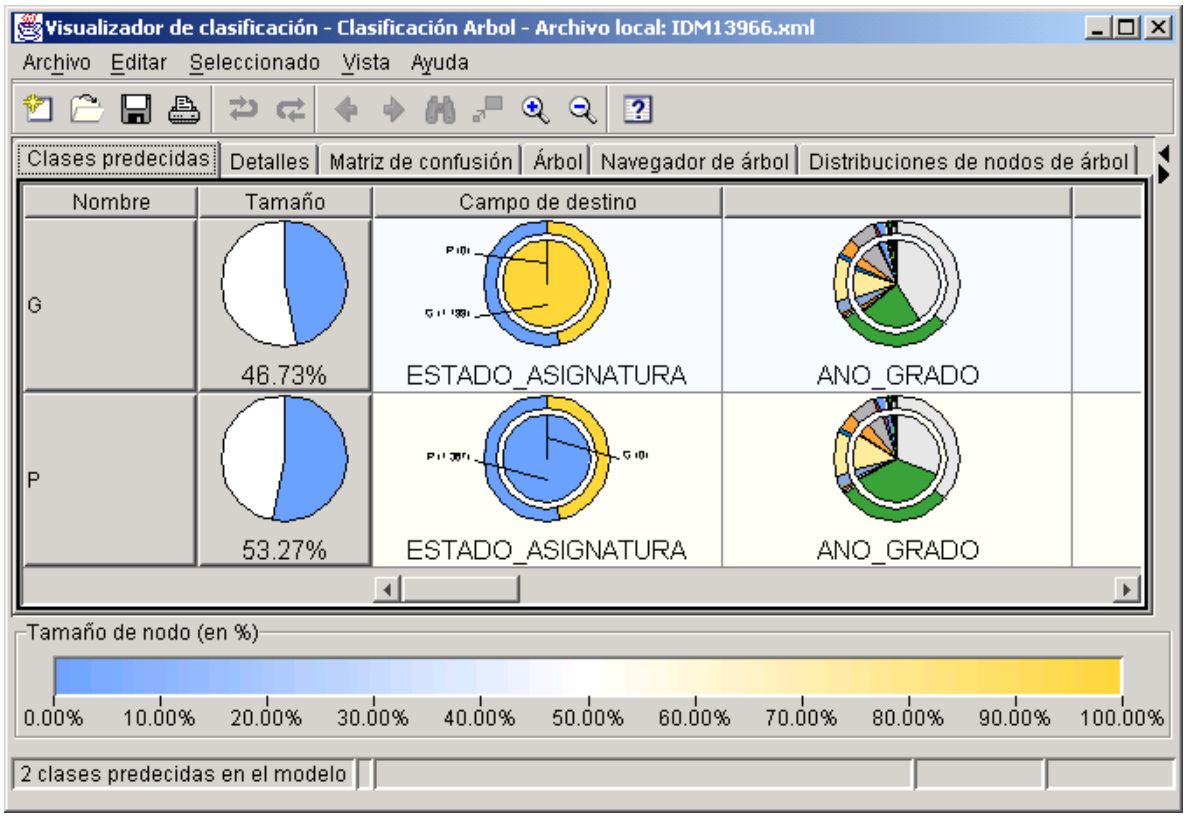


Figura 43. Despliegue de resultados.

6.3.4.5. ANÁLISIS DE RESULTADOS

En la figura anterior se puede notar que el 53.27% de la población evaluada perdió el curso de *Matemáticas I*, y sobre este subconjunto se evaluarán la distribución de las diferentes variables del modelo.

La clasificación de los estudiantes que perdieron el curso de *Matemáticas I* se presenta de la siguiente manera:

Figura 44. Distribución de variables para alumnos que perdieron el curso de Matemáticas I.

En el análisis básico de estos datos se encuentra que no existe una tendencia significativa que marque el comportamiento de los alumnos que pierden la

asignatura analizada en función de las variables descritas. Sin embargo entre las variables que muestran una leve diferencia de comportamiento, y que valdría la pena que se pusieran en estado de observación por parte de las autoridades académicas de la Universidad están las siguientes:

- El **año de grado** del estudiante en su colegio, ya que el análisis muestra que mientras que el 36% de la gente evaluada se graduó en 2001, el porcentaje de los estudiantes que perdieron la asignatura y que se graduaron en ese año fue del 41%. De igual forma se nota que mientras que el 29.8% de la población evaluada se graduó en el año 2000, el porcentaje de los que perdieron la asignatura y se graduaron en el año 2000 fue del 22.6%. Estas cifras pueden llevar a mostrar que existe una tendencia de relación inversa entre el año de graduación en el colegio y la pérdida de la asignatura de Matemáticas I, debida quizás a los menores niveles de exigencia que se presentan en los colegios inducidos por las reformas académicas y el sistema de logros que se aplica en los mismos.

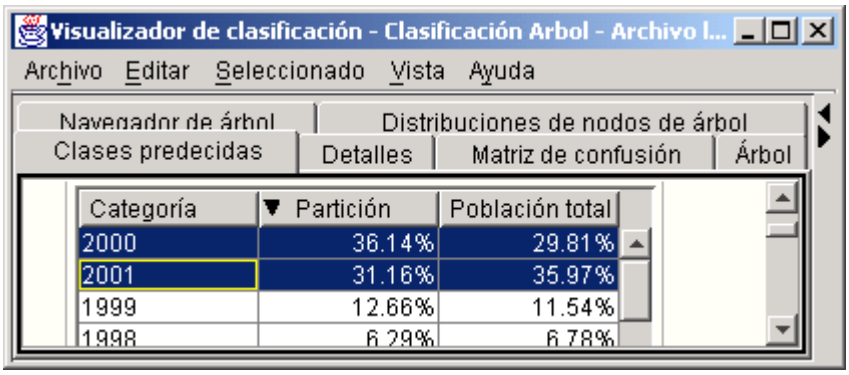


Figura 45. Distribución por año de grado.

- La **edad** de los estudiantes, ya que el análisis muestra que las edades que más aportan en la población son 17, 18, 19 y 20 en ese orden y se nota una relación inversa entre los estudiantes con edades de 18, 19 y 20; para los de 17 años la relación es al contrario, esto puede indicar que los alumnos de menor edad asimilan más fácilmente las propuestas metodológicas de la División de Ciencias Básicas.

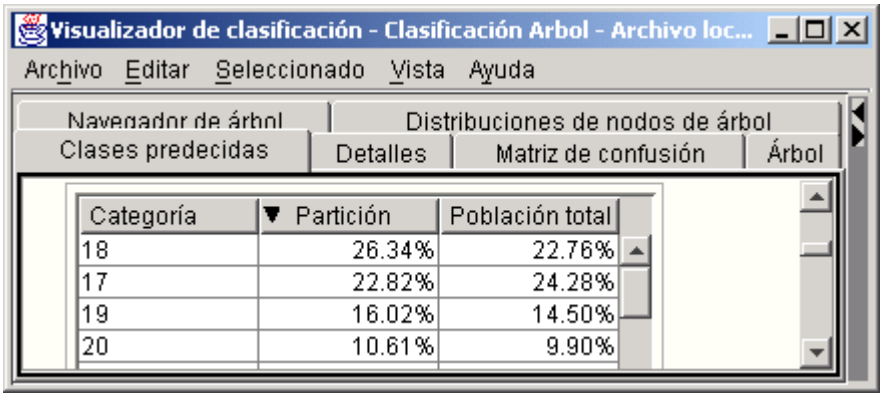
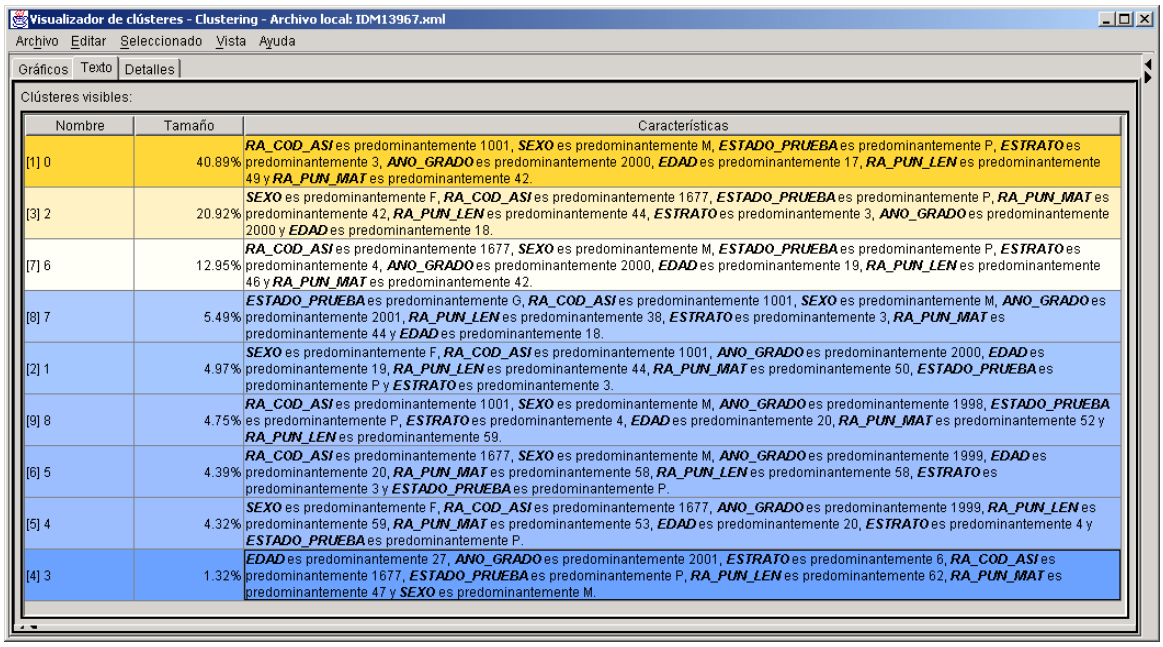


Figura 46. Distribución por edad.

- Con respecto al **resultado de la prueba diagnóstica** realizada por la División de Ciencias Básicas a los estudiantes del grupo, se encuentra que aunque el 86% de los estudiantes perdió dicha prueba, la asignatura fue finalmente reprobada por el 53%, lo que al parecer muestra que el resultado de dicha prueba ayuda a motivar a los estudiantes a superar sus deficiencias en el área de las matemáticas.

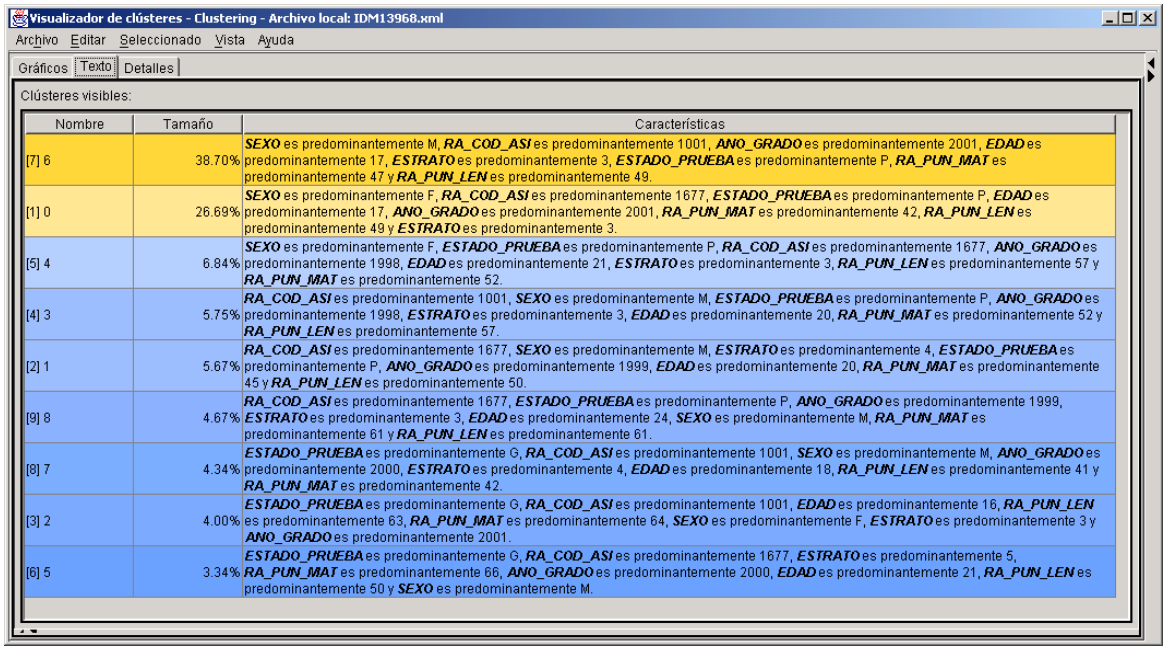
- El **resultado de la prueba de estado para Lenguaje y Aptitud verbal,** así como para **Matemáticas y Aptitud Matemática** muestran una tendencia normal en cuanto a que quienes obtuvieron mayores puntajes en dicha áreas no pierden la asignatura de *Matemáticas I*.

Se realizó una partición del universo de estudio en las dos categorías (estudiantes que aprobaron y estudiantes que reprobaron la asignatura *Matemáticas I*), y sobre cada una de esas particiones se ejecutó el modelo de minería de *agrupamiento* (*clustering*), encontrándose resultados muy similares en la confección de los patrones de agrupamiento, lo que corrobora la hipótesis de que no existe una correlación alta entre las variables del modelo y su incidencia en la mortalidad académica. A continuación se muestran los informes presentados por el software *Intelligent Miner*.



Nombre	Tamaño	Características
[1] 0	40.89%	***RA_COD_ASI*** es predominantemente 1001, ***SEXO*** es predominantemente M, ***ESTADO_PRUEBA*** es predominantemente P, ***ESTRATO*** es predominantemente 3, ***ANO_GRADO*** es predominantemente 2000, ***EDAD*** es predominantemente 17, ***RA_PUN_LEN*** es predominantemente 49 y ***RA_PUN_MAT*** es predominantemente 42.
[3] 2	20.92%	***SEXO*** es predominantemente F, ***RA_COD_ASI*** es predominantemente 1677, ***ESTADO_PRUEBA*** es predominantemente P, ***RA_PUN_MAT*** es predominantemente 42, ***RA_PUN_LEN*** es predominantemente 44, ***ESTRATO*** es predominantemente 3, ***ANO_GRADO*** es predominantemente 2000 y ***EDAD*** es predominantemente 18.
[7] 6	12.95%	***RA_COD_ASI*** es predominantemente 1677, ***SEXO*** es predominantemente M, ***ESTADO_PRUEBA*** es predominantemente P, ***ESTRATO*** es predominantemente 4, ***ANO_GRADO*** es predominantemente 2000, ***EDAD*** es predominantemente 19, ***RA_PUN_LEN*** es predominantemente 46 y ***RA_PUN_MAT*** es predominantemente 42.
[8] 7	5.49%	***ESTADO_PRUEBA*** es predominantemente G, ***RA_COD_ASI*** es predominantemente 1001, ***SEXO*** es predominantemente M, ***ANO_GRADO*** es predominantemente 2001, ***RA_PUN_LEN*** es predominantemente 38, ***ESTRATO*** es predominantemente 3, ***RA_PUN_MAT*** es predominantemente 44 y ***EDAD*** es predominantemente 18.
[2] 1	4.97%	***SEXO*** es predominantemente F, ***RA_COD_ASI*** es predominantemente 1001, ***ANO_GRADO*** es predominantemente 2000, ***EDAD*** es predominantemente 19, ***RA_PUN_LEN*** es predominantemente 44, ***RA_PUN_MAT*** es predominantemente 50, ***ESTADO_PRUEBA*** es predominantemente P y ***ESTRATO*** es predominantemente 3.
[9] 8	4.75%	***RA_COD_ASI*** es predominantemente 1001, ***SEXO*** es predominantemente M, ***ANO_GRADO*** es predominantemente 1998, ***ESTADO_PRUEBA*** es predominantemente P, ***ESTRATO*** es predominantemente 4, ***EDAD*** es predominantemente 20, ***RA_PUN_MAT*** es predominantemente 52 y ***RA_PUN_LEN*** es predominantemente 59.
[6] 5	4.39%	***RA_COD_ASI*** es predominantemente 1677, ***SEXO*** es predominantemente M, ***ANO_GRADO*** es predominantemente 1999, ***EDAD*** es predominantemente 20, ***RA_PUN_MAT*** es predominantemente 58, ***RA_PUN_LEN*** es predominantemente 58, ***ESTRATO*** es predominantemente 3 y ***ESTADO_PRUEBA*** es predominantemente P.
[5] 4	4.32%	***SEXO*** es predominantemente F, ***RA_COD_ASI*** es predominantemente 1677, ***ANO_GRADO*** es predominantemente 1999, ***RA_PUN_LEN*** es predominantemente 59, ***RA_PUN_MAT*** es predominantemente 53, ***EDAD*** es predominantemente 20, ***ESTRATO*** es predominantemente 4 y ***ESTADO_PRUEBA*** es predominantemente P.
[4] 3	1.32%	***EDAD*** es predominantemente 27, ***ANO_GRADO*** es predominantemente 2001, ***ESTRATO*** es predominantemente 6, ***RA_COD_ASI*** es predominantemente 1677, ***ESTADO_PRUEBA*** es predominantemente P, ***RA_PUN_LEN*** es predominantemente 62, ***RA_PUN_MAT*** es predominantemente 47 y ***SEXO*** es predominantemente M.

Figura 47. Análisis de agrupación para categoría de asignatura reprobada.



Nombre	Tamaño	Características
[7] 6	38.70%	***SEXO*** es predominantemente M, ***RA_COD_ASI*** es predominantemente 1001, ***ANO_GRADO*** es predominantemente 2001, ***EDAD*** es predominantemente 17, ***ESTRATO*** es predominantemente 3, ***ESTADO_PRUEBA*** es predominantemente P, ***RA_PUN_MAT*** es predominantemente 47 y ***RA_PUN_LEN*** es predominantemente 49.
[1] 0	26.69%	***SEXO*** es predominantemente F, ***RA_COD_ASI*** es predominantemente 1677, ***ESTADO_PRUEBA*** es predominantemente P, ***EDAD*** es predominantemente 17, ***ANO_GRADO*** es predominantemente 2001, ***RA_PUN_MAT*** es predominantemente 42, ***RA_PUN_LEN*** es predominantemente 49 y ***ESTRATO*** es predominantemente 3.
[5] 4	6.84%	***SEXO*** es predominantemente F, ***ESTADO_PRUEBA*** es predominantemente P, ***RA_COD_ASI*** es predominantemente 1677, ***ANO_GRADO*** es predominantemente 1998, ***EDAD*** es predominantemente 21, ***ESTRATO*** es predominantemente 3, ***RA_PUN_LEN*** es predominantemente 57 y ***RA_PUN_MAT*** es predominantemente 52.
[4] 3	5.75%	***RA_COD_ASI*** es predominantemente 1001, ***SEXO*** es predominantemente M, ***ESTADO_PRUEBA*** es predominantemente P, ***ANO_GRADO*** es predominantemente 1998, ***ESTRATO*** es predominantemente 3, ***EDAD*** es predominantemente 20, ***RA_PUN_MAT*** es predominantemente 52 y ***RA_PUN_LEN*** es predominantemente 57.
[2] 1	5.67%	***RA_COD_ASI*** es predominantemente 1677, ***SEXO*** es predominantemente M, ***ESTRATO*** es predominantemente 4, ***ESTADO_PRUEBA*** es predominantemente P, ***ANO_GRADO*** es predominantemente 1999, ***EDAD*** es predominantemente 20, ***RA_PUN_MAT*** es predominantemente 45 y ***RA_PUN_LEN*** es predominantemente 50.
[9] 8	4.67%	***RA_COD_ASI*** es predominantemente 1677, ***ESTADO_PRUEBA*** es predominantemente P, ***ANO_GRADO*** es predominantemente 1999, ***ESTRATO*** es predominantemente 3, ***EDAD*** es predominantemente 24, ***SEXO*** es predominantemente M, ***RA_PUN_MAT*** es predominantemente 61 y ***RA_PUN_LEN*** es predominantemente 61.
[8] 7	4.34%	***ESTADO_PRUEBA*** es predominantemente G, ***RA_COD_ASI*** es predominantemente 1001, ***SEXO*** es predominantemente M, ***ANO_GRADO*** es predominantemente 2000, ***ESTRATO*** es predominantemente 4, ***EDAD*** es predominantemente 18, ***RA_PUN_LEN*** es predominantemente 41 y ***RA_PUN_MAT*** es predominantemente 42.
[3] 2	4.00%	***ESTADO_PRUEBA*** es predominantemente G, ***RA_COD_ASI*** es predominantemente 1001, ***EDAD*** es predominantemente 16, ***RA_PUN_LEN*** es predominantemente 63, ***RA_PUN_MAT*** es predominantemente 64, ***SEXO*** es predominantemente F, ***ESTRATO*** es predominantemente 3 y ***ANO_GRADO*** es predominantemente 2001.
[6] 5	3.34%	***ESTADO_PRUEBA*** es predominantemente G, ***RA_COD_ASI*** es predominantemente 1677, ***ESTRATO*** es predominantemente 5, ***RA_PUN_MAT*** es predominantemente 66, ***ANO_GRADO*** es predominantemente 2000, ***EDAD*** es predominantemente 21, ***RA_PUN_LEN*** es predominantemente 50 y ***SEXO*** es predominantemente M.

Figura 48. Análisis de agrupación para categoría de asignatura aprobada

7. CONCLUSIONES

- Si bien es cierto que las herramientas tecnológicas evolucionan constantemente, y por lo tanto la facilidad de integración entre las herramientas para minería de datos y descubrimiento de conocimiento con las bases de datos o bodegas de datos donde está almacenada la información es cada vez mayor, sigue siendo importante que el proceso se realice en forma guiada, por lo que el concurso de los expertos en el tópico de estudio es de vital importancia para el éxito del proceso, ya que es el experto quién tiene claro tanto el problema específico como el entorno del mismo, e igualmente es él quién debe plantear los posibles cursos de acción a seguir.

- Las herramientas de minería de datos, aplicadas en el contexto de un problema específico, pueden arrojar resultados más completos que los obtenidos mediante la utilización de técnicas estadísticas clásicas, todo ello sin la necesidad de recurrir a un experto estadístico. Sigue siendo de utilidad

el contar con la participación de un experto en la temática propia del problema de estudio.

- La integración entre las herramientas de minería de datos y los manejadores de bases de datos permite automatizar las labores de carga y depuración de los datos para el modelo, así como la ejecución del mismo, de tal forma que los expertos temáticos se pueden centrar en el análisis de los resultados generados por la herramienta de minería de datos y no en los detalles matemáticos de la evaluación. Esto permite que se puedan aplicar las recomendaciones generadas por los expertos temáticos de una forma más ágil, por lo que se puede llegar a evaluar la efectividad de las mismas de manera más rápida.

- En el caso específico del análisis de la mortalidad académica para el curso de Matemáticas I se detectó que no existen patrones de comportamiento contundentes para el desempeño de un estudiante en el citado curso derivados a partir del análisis de las variables típicas que se manejan en las instituciones de educación superior, tales como edad, sexo, estrato socioeconómico o colegio de procedencia. Esto implica que se hace necesario involucrar nuevos elementos sobre el estudiante entre los que se podrían

incluir su situación económica, afectiva, social, etc o sobre el entorno académico como los docentes y sus características propias.

- El proceso elaborado para trabajar sobre la información académica de Estudiantes de la Corporación Universitaria Autónoma de Occidente en el curso de Matemáticas I puede ser fácilmente adaptado para incluir otros cursos, así como también nuevas variables de análisis, e incluso para trabajar con información de cualquier tipo de institución educativa o de grupos de instituciones a nivel local, regional o nacional.

BIBLIOGRAFIA

ALIBERAS, J. Didáctica de las Ciencias. Perspectivas Actuales. en Enseñanza de las Ciencias. 1989.

AMERICAN COUNCIL OF EDUCATION. Straight talk about college costs and prices. Report of the National Commission on the cost of Higher Education. Acenet. Washington D.C. 1998

ALBRECHT, D y ZIDERMAN, A. Funding mechanisms for Higher Education. WBDP, paper number 153. The World Bank. 1993

APODACA, P y GALLARRETA, L. Propuesta de diversos indicadores del acceso /demanda de estudios universitarios. En: Indicadores en la universidad,

información y decisiones. MEC /Consejo de Universidades. Fareso SA, Madrid. 1999

ARBOLEDA, Gonzalo, PICON, César. La Mortalidad y la Deserción estudiantil en EAFIT sus causas y posibles soluciones. Universidad EAFIT. Medellín, 1977.

BARNETSON, B. Key performance indicators in Higher Education. Alberta Colleges and Institutes Faculties Association. Alberta, Canadá. 1999

BATISTA, Enrique y otros. Mortalidad y Deserción Académicos en los programas de Pregrado de la Universidad de Antioquia. Universidad de Antioquia. Medellín, 1994.

BERG, D. University decision making and management infomation systems. Manuscrito no publicado. Presentación en el Seminario: Los desafíos del siglo XXI, Montevideo. 1997

BOJALIL, Luis Felipe. Diagnóstico y Prospectiva de la educación Superior en México. Editado por Unidad Xochimilco e Instituto de Investigaciones Legislativas. Programa de Superación Académica. Universidad Autónoma de México. México

CAVE, M. , HANNEY, S. , HENKEL, M. Y KOGAN, M. The use of Performance Indicators in Higher Education. A critical analysis of developing practice. Jessica Kingsley Publisher. 1998

CHARLTON, C. La evaluación institucional en Gran Bretaña. Manuscrito no publicado. Presentación en el Seminario: La evaluación universitaria. Montevideo. 1993

COLLAZOS, J. y GENSINI F. "La eficiencia del Sistema Universitario Colombiano". En: Revista Mundo Universitario. No.5, pp. 77-103. Bogotá, 1973.

COMMITTEE OF VICECHANCELLORS AND PRINCIPALS –CVCP. Report of the steering committee for efficiency studies in universities (JARRATT REPORT). William Lea Group, London. 1985

CONSEJO NACIONAL DE ACREDITACIÓN. Lineamientos para la Acreditación. Tercera Edición. Santafé de Bogotá, 1998.

CVCP/UFC. University management statistics and performance indicators, UK.1987

DE MIGUEL, M. La evaluación de la enseñanza. Propuesta de indicadores para las titulaciones. En: Indicadores en la universidad, información y decisiones. MEC /Consejo de universidades. Fareso SA, Madrid. 1999

DELORS, Jacques. Informe a la UNESCO de la Comisión Internacional sobre la Educación para el siglo XXI. La Educación encierra un tesoro. Madrid: Santillana. Ediciones UNESCO. 1996.

División de Investigaciones e Innovaciones Educativas. Deserción Escolar: Estrategias efectivas. Revista Educación. Número 56. Noviembre. Santiago de Chile. 1993.

DRIVER, R. Un enfoque constructivista para el desarrollo del currículo en ciencias. En Enseñanza de las Ciencias. 1988.

FAYYAD, Usama et al. Knowledge Discovery in Databases. AAAI/MIT Press, 1991.

FLOREZ y otros. "Deserción en los programas tecnológicos del SED 1983- 1986". Universidad de Antioquia. Medellín, 1987.

FURIO MAS, C. J. Tendencias actuales en la formación del profesorado de ciencias. Departamento de didáctica de las ciencias experimentales y sociales. Universidad de Valencia, España, 1994

GARCIA, T. La evaluación y el control de la eficiencia en la universidad. Tesis doctoral. Facultad de Ciencias Económicas y Empresariales. Universidad de Cádiz. España. 1993

GAITHER, G. Measuring up: the promise and pitfalls of performance indicators. ERIC Digest. 1995

GALLEGOS, G. Planeación de un proceso de autoevaluación. Manuscrito no publicado. Presentación en Taller de Coordinadores de Autoevaluación. Montevideo. 1999

GIL, D. Tres paradigmas básicos para la enseñanza de las ciencias. en: Enseñanza de las Ciencias, Vol. 1, 1. 1983

GINES MORA, J. Indicadores y decisiones en las universidades. En: Indicadores en la universidad, información y decisiones. MEC /Consejo de universidades. Fareso SA, Madrid. 1999

GINESTAR, A Costos educacionales para la gerencia universitaria. EDIUNC. Argentina. 1990

GOMEZ, Buendía Hernando. Educación la Agenda del Siglo XXI. Hacia un desarrollo Humano. Programa de Naciones Unidas para el desarrollo. PNUD. Santafé de Bogotá: TM Editores. 1998.

GRACIARENA, Jorge. "La Deserción y el Retraso en los Estudios Universitarios en Uruguay". En: América latina. No.13 V.1 Pp. 45-63. Río de Janeiro, 1970.

GRAO, J y WINTER, R. Indicadores para la calidad y calidad de los indicadores. En: Indicadores en la universidad, información y decisiones. MEC/ Consejo de universidades. Fareso SA, Madrid. 1999

GUTIERREZ, R. Sicología y aprendizaje de las ciencias: El Modelo de Ausubel. en Enseñanza de las Ciencias. Vol 5, 2. 1987

HAN, Jiawei y KAMBER, Micheline. Data Mining. Morgan Kaufmann.

HESESS. Request 97/24. Higher Education Funding Council for England (HEFCE). 1997

HEFCE. Study of comparative costs of first degree and sub degree provisions. HEFCE, UK. 1995

HEFCE. Funding method for teaching from 1998-1999. How HEFCE allocates its funds. HEFCE, UK. 1996

HEFCE. International comparisons of the cost of teaching in Higher Education. HEFCE, UK. 1997

HEFCE. Funding Higher Education in England. How HEFCE allocates its funds. HEFCE, UK. 1998

INTERNATIONAL BUSINESS MACHINES – IBM. Uso de Intelligent Miner for Data. Dinamarca, 2002.

JONES, J y TAYLOR, J. Performance indicators in Higher Education. SRHE, UK. 1990

KLUBITSCHKO, Doris. "El cambio de carrera como deserción interna en la Universidad Católica de Chile", en: Universidad Contemporánea un intento de análisis empírico. Corporación de Promoción Universitaria. Santiago de Chile, 1974. Pp. 275 -213.

KORTH, Henry F. Y SOLBERSCHATZ, Abraham. Fundamentos de bases de datos. McGraw-Hill, 1993.

LEVESQUE, K., BRADBY, D. Y ROSSI, K. Using data for program improvement: how do we encourage schools to do it? Centerfocus, number 12. 1996

LEWIS, D. Possibilities for expanding university internal efficiency. Manuscrito no publicado. Presentación en: Seminario los desafíos del siglo XXI. Montevideo. 1997

MACHUCA, Fernando. Optimization of Query Evaluation in Object Database Systems, Paris, 1995. Tesis (doctor en informática). Universidad de Versalles.

MALO ALVAREZ, S. Los indicadores en la evaluación de la educación superior. Un recurso para la toma de decisiones y la promoción de la calidad. UNAM. México. 1998

MARTIN, James y ODELL, James J. Análisis y diseño orientado a objetos. Prentice Hall Hispanoamericana, 1994.

MARTINEZ SANDRES, F. Sistemas de información y de evaluación universitaria. Aplicación y contribución en materia de educación superior. FCU-CSIC. 1999

McKEOWN, M. A view from the States. A survey of the collection and use of cost data by States. En: Cost measurement: public policy issues, options and strategies. IHEP, Washington D.C. 1999

MORENO, Cecilia y otros. "Factores que influyen en la elección de carrera profesional en alumnos de sexto de bachillerato de Medellín", en: Educación UPB. No 5. Universidad Pontificia Bolivariana. Medellín, 1980.

NACUBO. Congress addresses college costs as part of HEA reauthorizations. NACUBO. Washington D.C. 1996

NICHOLLS, J. Academic development and quality control. Manuscrito no publicado. Presentación en Seminario: Los cambios en la educación superior. Montevideo. 1992

NOVAK, J. D. Constructivismo Humano: un consenso emergente. en Enseñanza de las Ciencias. Vol. 6, 3. 1988

NOWACZYK, R y UNDERWOOD, D. Possible indicators of research quality for colleges and universities. En: Educational Policy Analysis Archives. Vol 3, number 20. 1995

PARAMO, Gabriel y CORREA, Carlos. "Deserción Estudiantil Universitaria. Conceptualización", en: Revista Universidad EAFIT. No.114. Medellín, 1999.

PEREZ LINDO, A. Evaluación del rendimiento de las universidades. En: Propuesta educativa. FLACSO Año 2, número 2. 1990

PEREZ, Teodoro. Proyectos pedagógicos para el cambio cultural: una aproximación conceptual y metodológica. En: Revista Avanzada. Número 3. 1998.

PERIE, Marianne, Z. Jing, R. Pearson y J. Sherman. U.S. Department of Education. International Education Indicators. A time Series Perspective. National Center for Education Statistics. Washington, DC. 1997.

PETREl, H. Costos de la educación universitaria en la Argentina. En: Ensayos en economía de la educación. Buenos Aires. 1989

RAMIREZ, Mariano. "Deserción estudiantil universitaria". En: Revista Mundo Universitario. No.6, pp. 9 -32. Bogotá, 1974.

RESTREPO, Luis Carlos. Ambientes educativos y estética Social. Intervención en Planteamiento de Planteamientos, Realizado en el Planetario Distrital, Santafé de Bogotá. 1993.

RICHARDS TORRES, Cecilia. Deserción escolar y circuito callejero. En: Revista Educación. No. 243. 1977.

ROLDAN, Ofelia, S. V, ALVARADO, C.M. Hincapié, E. Ocampo, J.E. Ramírez, M.R. Mejía y H.F Ospina. Educar el desafío de hoy. Construyendo posibilidades y alternativas. Santafé de Bogotá: Cooperativa Editorial Magisterio. 1999.

RODRIGUEZ A., Miguel A. Bases de datos. McGraw-Hill, 1992.

RODRIGUEZ ESPINAR, S. Información cualitativa y cuantitativa en el Plan Nacional de Evaluación. En: Indicadores en la universidad, información y decisiones. MECI Consejo de universidades. Fareso SA, Madrid. 1999

RUIZ, R. Evaluación académica y educación superior. En: Evaluar para transformar. IESALC UNESCO. 1999

SALAZAR, María C. MUÑOZ, Cecilia. "Aspectos de la deserción estudiantil en el Departamento de Sociología de la Universidad Nacional de Colombia", en: Revista América Latina. No.4, pp. 85-109. Río de Janeiro, 1968.

SARMIENTO, Doris y Patricia Giraldo: La Deserción académica en EAFIT y , sus causas. Medellín: Universidad EAFIT, 1989.

SHEFC. Management information for decision making: costing guidelines for HEI. Scotland. 1996

SMITH, T. Comparaciones internacionales sobre educación terciaria. En: Indicadores en la universidad, información y decisiones. MECI Consejo de universidades. Fareso SA. Madrid. 1999

SYVERSON, P y MAGUIRE, M. Estimating institutional cost of graduate education. Reports of three States demonstrate promise and pitfalls of cost studies. VCR. Council of Graduate Schools. 1997

TIBERGHIEN, A. Ideas Científicas en la Infancia y la Adolescencia. Ed. Morata, Madrid. 1985

THE JOINT COMMITTEE ON STANDARDS OF EDUCATIONAL EVALUATION. The program evaluation standards. Sage Publications Inc. 1994

TORRES Alvarez, Germán. Diagnóstico de la Educación superior. Postgrado en Gestión de la calidad Universitaria USB.

UNIVERSIDAD EAFIT. Documento de Archivo 1995 – 1998 (actas egresados No. 345/95 a 371/97; listados semestrales sobre estudiantes de pregrados retirados. Oficina de Admisiones y Registro y de Archivo.

UNIVERSIDAD EAFIT. Boletín Estadístico 1997.

UNIVERSIDAD EAFIT. Documento de Archivo 1995 – 1998 (actas egresados No. 345/95 a 371/97; listados semestrales sobre estudiantes de pregrados retirados. Oficina de Admisiones y Registro y de Archivo.

UNIVERSIDAD EAFIT. Boletín Estadístico 1997.

UNIVERSITY OF BRITISH COLUMBIA. Accountability in the UBC. PAIR-UBC, Canadá. 2000

UNIVERSITY OF MINNESOTA. Critical measures and performance goals. OVPAC, U of MN. Twin cities Minnesota. 1995

Universidad de los Andes. Oficina de Planeación. Indicadores de la población Pregrado. Santafé de Bogotá. 1998.

Universidad del Cuyo. Investigaciones Orientación Vocacional. Consejo de Investigaciones Científicas de Mendoza (CONICMEN). 1998.

URRUTIA Montoya, Miguel. Revista del Banco de la República. Inflación y Pensiones Escolares. Nota Editorial. En: Revista del Banco de la República.

VELEZ, Guillermo. Reflexiones introductorias sobre la deserción forzosa en la Universidad EAFIT, Dirección de Planeación Integral Universidad EAFIT. Medellín, 1986.

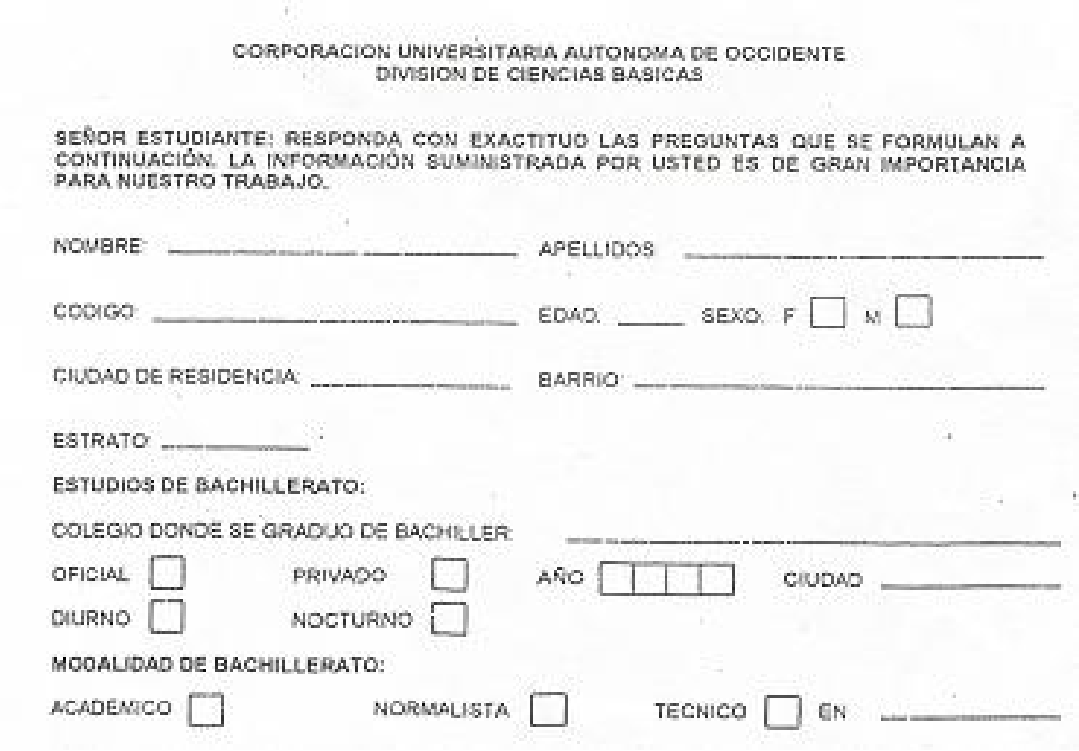
VELEZ, Guillermo, Mariano Ramírez y Jairo Rivera. Deserción Estudiantil Universitaria. Revista Mundo Universitario. Número 6. Medellín. 1974.

VIDAL, J. Indicadores de rendimiento para las universidades españolas: necesidades y disponibilidad. En: Indicadores en la universidad, Información y decisiones. MEC/ Consejo de Universidades. Fareso SA, Madrid. 1999

VILLAREAL M. La utilización de los indicadores de rendimiento en la financiación de la educación superior. En: Indicadores en la universidad, Información y decisiones. Fareso SA, Madrid. 1999

WITTEN, Ian H. y FRANK, Eibe. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, 1999.

ANEXO A. Formato de Recolección de Datos



CORPORACION UNIVERSITARIA AUTONOMA DE OCCIDENTE
DIVISION DE CIENCIAS BASICAS

SEÑOR ESTUDIANTE: RESPONDA CON EXACTITUD LAS PREGUNTAS QUE SE FORMULAN A CONTINUACIÓN. LA INFORMACIÓN SUMINISTRADA POR USTED ES DE GRAN IMPORTANCIA PARA NUESTRO TRABAJO.

NOMBRE: _____ APELLIDOS: _____

CODIGO: _____ EDAD: _____ SEXO: F ☐ M ☐

CIUDAD DE RESIDENCIA: _____ BARRIO: _____

ESTRATO: _____

ESTUDIOS DE BACHILLERATO:

COLEGIO DONDE SE GRADUO DE BACHILLER: _____

OFICIAL ☐ PRIVADO ☐ AÑO ☐☐☐☐ CIUDAD _____

DIURNO ☐ NOCTURNO ☐

MODALIDAD DE BACHILLERATO:

ACADEMICO ☐ NORMALISTA ☐ TECNICO ☐ EN _____

Fuente. División de Ciencias Básicas CUAO.

ANEXO B. Formato de Prueba Diagnóstica

PRUEBA DIAGNOSTICA 2003 A

Responda las siguientes 20 preguntas marcando únicamente en la hoja de respuestas la opción que usted considere correcta.

1. Dos Personas alquilaron un vehículo por un día. Una de ellas aportó 50.000 pesos y la otra aportó $\frac{3}{4}$ del alquiler total. El valor total del alquiler en pesos es:

A	B	C	D
250.000	200.000	150.000	100.000

2. El precio de entrada al cine pasó de 8.000 a 9.600 pesos. El porcentaje en que se incrementó el precio fue del:

A	B	C	D
100%	120%	96%	20%

3. El resultado de efectuar las operaciones indicadas en $1 - \left[\frac{1}{1+\frac{1}{2}}\right]$ es:

A	B	C	D
-1	3	$\frac{1}{3}$	$\frac{5}{3}$

4. La expresión $\left(2^2\right)^3 + \left(\frac{1}{2}\right)^3$ es equivalente a:

A	B	C	D
$2^6 + \frac{1}{8}$	$2^5 + \frac{3}{6}$	$2^{12} + \frac{1}{6}$	$2^8 + \frac{1}{8}$

5. De las siguientes expresiones indique cuál es la correcta.

A	B	C	D
$\sqrt{4+9}=2+3$	$3\sqrt{3}-\sqrt{2}=3$	$\sqrt{5+4}=3$	$2\sqrt{3}+\sqrt{3}=3\sqrt{6}$

6. Cuatro hombres hacen una obra en 12 días. ¿En cuántos días podrán hacer la misma obra seis hombres?

A	B	C	D
8	6	18	4

7.

Las partes sombreadas en el dibujo corresponden a:

A	B	C	D
3/3 del total	6/5 del total	3/10 del total	3/5 del total

8. El resultado de efectuar las operaciones indicadas en $-3+3\left[-2(1-2)-2(2-1)\right]$ es:

A	B	C	D
0	-3	-9	1

9. ¿Cuál de las siguientes opciones es equivalente a la expresión $(x^{-1}+x)^{-1}$?

A	B	C	D
$\frac{1+x^2}{x}$	$x+\frac{1}{x}$	$\frac{x}{1+x^2}$	$\frac{1}{1+x}$

10. Al despejar x de la ecuación $\frac{6x-3}{2}=1$ se obtiene:

A	B	C	D
$x=\frac{4}{3}$	$x=\frac{5}{6}$	$x=\frac{2}{3}$	$x=\frac{11}{3}$

11. Al simplificar la expresión $\frac{a^2-b^2}{a-b}$, donde $a-b$ es diferente de cero, el resultado es:

A	B	C	D
1	$a-b$	$b-a$	$a+b$

12. Al desarrollar la expresión $(x-y)^2+2xy$ obtenemos como resultado:

A	B	C	D
x^2+y^2	x^2-y^2	x^2+y^2+2xy	x^2-y^2+2xy

13. ¿Cuál de las siguientes expresiones debe sumarse a $5+8x+x^2$ para que dé como resultado x^2-7?

A	B	C	D
$2-8x$	$-12-8x$	$-2+8x+2x^2$	$12+8x$

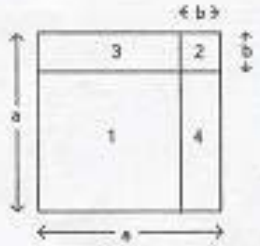
14. Si $p = x^3 - 3x - 2$, ¿Cuál es el valor de p cuando $x = -2$?

A	B	C	D
0	-2	3	-4

15. Sean a y b dos números tales que el cuadrado de su suma es 100 y uno de ellos el doble del otro. ¿Cuál de las siguientes opciones representa el problema?

A	B	C	D
$(a+b)^2 = 100$ $a = 2b$	$a^2 + b^2 = 100$ $a = 2b$	$(a+b)^2 = 100$ $a = b+2$	$(a+b)^2 = 100$ $a+b = 2$

Las preguntas 16 y 17 se deben responder de acuerdo a la figura que se muestra a continuación.



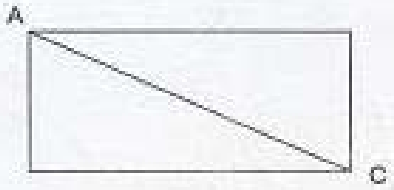
16. El área de la región demarcada con el número 1 es:

A	B	C	D
$a^2 - b^2$	$a^2 + b^2 - ab$	$a^2 + b^2 - 2ab$	$a^2 - b^2 + 2ab$

17. El valor de la diagonal del rectángulo demarcado con el número 4 es:

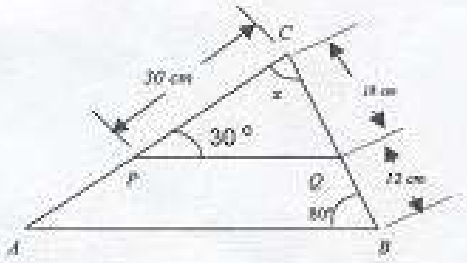
A	B	C	D
$\sqrt{a^2 + 2b^2 - 2ab}$	$a^2 + 2b^2 + 2ab$	$a^2 + 2b^2 - 2ab$	$\sqrt{a^2 + (a^2 - b^2)}$

19. El área de un rectángulo es de 12 cm^2 y su perímetro es de 14 cm, la longitud de la diagonal AC es:



A	B	C	D
7 cm	5 cm	52 cm	25 cm

Las preguntas 19 y 20 se deben responder de acuerdo a la figura que se muestra a continuación.



19. Si los segmentos AB y PQ son paralelos, la medida del ángulo x es:

A	B	C	D
30°	45°	50°	70°

20. Si los segmentos AB y PQ son paralelos, la longitud del segmento AP es:

A	B	C	D
20 cm	22 cm	30 cm	12 cm

Fuente. División de Ciencias Básicas – CUAO.

ANEXO C. Estudio Estadístico a los datos de prueba diagnóstica aplicada por la División de Ciencias Básicas a los estudiantes de los cursos de Matemáticas I de la CUAO

CORPORACIÓN UNIVERSITARIA AUTÓNOMA DE OCCIDENTE

APLICACIÓN DE TÉCNICAS MULTIVARIADAS EN EL PROCESO KDD A LOS DATOS DE LA PRUEBA DIAGNÓSTICA APLICADA A LOS ESTUDIANTES DE LOS CURSOS DE MATEMÁTICAS 1 DE LA CUAO.

M.g. VICTOR MANUEL GONZALEZ

CALI, OCTUBRE 29 DE 2002

APLICACIÓN DE TÉCNICAS MULTIVARIADAS EN EL PROCESO KDD.

Ha sido de interes reciente el desarrollo de temas de investigación como el de Descubrimiento de Conocimiento en Bases de Datos, KDD (knowledge Discovery in Databases), el cual es básicamente un proceso automatico que consiste en identificar patrones de interés novedosos y útiles en forma de reglas o funciones a partir de los datos para el respectivo análisis por parte de los usuarios. El proceso de descubrir conocimiento en bases de datos implica integrar en una herramienta, interrogaciones, análisis y métodos de visualización que le faciliten al usuario entender e interpretar los datos.

En el proceso de KDD, se siguen varias etapas dentro de las cuales encontramos las siguientes:

1. Comprensión de los objetivos y dominio de la aplicación.

2. Selección de los datos objetivo: Censo o Muestreo.

3. Validación de los datos: Remoción de ruido, Outliers y puntos de influencia, Tratamiento a datos faltantes.

4. Reducción de la dimensionalidad del espacio de las variables mediante métodos factoriales o técnicas *"fuzzy"* que permitan sintetizar grandes volúmenes de datos.

5. *Data mining* basadas en clasificación, regresión, agrupamiento etc, utiliza técnicas estadísticas clásicas (Análisis Discriminante, Cluster análisis, Regresión múltiple...) y otras como Conjuntos aproximados (*rough sets*), Redes neuronales, Árboles de decisión.

6. Interpretación y consolidación del conocimiento descubierto.

Observe que en el desarrollo de KDD, los pasos 2,3,4 y 5 requieren para su aplicación la intervención de profesionales formados especialmente en el área de estadística multivariada.

La minería de datos (*data mining*) es un paso en el proceso de KDD consistente en algoritmos de búsqueda en los datos que producen una enumeración particular de patrones encontrados. Las dos metas primarias de la minería de datos son la predicción y descripción.

Si el interés estadístico radica en establecer relaciones funcionales entre variables predictoras y predictivas pueden considerarse métodos como el Análisis de regresión múltiple, Análisis multivariado de varianza (MANOVA), Análisis discriminante, Modelos logit y probit, y Análisis de correlación canónica.

De otra parte, si el interés radica en reducir y describir las interrelaciones entre variables e individuos, los métodos factoriales como el Análisis de Componentes Principales (ACP), el Análisis de Correspondencias e inclusive el Análisis Cluster para agrupamiento son de mucha utilidad.

Cada uno de los métodos anteriores se implementa de acuerdo al tipo de variables: Métricas (cuantitativas) o no métricas (cualitativas).

La Corporación Universitaria Autónoma de Occidente CUAO, a través de la Escuela de Postgrados contribuye en estos procesos de investigación y formación de sus docentes mediante la realización de la maestría en " Ciencias computacionales con énfasis en redes ".

Para optar el titulo de maestría, se desarrollará la tesis "**Aplicación del descubrimiento de conocimiento en el análisis de mortalidad académica en la Ciencias Básicas** " la cual en su fase de aplicación toma como bases de datos principales las evaluaciones de la "prueba diagnostica" en conocimiento matemático realizada a los estudiantes que cursan matemáticas I en la CUAO correspondiente a los períodos I y II semestre del año 2000, I y II semestre del 2001 y I semestre 2002 (estudiantes de matemáticas II).

En este caso y como primera instancia se busca definir sobre la base de datos el modelo matemático y otros indicadores que permitan interpretar, describir o caracterizar el problema de deserción o mortalidad estudiantil en la CUAO. A partir de estos resultados se evaluarán patrones de comportamiento de los estudiantes como aporte a la búsqueda de nuevo conocimiento.

Sobre esta base, se desarrollarán los siguientes procedimientos:

- Análisis de Regresión Múltiple.

    1. Detección de Outliers y Puntos de influencia
    2. Corrección de Multicolinealidad
    3. Validación de supuestos

- Análisis de Componentes Principales.

    1. Reducción
    2. Descripción

ANALISIS DE RESULTADOS

Estadísticas Descriptivas

Obviamente el primer paso fue estructurar la base de datos conformada por 1608 estudiantes de la CUAO que presentaron las pruebas diagnosticas n° 2, 3, 4, 5 y que además retienen el 50.4%, 15%, 33.1%, y 1.5% respectivamente. Esto "garantiza" la calidad de los datos y evita *Outliers y Puntos de influencia* por errores de digitación.

La prueba n° 2 se realizó en el primer semestre del año 2000, la n° 3 en el segundo semestre del mismo año y asi sucesivamente. El objetivo de realizar esta prueba diagnostica a los estudiantes de los diferentes programas matriculados en matemáticas, es evaluar el nivel de conocimientos básicos tanto en esta área como en otras relacionadas tales como álgebra, geometría y gráficas. Posteriormente y de acuerdo con los procesos KDD se buscarán modelos estadisticos que expliquen las relaciones " existentes" en las estructuras de los datos.

Precisamente, al observar los resultados de la prueba diagnostica del total de estudiantes es evidente que solo en aritmética el 70.5% de lo estudiantes superan 3 o más aciertos , que los resultados en álgebra son regulares mientras que en geometría y gráficas los estudiantes se rajan. En esta última área el 87.2% de los evaluados obtienen 2 o menos aciertos.

Posesionados en este punto, revisamos algunas circunstancias como antecedentes y otras a posteriori buscando una explicación al resultado en cuestion. Se examinó el año de grado bachiller y se encontró que aproximadamente el 90% de los estudiantes poseen la característica " recien egresados" porque tenían menos de 4 años de graduados; igualmente se observó que el 74% de los estudiantes obtuvo nota baja en matemáticas-Icfes por debajo de 50 puntos.

Posteriormente, se evaluó el curso de matemáticas que reciben los estudiantes en los primeros semestres de su carrera profesional , encontrándose que solo el 40% de los estudiantes aprobaron el curso con nota superior a 3.0 mientras el 14.2% obtuvieron exactamente 3.0.

De otro lado, y considerando esta descripción como una radiografia rápida de las características más importantes, se considera que la población de estudiantes es *joven* pues casi todos (93.5%) presentan edades que oscilan entre los 16 y 24 años y el 87.5% viven en estratos socioeconómicos 3, 4 y 5. Es claro que el 80% de los estudiantes provienen de colegios privados.

Estadísticas Básicas Poblacion Estudiantil CUAO

Estadísticos

	N					
	Válidos	Perdidos	Media	Desv. típ.	Mínimo	Máximo
ALGEBRA	1608	0	2,44	1,48	0	7
AÑOGRADO	1542	66	1999,28	2,85	1978	2001
Aritmética	1608	0	3,33	1,57	0	8
CUAOMAT	1608	0	2,789	,973	,0	5,0
EDAD	1556	52	19,00	3,38	15	42
ESTRATO	1472	136	3,61	1,01	1	6
Geometria	1608	0	1,61	1,22	0	5
GRÁFICAS	1052	556	1,35	1,16	0	5
JORNADA	1608	0				
LAPSO	1542	66	1,7173	2,8520	,00	23,00
MATICFES	1608	0	46,697	7,788	28,0	75,5
NOTAPD	1608	0	2,063	,883	,0	5,0
PROGRAMA	1608	0	198,73	53,20	143	259
Prueba No	1608	0	2,86	,94	2	5
SEXO	1608	0				
Tipo Colegio	1608	0				
Bachillerato	1608	0				

Tabla de frecuencia ALGEBRA

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	0	124	7,7	7,7	7,7
	1	352	21,9	21,9	29,6
	2	404	25,1	25,1	54,7
	3	364	22,6	22,6	77,4
	4	206	12,8	12,8	90,2
	5	115	7,2	7,2	97,3
	6	36	2,2	2,2	99,6
	7	7	,4	,4	100,0
	Total	1608	100,0	100,0	
Total		1608	100,0		

Tabla de frecuencia AÑOGRADO

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	1978	1	,1	,1	,1
	1980	1	,1	,1	,1
	1981	1	,1	,1	,2
	1983	3	,2	,2	,4
	1984	2	,1	,1	,5
	1985	2	,1	,1	,6
	1986	7	,4	,5	1,1
	1987	3	,2	,2	1,3
	1988	6	,4	,4	1,7
	1989	8	,5	,5	2,2
	1990	6	,4	,4	2,6
	1991	9	,6	,6	3,2
	1992	11	,7	,7	3,9
	1993	15	,9	1,0	4,9
	1994	23	1,4	1,5	6,4
	1995	22	1,4	1,4	7,8
	1996	40	2,5	2,6	10,4
	1997	56	3,5	3,6	14,0
	1998	106	6,6	6,9	20,9
	1999	171	10,6	11,1	32,0
	2000	391	24,3	25,4	57,3
	2001	658	40,9	42,7	100,0
	Total	1542	95,9	100,0	
Perdidos	Perdidos del sistema	66	4,1		
	Total	66	4,1		
Total		1608	100,0		

Tabla de frecuencia Aritmética

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	0	60	3,7	3,7	3,7
	1	143	8,9	8,9	12,6
	2	272	16,9	16,9	29,5
	3	392	24,4	24,4	53,9
	4	374	23,3	23,3	77,2
	5	266	16,5	16,5	93,7
	6	59	3,7	3,7	97,4
	7	29	1,8	1,8	99,2
	8	13	,8	,8	100,0
	Total	1608	100,0	100,0	
Total		1608	100,0		

Tabla de frecuencia CUAOMAT

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	,0	10	,6	,6	,6
	,1	4	,2	,2	,9
	,2	2	,1	,1	1,0
	,3	2	,1	,1	1,1
	,4	9	,6	,6	1,7
	,5	9	,6	,6	2,2
	,6	11	,7	,7	2,9
	,7	8	,5	,5	3,4
	,8	14	,9	,9	4,3
	,9	10	,6	,6	4,9
	1,0	17	1,1	1,1	6,0
	1,1	14	,9	,9	6,8
	1,2	18	1,1	1,1	8,0
	1,3	13	,8	,8	8,8
	1,4	17	1,1	1,1	9,8
	1,5	24	1,5	1,5	11,3
	1,6	26	1,6	1,6	12,9
	1,7	25	1,6	1,6	14,5
	1,8	25	1,6	1,6	16,0
	1,9	29	1,8	1,8	17,8
	2,0	50	3,1	3,1	21,0
	2,1	45	2,8	2,8	23,8
	2,2	37	2,3	2,3	26,1
	2,3	41	2,5	2,5	28,6
	2,4	37	2,3	2,3	30,9
	2,5	113	7,0	7,0	37,9
	2,6	58	3,6	3,6	41,5
	2,7	33	2,1	2,1	43,6
	2,8	28	1,7	1,7	45,3
	2,9	5	,3	,3	45,6
	3,0	229	14,2	14,2	59,9
	3,1	91	5,7	5,7	65,5
	3,2	53	3,3	3,3	68,8
	3,3	69	4,3	4,3	73,1
	3,4	55	3,4	3,4	76,6
	3,5	59	3,7	3,7	80,2
	3,6	40	2,5	2,5	82,7
	3,7	37	2,3	2,3	85,0
	3,8	45	2,8	2,8	87,8
	3,9	31	1,9	1,9	89,7
	4,0	28	1,7	1,7	91,5
	4,1	28	1,7	1,7	93,2
	4,2	18	1,1	1,1	94,3
	4,3	17	1,1	1,1	95,4
	4,4	17	1,1	1,1	96,5
	4,5	8	,5	,5	97,0
	4,6	14	,9	,9	97,8

Tabla de frecuencia CUAOMAT

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	4,7	11	,7	,7	98,5
	4,8	7	,4	,4	98,9
	4,9	6	,4	,4	99,3
	5,0	11	,7	,7	100,0
	Total	1608	100,0	100,0	
Total		1608	100,0		

Tabla de frecuencia EDAD

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	15	14	,9	,9	,9
	16	174	10,8	11,2	12,1
	17	412	25,6	26,5	38,6
	18	326	20,3	21,0	59,5
	19	201	12,5	12,9	72,4
	20	141	8,8	9,1	81,5
	21	78	4,9	5,0	86,5
	22	52	3,2	3,3	89,8
	23	35	2,2	2,2	92,1
	24	36	2,2	2,3	94,4
	25	11	,7	,7	95,1
	26	11	,7	,7	95,8
	27	9	,6	,6	96,4
	28	5	,3	,3	96,7
	29	6	,4	,4	97,1
	30	7	,4	,4	97,6
	31	10	,6	,6	98,2
	32	6	,4	,4	98,6
	33	8	,5	,5	99,1
	34	4	,2	,3	99,4
	35	3	,2	,2	99,6
	36	2	,1	,1	99,7
	37	1	,1	,1	99,7
	38	1	,1	,1	99,8
	39	2	,1	,1	99,9
	42	1	,1	,1	100,0
	Total	1556	96,8	100,0	
Perdidos	Perdidos del sistema	52	3,2		
	Total	52	3,2		
Total		1608	100,0		

Tabla de frecuencia ESTRATO

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	1	7	,4	,5	,5
	2	125	7,8	8,5	9,0
	3	686	42,7	46,6	55,6
	4	326	20,3	22,1	77,7
	5	277	17,2	18,8	96,5
	6	51	3,2	3,5	100,0
	Total	1472	91,5	100,0	
Perdidos	Perdidos del sistema	136	8,5		
	Total	136	8,5		
Total		1608	100,0		

Tabla de frecuencia Geometría

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	0	316	19,7	19,7	19,7
	1	498	31,0	31,0	50,6
	2	452	28,1	28,1	78,7
	3	220	13,7	13,7	92,4
	4	91	5,7	5,7	98,1
	5	31	1,9	1,9	100,0
	Total	1608	100,0	100,0	
Total		1608	100,0		

Tabla de frecuencia GRÁFICAS

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	0	254	15,8	24,1	24,1
	1	396	24,6	37,6	61,8
	2	267	16,6	25,4	87,2
	3	66	4,1	6,3	93,4
	4	48	3,0	4,6	98,0
	5	21	1,3	2,0	100,0
	Total	1052	65,4	100,0	
Perdidos	Perdidos del sistema	556	34,6		
	Total	556	34,6		
Total		1608	100,0		

Tabla de frecuencia JORNADA

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	D	1198	74,5	74,5	74,5
	N	410	25,5	25,5	100,0
	Total	1608	100,0	100,0	
Total		1608	100,0		

Tabla de frecuencia LAPSO

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	,00	658	40,9	42,7	42,7
	1,00	391	24,3	25,4	68,0
	2,00	171	10,6	11,1	79,1
	3,00	106	6,6	6,9	86,0
	4,00	56	3,5	3,6	89,6
	5,00	40	2,5	2,6	92,2
	6,00	22	1,4	1,4	93,6
	7,00	23	1,4	1,5	95,1
	8,00	15	,9	1,0	96,1
	9,00	11	,7	,7	96,8
	10,00	9	,6	,6	97,4
	11,00	6	,4	,4	97,8
	12,00	8	,5	,5	98,3
	13,00	6	,4	,4	98,7
	14,00	3	,2	,2	98,9
	15,00	7	,4	,5	99,4
	16,00	2	,1	,1	99,5
	17,00	2	,1	,1	99,6
	18,00	3	,2	,2	99,8
	20,00	1	,1	,1	99,9
	21,00	1	,1	,1	99,9
	23,00	1	,1	,1	100,0
	Total	1542	95,9	100,0	
Perdidos	Perdidos del sistema	66	4,1		
	Total	66	4,1		
Total		1608	100,0		

Tabla de frecuencia MATICFES

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	28,0	2	,1	,1	,1
	29,0	5	,3	,3	,4
	30,0	1	,1	,1	,5
	31,0	9	,6	,6	1,1
	32,0	4	,2	,2	1,3
	33,0	13	,8	,8	2,1

Tabla de frecuencia MATICFES

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	34,0	1	,1	,1	2,2
	34,5	2	,1	,1	2,3
	35,0	24	1,5	1,5	3,8
	35,5	1	,1	,1	3,9
	36,0	24	1,5	1,5	5,3
	36,5	2	,1	,1	5,5
	37,0	8	,5	,5	6,0
	37,5	2	,1	,1	6,1
	38,0	57	3,5	3,5	9,6
	38,5	2	,1	,1	9,8
	39,0	121	7,5	7,5	17,3
	39,5	3	,2	,2	17,5
	40,0	14	,9	,9	18,3
	40,5	2	,1	,1	18,5
	41,0	114	7,1	7,1	25,6
	41,5	5	,3	,3	25,9
	42,0	162	10,1	10,1	35,9
	42,5	5	,3	,3	36,3
	43,0	12	,7	,7	37,0
	43,5	5	,3	,3	37,3
	44,0	132	8,2	8,2	45,5
	44,5	2	,1	,1	45,6
	45,0	124	7,7	7,7	53,4
	45,5	10	,6	,6	54,0
	46,0	12	,7	,7	54,7
	46,5	3	,2	,2	54,9
	47,0	116	7,2	7,2	62,1
	47,5	3	,2	,2	62,3
	48,0	34	2,1	2,1	64,4
	48,5	5	,3	,3	64,7
	49,0	58	3,6	3,6	68,3
	49,5	13	,8	,8	69,2
	50,0	78	4,9	4,9	74,0
	50,5	9	,6	,6	74,6
	51,0	20	1,2	1,2	75,8
	51,5	4	,2	,2	76,1
	52,0	63	3,9	3,9	80,0
	52,5	10	,6	,6	80,6
	53,0	18	1,1	1,1	81,7
	53,5	11	,7	,7	82,4
	54,0	40	2,5	2,5	84,9
	54,5	6	,4	,4	85,3
	55,0	17	1,1	1,1	86,3
	55,5	7	,4	,4	86,8
	56,0	27	1,7	1,7	88,4
	56,5	9	,6	,6	89,0
	57,0	16	1,0	1,0	90,0

Tabla de frecuencia MATICFES

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	57,5	10	,6	,6	90,6
	58,0	14	,9	,9	91,5
	58,5	6	,4	,4	91,9
	59,0	5	,3	,3	92,2
	59,5	7	,4	,4	92,6
	60,0	9	,6	,6	93,2
	60,5	12	,7	,7	93,9
	61,0	8	,5	,5	94,4
	61,5	4	,2	,2	94,7
	62,0	10	,6	,6	95,3
	62,5	4	,2	,2	95,5
	63,0	7	,4	,4	96,0
	63,5	8	,5	,5	96,5
	64,0	6	,4	,4	96,8
	64,5	6	,4	,4	97,2
	65,0	4	,2	,2	97,5
	65,5	2	,1	,1	97,6
	66,0	2	,1	,1	97,7
	66,5	6	,4	,4	98,1
	67,0	5	,3	,3	98,4
	67,5	3	,2	,2	98,6
	68,0	2	,1	,1	98,7
	69,0	2	,1	,1	98,8
	69,5	1	,1	,1	98,9
	70,0	4	,2	,2	99,1
	70,5	3	,2	,2	99,3
	71,0	1	,1	,1	99,4
	71,5	3	,2	,2	99,6
	72,5	1	,1	,1	99,6
	73,0	1	,1	,1	99,7
	73,5	1	,1	,1	99,8
	74,5	2	,1	,1	99,9
	75,0	1	,1	,1	99,9
	75,5	1	,1	,1	100,0
	Total	1608	100,0	100,0	
Total		1608	100,0		

Tabla de frecuencia NOTAPD

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	,0	1	,1	,1	,1
	,3	9	,6	,6	,6
	,5	31	1,9	1,9	2,5
	,8	61	3,8	3,8	6,3
	1,0	107	6,7	6,7	13,0
	1,3	139	8,6	8,6	21,6
	1,5	182	11,3	11,3	33,0
	1,8	212	13,2	13,2	46,1
	1,8	1	,1	,1	46,2
	2,0	193	12,0	12,0	58,2
	2,3	171	10,6	10,6	68,8
	2,5	147	9,1	9,1	78,0
	2,8	82	5,1	5,1	83,1
	3,0	78	4,9	4,9	87,9
	3,3	55	3,4	3,4	91,4
	3,5	42	2,6	2,6	94,0
	3,8	29	1,8	1,8	95,8
	4,0	23	1,4	1,4	97,2
	4,3	17	1,1	1,1	98,3
	4,5	12	,7	,7	99,0
	4,8	13	,8	,8	99,8
	5,0	3	,2	,2	100,0
	Total	1608	100,0	100,0	
Total		1608	100,0		

Tabla de frecuencia PROGRAMA

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	143	124	7,7	7,7	7,7
	144	168	10,4	10,4	18,2
	145	26	1,6	1,6	19,8
	146	143	8,9	8,9	28,7
	147	3	,2	,2	28,9
	148	119	7,4	7,4	36,3
	149	120	7,5	7,5	43,7
	150	15	,9	,9	44,7
	151	28	1,7	1,7	46,4
	152	72	4,5	4,5	50,9
	241	89	5,5	5,5	56,4
	248	3	,2	,2	56,6
	249	3	,2	,2	56,8
	250	16	1,0	1,0	57,8
	251	38	2,4	2,4	60,1
	252	118	7,3	7,3	67,5
	253	243	15,1	15,1	82,6
	254	51	3,2	3,2	85,8
	256	49	3,0	3,0	88,8
	257	67	4,2	4,2	93,0
	258	43	2,7	2,7	95,6
	259	70	4,4	4,4	100,0
	Total	1608	100,0	100,0	
Total		1608	100,0		

Tabla de frecuencia Prueba No

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	2	811	50,4	50,4	50,4
	3	241	15,0	15,0	65,4
	4	532	33,1	33,1	98,5
	5	24	1,5	1,5	100,0
	Total	1608	100,0	100,0	
Total		1608	100,0		

Tabla de frecuencia SEXO

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos		11	,7	,7	,7
	F	589	36,6	36,6	37,3
	M	1008	62,7	62,7	100,0
	Total	1608	100,0	100,0	
Total		1608	100,0		

Tabla de frecuencia Tipo Colegio

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos		51	3,2	3,2	3,2
	O	271	16,9	16,9	20,0
	P	1286	80,0	80,0	100,0
	Total	1608	100,0	100,0	
Total		1608	100,0		

Tabla de frecuencia Bachillerato

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos		73	4,5	4,5	4,5
	A	1070	66,5	66,5	71,1
	N	15	,9	,9	72,0
	T	450	28,0	28,0	100,0
	Total	1608	100,0	100,0	
Total		1608	100,0		

Regresión Múltiple

De acuerdo con el desarrollo reciente de temas como KDD, es de suma importancia descubrir relaciones funcionales en que permitan entender la estructura o patrones dentro del sistema de información que expliquen especialmente la relación deserción – rendimiento matemático. Para lograr esto se plantea un modelo de regresión múltiple de la forma

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$, donde los componentes aleatorios son Y, ε . además,

Y: Nota curso matemáticas CUAOMAT es la variable dependiente; y las variables independientes
X_1: Puntaje matemáticas MATICFES
X_2: Nota prueba diagnostica, NOTAPD
X_3: Tipo colegio, TICOL (1=Oficial, 0=Privado)

Se busca asi, explicar la nota obtenida en la materia de matemáticas recibida en la universidad CUAO a partir del puntaje en matemáticas Icfes, de la prueba diagnostica y de la variable dummy tipo de colegio.

Una vez se corre el modelo para la población total de estudiantes (pruebas 2,3,4 ,5) se observa que la mayor correlación 0.416 se da entre CUAOMAT y NOTAPD. La variable TICOL no es significativa ya que la prueba t=-0.678 asociada a la hipótesis H_0: β_3=0 no se rechaza ya que su nivel de significancia es de 0.498, orientado a que la variable TICOL no aporta al modelo.

Respecto a los diagnosticos de multicolinealidad ni los factores de inflación de varianza FIV (aproximadamene iguales a 1) ni los índices de condición (menores a 20) sugieren la presencia de esta patología tal que afecten la estimación de los β^s .Sin embargo, el $R^2 = 0.191$ inicialmente indica que el modelo desarrollado en la población no explica buena parte de la variación total en Y.

A partir de aquí, y en un intento por mejorar el modelo, se segmentó el archivo de datos por n° de prueba y se corrio en cada uno de estos items sin que se tuviera un incremento significativo en el coeficiente de determinación R^2. Luego se involucrarón otras variables explicativas como estrato y clase de bachillerato sin que se experimentara mejoría alguna.

Finalmente, se opto por correr el modelo solo con estudiantes de ingeniería obteniéndose un $R^2 = 0.258$; este fue el mejor modelo linel obtenido : $\hat{Y} = 0.93 + 0.017 MATICFES + 0.468 NOTAPD$ .Es claro en esta etapa que bajo las características estudiadas *no existe un mejor modelo lineal* ni si quiera de tipo cuadrático cuyos resultados fuerón menos esperanzadores.

Regresión Poblacion Total, Pruebas 2, 3, 4, 5.

Estadísticos descriptivos

	Media	Desviación típ.	N
CUAOMAT	2,791	,969	1557
MATICFES	46,659	7,788	1557
NOTAPD	2,070	,885	1557
TICOL	,8259	,3793	1557

Correlaciones

		CUAOMAT	MATICFES	NOTAPD	TICOL
Correlación de Pearson	CUAOMAT	1,000	,214	,416	,003
	MATICFES	,214	1,000	,204	-,042
	NOTAPD	,416	,204	1,000	,062
	TICOL	,003	-,042	,062	1,000
Sig. (unilateral)	CUAOMAT	.	,000	,000	,454
	MATICFES	,000	.	,000	,048
	NOTAPD	,000	,000	.	,007
	TICOL	,454	,048	,007	.
N	CUAOMAT	1557	1557	1557	1557
	MATICFES	1557	1557	1557	1557
	NOTAPD	1557	1557	1557	1557
	TICOL	1557	1557	1557	1557

Variables introducidas/eliminadas[b]

Modelo	Variables introducidas	Variables eliminadas	Método
1	TICOL, MATICFES, NOTAPD[a]	.	Introducir

a. Todas las variables solicitadas introducidas

b. Variable dependiente: CUAOMAT

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	,437[a]	,191	,189	,872	1,804

a. Variables predictoras: (Constante), TICOL, MATICFES, NOTAPD

b. Variable dependiente: CUAOMAT

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	278,589	3	92,863	122,000	,000[a]
	Residual	1182,096	1553	,761		
	Total	1460,685	1556			

a. Variables predictoras: (Constante), TICOL, MATICFES, NOTAPD

b. Variable dependiente: CUAOMAT

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Estadísticos de colinealidad	
		B	Error típ.	Beta			Tolerancia	FIV
1	(Constante)	1,165	,146		7,963	,000		
	MATICFES	1,663E-02	,003	,134	5,725	,000	,956	1,047
	NOTAPD	,427	,026	,390	16,676	,000	,954	1,049
	TICOL	-3,968E-02	,059	-,016	-,678	,498	,993	1,007

a. Variable dependiente: CUAOMAT

Diagnósticos de colinealidad[a]

Modelo	Dimensión	Autovalor	Indice de condición	Proporciones de la varianza			
				(Constante)	MATICFES	NOTAPD	TICOL
1	1	3,738	1,000	,00	,00	,01	,01
	2	,157	4,874	,00	,00	,32	,70
	3	9,125E-02	6,401	,05	,07	,67	,23
	4	1,292E-02	17,012	,95	,92	,00	,06

a. Variable dependiente: CUAOMAT

Estadísticos sobre los residuos[a]

	Mínimo	Máximo	Media	Desviación típ.	N
Valor pronosticado	1,864	4,506	2,791	,423	1557
Residual	-3,283	2,176	-4,066E-15	,872	1557
Valor pronosticado tip.	-2,192	4,052	,000	1,000	1557
Residuo tip.	-3,763	2,495	,000	,999	1557

a. Variable dependiente: CUAOMAT

Regresión Ingenieria (Pruebas 2,3,4,5)

Estadísticos descriptivos

	Media	Desviación típ.	N
CUAOMAT	2,778	,976	804
MATICFES	47,099	7,974	804
NOTAPD	2,305	,956	804
TICOL	,7811	,4138	804

Correlaciones

		CUAOMAT	MATICFES	NOTAPD	TICOL
Correlación de Pearson	CUAOMAT	1,000	,248	,488	,028
	MATICFES	,248	1,000	,228	-,013
	NOTAPD	,488	,228	1,000	,128
	TICOL	,028	-,013	,128	1,000
Sig. (unilateral)	CUAOMAT	.	,000	,000	,211
	MATICFES	,000	.	,000	,354
	NOTAPD	,000	,000	.	,000
	TICOL	,211	,354	,000	.
N	CUAOMAT	804	804	804	804
	MATICFES	804	804	804	804
	NOTAPD	804	804	804	804
	TICOL	804	804	804	804

Variables introducidas/eliminadas[b]

Modelo	Variables introducidas	Variables eliminadas	Método
1	TICOL, MATICFES, NOTAPD[a]	.	Introducir

a. Todas las variables solicitadas introducidas

b. Variable dependiente: CUAOMAT

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,508[a]	,258	,255	,842

a. Variables predictoras: (Constante), TICOL, MATICFES, NOTAPD

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	197,335	3	65,778	92,774	,000[a]
	Residual	567,211	800	,709		
	Total	764,546	803			

a. Variables predictoras: (Constante), TICOL, MATICFES, NOTAPD

b. Variable dependiente: CUAOMAT

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Estadísticos de colinealidad	
		B	Error típ.	Beta			Tolerancia	FIV
1	(Constante)	,930	,189		4,919	,000		
	MATICFES	1,745E-02	,004	,143	4,555	,000	,946	1,057
	NOTAPD	,468	,032	,459	14,530	,000	,931	1,074
	TICOL	-6,697E-02	,072	-,026	-,924	,356	,982	1,019

a. Variable dependiente: CUAOMAT

Diagnósticos de colinealidad[a]

Modelo	Dimensión	Autovalor	Indice de condición	Proporciones de la varianza			
				(Constante)	MATICFES	NOTAPD	TICOL
1	1	3,716	1,000	,00	,00	,01	,01
	2	,177	4,583	,00	,01	,13	,90
	3	9,337E-02	6,309	,05	,06	,86	,05
	4	1,355E-02	16,562	,94	,93	,00	,04

a. Variable dependiente: CUAOMAT

Análisis en Componentes Principales ACP

En esta fase se intenta describir todas las interrelaciones entre individuos, entre variables y entre variables e individuos en espacios de menor dimensión.

La estructura resultante del ACP aplicado a la población total es muy similar a la obtenida en cada prueba 2, 3 o4. Por ello se adjunta tambien el ACP al segmento conformado por la prueba n°3.

Luego de filtrarsen los individuos que no respondieron (asumieron) algunas variables,el archivo total finamente quedo conformado por 1417 estudiantes y por las variables activas Nota prueba diagnostica NPD, Aritmética ARITM, Cuso de matemáticas CUAO y matemáticas icfes MATI. Como variables suplementarias que ayudarán a la interpretación se tomaron los programas académicos (143=1, 144=2,.....259=19), Tipos de colegio (Ofic – Priv) y Bachillerato (Acad – Técn); estos tipos solo se aplicarón en la prueba n°3 .

En la población total al observar los valores propios mayores que el promedio (λ_1=2.0569 >1) y bajo los criterios de varianza explicada (51.42%), se debe elegir el primer eje para el análisis. Sin embargo los eje segundo 2° y tercero 3° aportan el 22.27% y 18.14% de la varianza respectivamente. En este momento ya se insinúa la no presencia de estructuras relacionales importantes, ya que es fácil intuir que se requieren 3 ejes para describir 4 variables.

En efecto, si observamos las correlaciones (0.86 , 0.81) entre las variables y los factores encontramos que las pruebas diagnosticas NPD y ARITM caracterizan el primer eje.Análogamente MATI con 0.86 califica el segundo eje; el tercer eje esta descrito por el curso de matemáticas CUAO cuya correlación asociada es 0.73. Note que esta "misma " estructura se mantiene en las componentes de los vectores propios o antiguos ejes unitarios como trazadores de los vectores directores de las respectivas variables.

En la representación del primer plano factorial no se observan características importantes con los programas académicos, excepto el 147, 150 y 249 que deben ser analizados con cierta reserva ya que su frecuencia absoluta representa solo 3, 15 y 3 estudiantes respectivamente. Los programas 251, 253,... presentan en promedio notas relativamente bajas en MATI, NPD y ARITM, mientras que programas como el 144, 146, 148,.. presentan las mejores notas en las pruebas diagnosticas. Asi mismo, los estudiantes que conforman los programas de estudio 152, 252, ...presentarón las mejores pruebas de matemáticas en el Icfes.

Finalmente, en el plano F1, F3 se puede apreciar que programas como el 241, 254, 257, ... obtuvieron buenas notas en la asignatura institucional de matemáticas CUAOM.

El análisis de Componentes Principales en el segmento dado por la prueba n°3 , se puede desarrollar en forma anloga al relizado para la población total. Tres cosas vamos a destacar en este espacio.

La primera es que tambien se requieren prácticamente el mismo número de ejes, tres, para describir la variación de las cuatro variables activas. La segunda es que prácticamente son indistinguibles tanto los tipos de colegio Ofic. y Priv. Como los bachilleratos Tecn y Acad. La última característica se presenta como ayuda a la interpretación en el sentido de que por ejemplo los estudiantes con codigo 63, 142,... presentan notas altas en las pruebas diagnosticas ARITMETI y NOTAPD oponiendo a los estudiantes 47, 54, 100 que obtuvieron notas bajas en estos mismos test.

Nota: La contribucion y calidad (Coseno^2) de estos puntos es alta en el primer eje !

ANALYSE EN COMPOSANTES PRINCIPALES

STATISTIQUES SOMMAIRES DES VARIABLES CONTINUES
EFFECTIF TOTAL : 1417 POIDS TOTAL : 1417.00

NUM . IDEN - LIBELLE	EFFECTIF	POIDS	MOYENNE	ECART-TYPE	MINIMUM	MAXIMUM
2 . NOTA - NPD	1417	1417.00	2.08	0.88	0.25	5.00
5 . ARIT - ARITM	1364	1364.00	3.47	1.46	1.00	8.00
6 . CUAO - CUAOM	1408	1408.00	2.81	0.95	0.10	5.00
7 . MATI - ICFESm	1417	1417.00	46.58	7.82	28.00	75.50

MATRICE DES CORRELATIONS

	NOTA	ARIT	CUAO	MATI
NOTA	1.00			
ARIT	0.65	1.00		
CUAO	0.43	0.29	1.00	
MATI	0.22	0.20	0.21	1.00
	NOTA	ARIT	CUAO	MATI

MATRICE DES VALEURS-TESTS

	NOTA	ARIT	CUAO	MATI
NOTA	99.99			
ARIT	28.91	99.99		
CUAO	17.20	11.23	99.99	
MATI	8.30	7.31	8.04	99.99
	NOTA	ARIT	CUAO	MATI

VALEURS PROPRES
APERCU DE LA PRECISION DES CALCULS : TRACE AVANT DIAGONALISATION .. 4.0000
SOMME DES VALEURS PROPRES 4.0000

HISTOGRAMME DES 4 PREMIERES VALEURS PROPRES

NUMERO	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE	
1	2.0569	51.42	51.42	**********
2	0.8907	22.27	73.69	**********
3	0.7255	18.14	91.83	**********
4	0.3268	8.17	100.00	**********

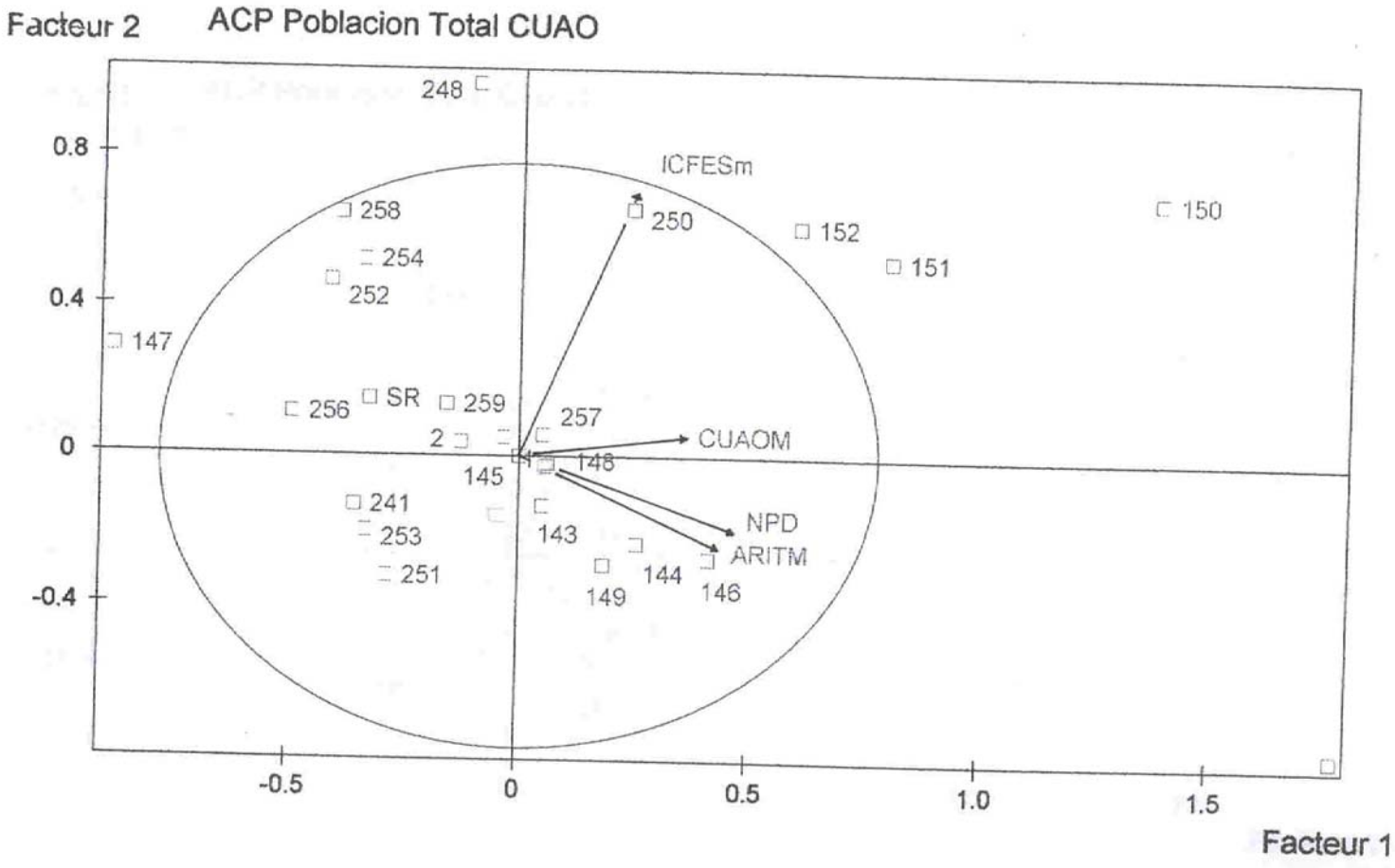
INTERVALLES LAPLACIENS D'ANDERSON
INTERVALLES AU SEUIL 0.95

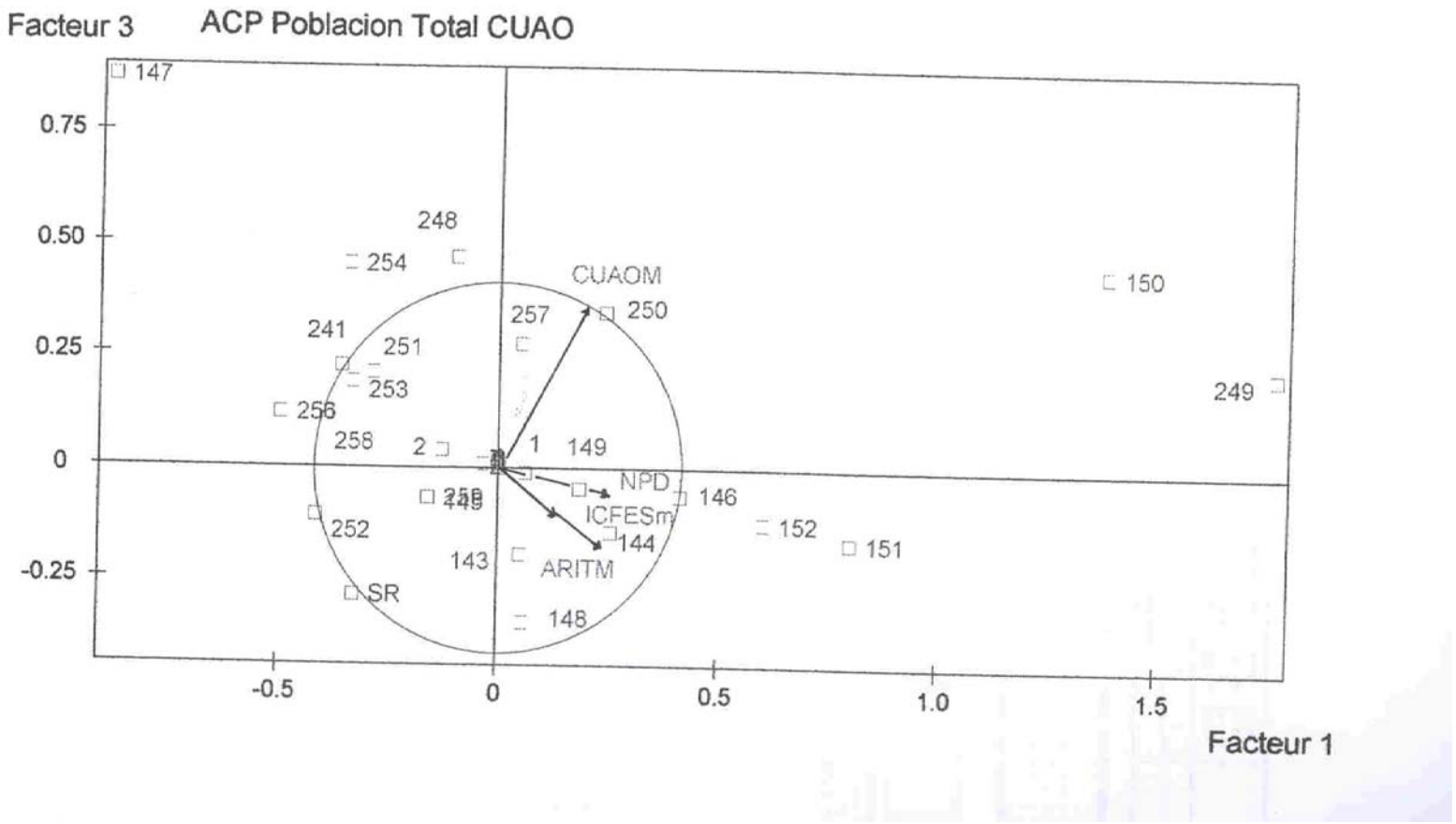
NUMERO	BORNE INFERIEURE	VALEUR PROPRE	BORNE SUPERIEURE
1	1.9054	2.0569	2.2084
2	0.8251	0.8907	0.9564
3	0.6721	0.7255	0.7790
4	0.3027	0.3268	0.3509

ETENDUE ET POSITION RELATIVE DES INTERVALLES

COORDONNEES DES VARIABLES SUR LES AXES 1 A 4
VARIABLES ACTIVES

VARIABLES	COORDONNEES					CORRELATIONS VARIABLE-FACTEUR					ANCIENS AXES UNITAIRES				
IDEN - LIBELLE COURT	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0
NOTA - NPD	0.86	-0.24	-0.11	0.43	0.00	0.86	-0.24	-0.11	0.43	0.00	0.60	-0.26	-0.13	0.74	0.00
ARIT - ARITM	0.81	-0.29	-0.36	-0.37	0.00	0.81	-0.29	-0.36	-0.37	0.00	0.56	-0.31	-0.42	-0.64	0.00
CUAO - CUAOM	0.67	0.07	0.73	-0.11	0.00	0.67	0.07	0.73	-0.11	0.00	0.47	0.08	0.86	-0.20	0.00
MATI - ICFESm	0.46	0.86	-0.22	0.00	0.00	0.46	0.86	-0.22	0.00	0.00	0.32	0.91	-0.25	0.01	0.00


Facteur 2
ACP Poblacion Total CUAO
248
0.8
ICFESm
258
250
150
152
254
151
0.4
252
147
256
SR
259
257
0
2
CUAOM
148
145
241
143
NPD
253
ARITM
144
251
149
146
-0.4
-0.5
0
0.5
1.0
1.5
Facteur 1


Facteur 3
ACP Poblacion Total CUAO
147
0.75
0.50
0.25
0
-0.25
248
254
CUAOM
257
250
241
251
253
256
258
2
1
149
NPD
146
ICFESm
144
252
143
ARITM
SR
148
150
249
152
151
-0.5
0
0.5
1.0
1.5
Facteur 1

ANALYSE EN COMPOSANTES PRINCIPALES Prueba nº 3
STATISTIQUES SOMMAIRES DES VARIABLES CONTINUES
EFFECTIF TOTAL : 209 POIDS TOTAL : 209.00

NUM . IDEN - LIBELLE	EFFECTIF	POIDS	MOYENNE	ECART-TYPE	MINIMUM	MAXIMUM
1 . NOTA - NOTAPD	209	209.00	1.97	0.85	0.50	4.75
2 . ARIT - ARITMÉTI	197	197.00	2.96	1.15	1.00	5.00
3 . CUAO - CUAOMAT	206	206.00	2.66	0.82	0.10	4.40
4 . MATI - Micfes	209	209.00	45.11	6.72	28.00	70.00

MATRICE DES CORRELATIONS

	NOTA	ARIT	CUAO	MATI
NOTA	1.00			
ARIT	0.69	1.00		
CUAO	0.11	0.04	1.00	
MATI	-0.01	0.06	-0.02	1.00

MATRICE DES VALEURS-TESTS

	NOTA	ARIT	CUAO	MATI
NOTA	99.99			
ARIT	11.80	99.99		
CUAO	1.55	0.61	99.99	
MATI	-0.20	0.86	-0.30	99.99

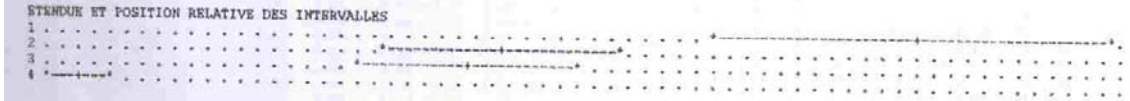
VALEURS PROPRES
APERCU DE LA PRECISION DES CALCULS : TRACE AVANT DIAGONALISATION .. 4.0000
SOMME DES VALEURS PROPRES 4.0000
HISTOGRAMME DES 4 PREMIERES VALEURS PROPRES

NUMERO	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE	
1	1.7035	42.59	42.59	**********
2	1.0250	25.62	68.21	**********
3	0.9644	24.11	92.32	**********
4	0.3071	7.68	100.00	**********

INTERVALLES LAPLACIENS D'ANDERSON
INTERVALLES AU SEUIL 0.95

NUMERO	BORNE INFERIEURE	VALEUR PROPRE	BORNE SUPERIEURE
1	1.3761	1.7035	2.0309
2	0.8280	1.0250	1.2220
3	0.7791	0.9644	1.1498
4	0.2481	0.3071	0.3661

ETENDUE ET POSITION RELATIVE DES INTERVALLES



COORDONNEES DES VARIABLES SUR LES AXES 1 A 4
VARIABLES ACTIVES

VARIABLES	COORDONNEES					CORRELATIONS VARIABLE-FACTEUR					ANCIENS AXES UNITAIRES				
IDEN - LIBELLE COURT	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0
NOTA - NOTAPD	-0.91	-0.03	0.10	0.39	0.00	-0.91	-0.03	0.10	0.39	0.00	-0.70	-0.03	0.10	0.71	0.00
ARIT - ARITMÉTI	-0.91	0.11	0.10	-0.39	0.00	-0.91	0.11	0.10	-0.39	0.00	-0.70	0.11	0.10	-0.70	0.00
CUAO - CUAOMAT	-0.19	-0.61	-0.77	-0.03	0.00	-0.19	-0.61	-0.77	-0.03	0.00	-0.15	-0.60	-0.78	-0.06	0.00
MATI - Micfes	-0.06	0.80	-0.59	0.04	0.00	-0.06	0.80	-0.59	0.04	0.00	-0.04	0.79	-0.61	0.07	0.00

COORDONNEES, CONTRIBUTIONS ET COSINUS CARRES DES INDIVIDUS
AXES 1 A 4

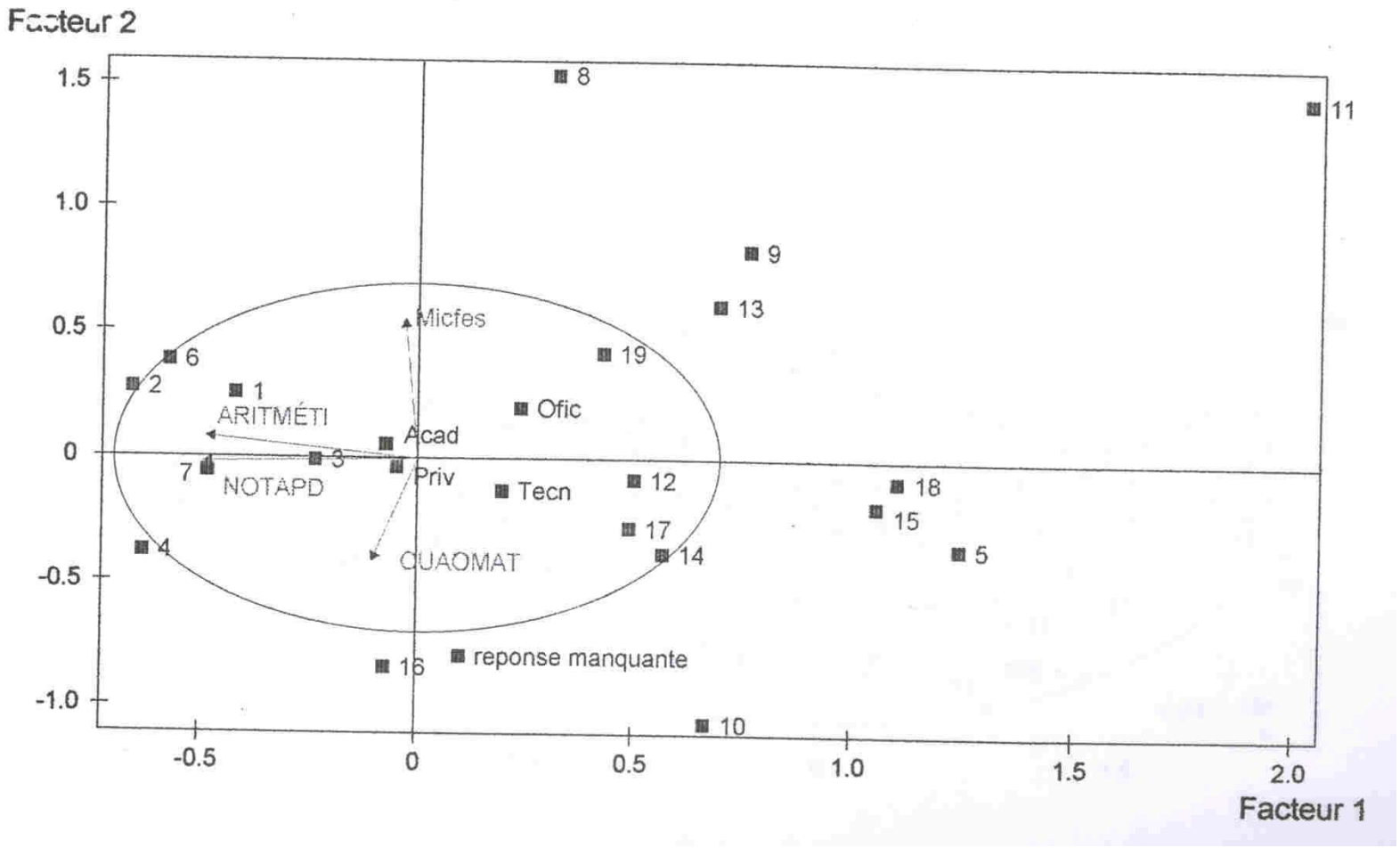
INDIVIDUS			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDENTIFICATEUR	P.REL	DISTO	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0
1	0.48	13.24	2.46	2.41	-1.15	0.26	0.00	1.7	2.7	0.7	0.1	0.0	0.46	0.44	0.10	0.01	0.00
2	0.48	9.76	1.00	-0.34	2.87	-0.61	0.00	0.3	0.1	4.1	0.6	0.0	0.10	0.01	0.85	0.04	0.00
3	0.48	1.22	0.83	-0.45	0.41	0.39	0.00	0.2	0.1	0.1	0.2	0.0	0.57	0.17	0.14	0.12	0.00
4	0.48	1.26	-0.72	-0.51	0.06	-0.69	0.00	0.1	0.1	0.0	0.7	0.0	0.41	0.21	0.00	0.38	0.00
5	0.48	4.17	0.38	1.89	-0.63	-0.23	0.00	0.0	1.7	0.2	0.1	0.0	0.04	0.86	0.10	0.01	0.00
6	0.48	3.22	1.44	-0.20	0.05	1.05	0.00	0.6	0.0	0.0	1.7	0.0	0.64	0.01	0.00	0.34	0.00
7	0.48	1.48	0.13	1.06	-0.58	-0.11	0.00	0.0	0.5	0.2	0.0	0.0	0.01	0.76	0.23	0.01	0.00
8	0.48	3.54	1.63	-0.48	0.00	0.82	0.00	0.7	0.1	0.0	1.0	0.0	0.75	0.07	0.00	0.19	0.00
9	0.48	0.27	0.10	-0.37	-0.24	-0.24	0.00	0.0	0.1	0.0	0.1	0.0	0.04	0.51	0.23	0.23	0.00
10	0.48	3.49	1.56	0.87	0.55	-0.10	0.00	0.7	0.4	0.1	0.0	0.0	0.70	0.22	0.09	0.00	0.00
11	0.48	2.85	1.64	-0.06	0.02	-0.40	0.00	0.8	0.0	0.0	0.3	0.0	0.94	0.00	0.00	0.06	0.00
12	0.48	3.34	-1.65	-0.14	-0.75	0.17	0.00	0.8	0.0	0.3	0.0	0.0	0.82	0.01	0.17	0.01	0.00
13	0.48	3.83	-1.80	0.38	-0.52	0.44	0.00	0.9	0.1	0.1	0.3	0.0	0.88	0.04	0.07	0.05	0.00
14	0.48	2.58	0.75	-1.05	0.21	-0.93	0.00	0.2	0.5	0.0	1.4	0.0	0.22	0.43	0.02	0.34	0.00
15	0.48	2.15	1.14	0.60	0.64	0.29	0.00	0.4	0.2	0.2	0.1	0.0	0.60	0.17	0.19	0.04	0.00
16	0.48	0.63	0.15	0.71	-0.31	-0.14	0.00	0.0	0.2	0.0	0.0	0.0	0.04	0.78	0.15	0.03	0.00
17	0.48	2.20	1.33	-0.32	-0.52	-0.22	0.00	0.5	0.0	0.1	0.1	0.0	0.81	0.05	0.12	0.02	0.00
18	0.48	1.66	-0.21	0.34	-1.22	0.03	0.00	0.0	0.1	0.7	0.0	0.0	0.03	0.07	0.90	0.00	0.00
19	0.48	0.94	0.52	-0.47	-0.31	0.59	0.00	0.1	0.1	0.0	0.5	0.0	0.29	0.24	0.10	0.37	0.00
20	0.48	7.37	1.45	1.89	0.79	-1.03	0.00	0.6	1.7	0.3	1.7	0.0	0.29	0.48	0.09	0.14	0.00
21	0.48	4.04	1.39	-1.38	0.30	-0.32	0.00	0.5	0.9	0.0	0.2	0.0	0.48	0.47	0.02	0.03	0.00
22	0.48	5.80	0.28	-1.85	-1.28	-0.81	0.00	0.0	1.6	0.8	1.0	0.0	0.01	0.59	0.28	0.11	0.00
23	0.48	2.41	-0.03	-0.62	0.53	-1.32	0.00	0.0	0.2	0.1	2.7	0.0	0.00	0.16	0.12	0.72	0.00
24	0.48	4.02	1.58	-0.34	1.17	-0.22	0.00	0.7	0.1	0.7	0.1	0.0	0.62	0.03	0.34	0.01	0.00
25	0.48	1.79	0.75	0.22	0.93	-0.57	0.00	0.2	0.0	0.4	0.5	0.0	0.31	0.03	0.48	0.18	0.00
26	0.48	0.68	-0.16	-0.60	-0.50	-0.06	0.00	0.0	0.2	0.1	0.0	0.0	0.04	0.56	0.39	0.01	0.00
27	0.48	5.82	2.15	-0.66	0.79	0.39	0.00	1.3	0.2	0.3	0.2	0.0	0.79	0.07	0.11	0.03	0.00
28	0.48	10.01	1.75	-2.27	-1.32	0.22	0.00	0.9	2.4	0.9	0.1	0.0	0.31	0.52	0.17	0.00	0.00
29	0.48	1.07	0.40	-0.42	0.28	0.81	0.00	0.0	0.1	0.0	1.0	0.0	0.15	0.17	0.07	0.61	0.00
30	0.48	4.27	1.23	0.53	1.25	-0.95	0.00	0.4	0.1	0.8	1.4	0.0	0.36	0.07	0.37	0.21	0.00
31	0.48	1.11	-0.28	-0.97	0.26	0.11	0.00	0.0	0.4	0.0	0.0	0.0	0.07	0.85	0.06	0.01	0.00
32	0.48	1.25	0.40	0.11	-0.76	0.65	0.00	0.1	0.0	0.3	0.7	0.0	0.19	0.01	0.46	0.34	0.00
33	0.48	6.88	2.33	-0.98	-0.68	-0.07	0.00	1.5	0.4	0.2	0.0	0.0	0.79	0.14	0.07	0.00	0.00
34	0.48	2.01	-0.82	0.21	-0.95	-0.62	0.00	0.2	0.0	0.4	0.6	0.0	0.34	0.02	0.45	0.19	0.00
35	0.48	1.26	1.01	-0.40	0.20	0.19	0.00	0.3	0.1	0.0	0.1	0.0	0.82	0.13	0.03	0.03	0.00
36	0.48	1.37	-0.92	-0.52	0.09	-0.48	0.00	0.2	0.1	0.0	0.4	0.0	0.62	0.20	0.01	0.17	0.00
37	0.48	3.70	0.56	-1.44	-1.11	0.29	0.00	0.1	1.0	0.6	0.1	0.0	0.08	0.56	0.34	0.02	0.00
38	0.48	8.09	-1.87	0.18	1.97	0.84	0.00	1.0	0.0	1.9	1.1	0.0	0.43	0.00	0.48	0.09	0.00
39	0.48	7.60	-2.49	-1.09	-0.28	-0.35	0.00	1.7	0.6	0.0	0.2	0.0	0.82	0.16	0.01	0.02	0.00
40	0.48	2.68	0.89	-1.34	-0.30	0.09	0.00	0.2	0.8	0.0	0.0	0.0	0.29	0.67	0.03	0.00	0.00
41	0.48	0.90	0.20	-0.66	0.59	-0.27	0.00	0.0	0.2	0.2	0.1	0.0	0.05	0.49	0.39	0.08	0.00
42	0.48	2.89	-0.27	-1.35	-0.99	-0.14	0.00	0.0	0.8	0.5	0.0	0.0	0.03	0.63	0.34	0.01	0.00
43	0.48	2.67	0.15	0.14	-1.57	-0.41	0.00	0.0	0.0	1.2	0.3	0.0	0.01	0.01	0.92	0.06	0.00
44	0.48	14.70	-1.04	3.02	-2.11	-0.15	0.00	0.3	4.3	2.2	0.0	0.0	0.07	0.62	0.30	0.00	0.00
45	0.48	3.50	0.97	-1.59	0.35	0.07	0.00	0.3	1.2	0.1	0.0	0.0	0.26	0.70	0.03	0.00	0.00
46	0.48	1.15	-0.94	-0.17	-0.18	-0.45	0.00	0.2	0.0	0.0	0.3	0.0	0.77	0.03	0.03	0.17	0.00
47	0.48	4.18	1.85	-0.61	0.16	0.60	0.00	1.0	0.2	0.0	0.6	0.0	0.82	0.09	0.01	0.09	0.00
48	0.48	8.21	0.18	0.75	2.74	0.30	0.00	0.0	0.3	3.7	0.1	0.0	0.00	0.07	0.92	0.01	0.00
49	0.48	0.61	0.12	-0.72	0.03	-0.28	0.00	0.0	0.2	0.0	0.1	0.0	0.02	0.85	0.00	0.13	0.00
50	0.48	4.16	-0.76	-1.83	0.20	0.45	0.00	0.2	1.6	0.0	0.3	0.0	0.14	0.80	0.01	0.05	0.00
51	0.48	7.63	0.85	-2.20	-0.28	1.26	0.00	0.2	2.4	0.0	2.5	0.0	0.10	0.69	0.01	0.21	0.00
52	0.48	0.27	0.10	-0.37	-0.24	-0.24	0.00	0.0	0.1	0.0	0.1	0.0	0.04	0.51	0.23	0.23	0.00
53	0.48	0.32	-0.12	-0.45	-0.31	-0.04	0.00	0.0	0.1	0.0	0.0	0.0	0.05	0.68	0.31	0.01	0.00
54	0.48	5.57	1.98	-0.52	1.01	0.61	0.00	1.1	0.1	0.5	0.6	0.0	0.70	0.05	0.18	0.07	0.00
55	0.48	1.02	0.31	-0.37	-0.37	0.81	0.00	0.0	0.1	0.1	1.0	0.0	0.09	0.13	0.13	0.64	0.00
56	0.48	2.00	-0.23	-0.63	0.56	-1.11	0.00	0.0	0.2	0.2	1.9	0.0	0.03	0.20	0.16	0.62	0.00
57	0.48	4.59	1.33	-1.37	-0.27	0.94	0.00	0.5	0.9	0.0	1.4	0.0	0.38	0.41	0.02	0.19	0.00
58	0.48	3.55	-0.58	-1.78	-0.01	0.24	0.00	0.1	1.5	0.0	0.1	0.0	0.09	0.89	0.00	0.02	0.00
59	0.48	2.01	0.83	0.87	-0.21	-0.72	0.00	0.2	0.4	0.0	0.8	0.0	0.34	0.38	0.02	0.26	0.00
60	0.48	1.65	0.42	0.02	-1.04	-0.63	0.00	0.1	0.0	0.5	0.6	0.0	0.11	0.00	0.65	0.24	0.00
61	0.48	0.62	0.29	-0.43	-0.37	-0.46	0.00	0.0	0.1	0.1	0.3	0.0	0.13	0.30	0.22	0.34	0.00
62	0.48	1.62	0.10	-1.22	0.11	-0.33	0.00	0.0	0.7	0.0	0.2	0.0	0.01	0.92	0.01	0.07	0.00
63	0.48	9.60	-2.95	0.82	0.14	0.47	0.00	2.4	0.3	0.0	0.3	0.0	0.90	0.07	0.00	0.02	0.00
64	0.48	7.77	-1.91	-0.28	1.96	-0.46	0.00	1.0	0.0	1.9	0.3	0.0	0.47	0.01	0.49	0.03	0.00
65	0.48	1.59	0.95	-0.81	-0.10	0.14	0.00	0.3	0.3	0.0	0.0	0.0	0.57	0.41	0.01	0.01	0.00
66	0.48	1.96	0.47	0.47	-1.03	0.68	0.00	0.1	0.1	0.5	0.7	0.0	0.11	0.11	0.54	0.24	0.00
67	0.48	1.70	1.24	-0.32	0.26	-0.01	0.00	0.4	0.0	0.0	0.0	0.0	0.90	0.06	0.04	0.00	0.00
68	0.48	2.79	-1.24	-0.97	-0.46	-0.32	0.00	0.4	0.4	0.1	0.2	0.0	0.55	0.34	0.07	0.04	0.00
69	0.48	1.06	0.83	-0.33	0.32	0.40	0.00	0.2	0.1	0.1	0.3	0.0	0.65	0.11	0.10	0.15	0.00
70	0.48	1.38	0.12	0.70	0.94	0.07	0.00	0.0	0.2	0.4	0.0	0.0	0.01	0.35	0.63	0.00	0.00
71	0.48	1.38	-0.21	0.23	-1.13	0.02	0.00	0.0	0.0	0.6	0.0	0.0	0.03	0.04	0.93	0.00	0.00
72	0.48	4.29	1.68	0.75	-0.10	0.94	0.00	0.8	0.3	0.0	1.4	0.0	0.66	0.13	0.00	0.21	0.00

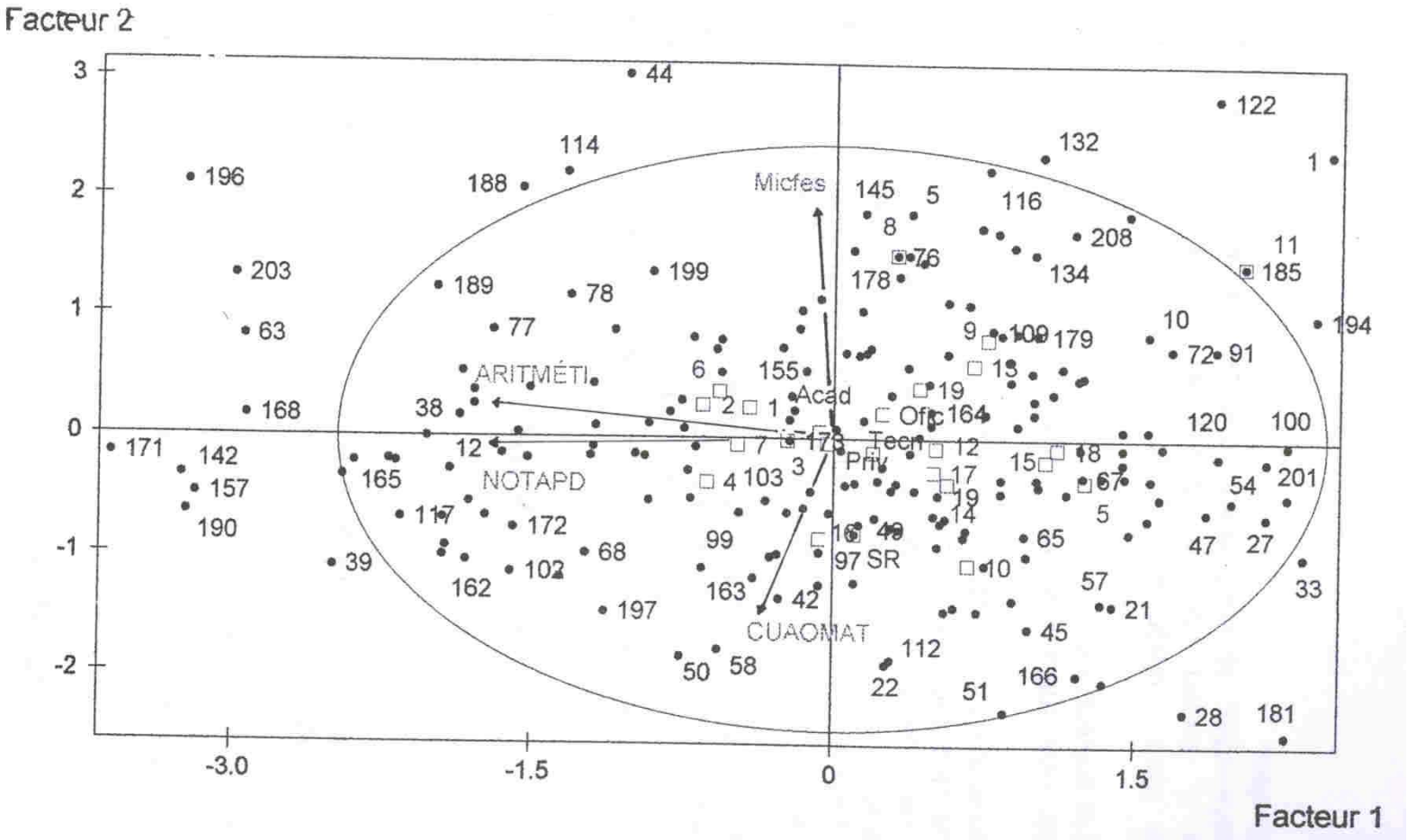
INDIVIDUS			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDENTIFICATEUR	P.REL	DISTO	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0
73	0.48	0.05	-2.22	-0.19	-1.62	-0.60	0.00	1.4	0.0	1.3	0.7	0.0	0.61	0.00	0.33	0.06	0.00
74	0.48	1.07	-0.73	-0.28	-0.12	-0.67	0.00	0.2	0.0	0.0	0.7	0.0	0.50	0.07	0.01	0.41	0.00
75	0.48	2.28	1.44	-0.31	0.23	-0.22	0.00	0.6	0.0	0.0	0.1	0.0	0.91	0.04	0.02	0.02	0.00
76	0.48	3.47	0.31	1.54	-0.97	-0.27	0.00	0.0	1.1	0.5	0.1	0.0	0.03	0.68	0.27	0.02	0.00
77	0.48	6.24	-1.70	0.90	1.49	-0.56	0.00	0.8	0.4	1.1	0.5	0.0	0.46	0.13	0.36	0.05	0.00
78	0.48	3.31	-1.32	1.18	-0.41	0.10	0.00	0.5	0.7	0.1	0.0	0.0	0.52	0.42	0.05	0.00	0.00
79	0.48	1.01	0.96	-0.09	0.04	0.13	0.00	0.3	0.5	0.0	0.0	0.0	0.48	0.51	0.00	0.01	0.00
80	0.48	9.67	0.67	1.13	-2.80	0.32	0.00	0.1	0.6	3.9	0.2	0.0	0.05	0.13	0.81	0.01	0.00
81	0.48	3.75	0.08	1.97	-1.12	-0.06	0.00	0.0	1.2	0.6	0.0	0.0	0.00	0.66	0.34	0.00	0.00
82	0.48	4.38	-1.86	0.54	0.45	0.66	0.00	1.0	0.1	0.1	0.7	0.0	0.79	0.07	0.05	0.10	0.00
83	0.48	3.12	0.60	-1.41	-0.83	0.29	0.00	0.1	0.9	0.3	0.1	0.0	0.11	0.64	0.22	0.03	0.00
84	0.48	4.10	-1.79	0.27	-0.33	-0.84	0.00	0.9	0.0	0.1	1.1	0.0	0.78	0.02	0.03	0.17	0.00
85	0.48	2.73	0.32	1.36	-0.83	-0.29	0.00	0.0	0.9	0.3	0.1	0.0	0.04	0.68	0.25	0.03	0.00
86	0.48	2.73	0.32	1.36	-0.83	-0.29	0.00	0.0	0.9	0.3	0.1	0.0	0.04	0.68	0.25	0.03	0.00
87	0.48	2.52	0.43	1.47	-0.32	-0.25	0.00	0.1	1.0	0.1	0.1	0.0	0.07	0.86	0.04	0.02	0.00
88	0.48	2.70	0.99	0.33	1.18	0.47	0.00	0.3	0.0	0.7	0.3	0.0	0.36	0.04	0.52	0.08	0.00
89	0.48	4.39	0.80	1.73	-0.55	-0.66	0.00	0.2	1.4	0.2	0.7	0.0	0.15	0.68	0.07	0.10	0.00
90	0.48	1.13	0.53	-0.71	-0.13	0.57	0.00	0.1	0.2	0.0	0.5	0.0	0.25	0.45	0.01	0.29	0.00
91	0.48	6.71	1.90	0.76	-1.51	0.52	0.00	1.0	0.3	1.1	0.4	0.0	0.54	0.09	0.34	0.04	0.00
92	0.48	3.83	-0.07	-1.23	-1.21	0.93	0.00	0.0	0.7	0.7	1.3	0.0	0.00	0.39	0.38	0.23	0.00
93	0.48	2.10	0.92	0.89	0.45	0.52	0.00	0.2	0.4	0.1	0.4	0.0	0.40	0.37	0.10	0.13	0.00
94	0.48	1.46	0.86	0.67	0.16	0.50	0.00	0.2	0.2	0.0	0.4	0.0	0.51	0.30	0.02	0.17	0.00
95	0.48	0.95	0.28	-0.74	-0.28	-0.49	0.00	0.0	0.3	0.0	0.4	0.0	0.08	0.58	0.08	0.25	0.00
96	0.48	1.90	-0.70	0.83	-0.47	0.70	0.00	0.1	0.3	0.1	0.8	0.0	0.26	0.37	0.12	0.26	0.00
97	0.48	0.73	0.10	-0.80	-0.07	-0.29	0.00	0.0	0.3	0.0	0.1	0.0	0.01	0.87	0.01	0.11	0.00
98	0.48	1.28	0.56	0.72	-0.36	-0.55	0.00	0.1	0.2	0.1	0.5	0.0	0.25	0.41	0.10	0.24	0.00
99	0.48	3.09	-0.66	-1.09	-0.73	-0.96	0.00	0.1	0.6	0.3	1.4	0.0	0.14	0.39	0.17	0.30	0.00
100	0.48	5.17	2.26	-0.04	-0.16	0.24	0.00	1.4	0.0	0.0	0.1	0.0	0.98	0.00	0.00	0.01	0.00
101	0.48	1.29	-0.17	1.06	0.23	0.31	0.00	0.0	0.5	0.0	0.2	0.0	0.02	0.86	0.04	0.08	0.00
102	0.48	7.03	-1.61	-1.13	1.50	-0.97	0.00	0.7	0.6	1.1	1.5	0.0	0.37	0.18	0.32	0.13	0.00
103	0.48	0.57	-0.35	-0.53	-0.38	0.15	0.00	0.0	0.1	0.1	0.0	0.0	0.21	0.50	0.25	0.04	0.00
104	0.48	2.34	-0.18	0.92	-1.21	0.08	0.00	0.0	0.4	0.7	0.0	0.0	0.01	0.36	0.62	0.00	0.00
105	0.48	3.35	-1.09	0.89	-0.31	1.13	0.00	0.3	0.4	0.0	2.0	0.0	0.36	0.24	0.03	0.38	0.00
106	0.48	3.60	-1.20	0.45	-0.89	1.08	0.00	0.4	0.1	0.4	1.8	0.0	0.40	0.06	0.22	0.32	0.00
107	0.48	0.13	0.00	0.06	0.36	0.01	0.00	0.0	0.0	0.1	0.0	0.0	0.00	0.03	0.97	0.00	0.00
108	0.48	1.52	1.23	-0.08	0.08	0.01	0.00	0.4	0.0	0.0	0.0	0.0	0.99	0.00	0.00	0.00	0.00
109	0.48	5.05	0.79	0.91	-1.87	0.30	0.00	0.2	0.4	1.7	0.1	0.0	0.12	0.16	0.69	0.02	0.00
110	0.48	1.77	-1.20	-0.09	-0.53	-0.23	0.00	0.4	0.0	0.1	0.1	0.0	0.81	0.00	0.16	0.03	0.00
111	0.48	9.94	0.56	-0.66	3.02	-0.26	0.00	0.1	0.2	4.5	0.1	0.0	0.03	0.04	0.92	0.01	0.00
112	0.48	5.82	0.26	-1.89	1.43	-0.38	0.00	0.0	1.7	1.0	0.2	0.0	0.01	0.61	0.35	0.03	0.00
113	0.48	2.26	0.88	0.48	-1.09	0.27	0.00	0.2	0.1	0.6	0.1	0.0	0.34	0.10	0.52	0.03	0.00
114	0.48	10.64	-1.34	2.22	1.87	0.63	0.00	0.5	2.3	1.7	0.6	0.0	0.17	0.46	0.33	0.04	0.00
115	0.48	4.02	1.47	-0.79	0.50	1.00	0.00	0.6	0.3	0.1	1.6	0.0	0.54	0.15	0.06	0.25	0.00
116	0.48	8.47	0.76	2.25	1.67	-0.16	0.00	0.2	2.4	1.4	0.0	0.0	0.07	0.60	0.33	0.00	0.00
117	0.48	5.79	-2.15	-0.69	0.35	0.75	0.00	1.3	0.2	0.1	0.9	0.0	0.80	0.08	0.02	0.10	0.00
118	0.48	1.38	0.21	-0.36	0.41	1.02	0.00	0.0	0.1	0.1	1.6	0.0	0.03	0.09	0.12	0.76	0.00
119	0.48	1.00	-0.07	-0.97	0.24	-0.09	0.00	0.0	0.4	0.0	0.0	0.0	0.01	0.93	0.06	0.01	0.00
120	0.48	4.27	1.91	-0.14	0.41	0.65	0.00	1.0	0.0	0.1	0.6	0.0	0.86	0.00	0.04	0.10	0.00
121	0.48	1.90	0.65	-0.81	0.71	0.56	0.00	0.1	0.3	0.2	0.5	0.0	0.22	0.35	0.26	0.17	0.00
122	0.48	12.26	1.90	2.86	0.67	-0.11	0.00	1.0	3.8	0.2	0.0	0.0	0.29	0.67	0.04	0.00	0.00
123	0.48	6.09	1.34	-2.03	0.18	-0.39	0.00	0.5	1.9	0.0	0.2	0.0	0.29	0.68	0.01	0.02	0.00
124	0.48	4.82	-1.36	-1.18	-1.21	-0.35	0.00	0.5	0.6	0.7	0.2	0.0	0.38	0.29	0.30	0.02	0.00
125	0.48	3.54	0.71	-1.45	0.06	-0.97	0.00	0.1	1.0	0.0	1.5	0.0	0.14	0.59	0.00	0.27	0.00
126	0.48	1.13	0.53	-0.71	-0.13	0.57	0.00	0.1	0.2	0.0	0.5	0.0	0.25	0.45	0.01	0.29	0.00
127	0.48	5.22	-1.95	-0.99	0.41	0.51	0.00	1.1	0.5	0.1	0.4	0.0	0.73	0.19	0.03	0.05	0.00
128	0.48	0.74	0.06	0.71	0.47	0.07	0.00	0.0	0.2	0.1	0.0	0.0	0.00	0.69	0.30	0.01	0.00
129	0.48	4.65	-1.95	-0.68	0.32	0.54	0.00	1.1	0.2	0.1	0.5	0.0	0.81	0.10	0.02	0.06	0.00
130	0.48	8.10	-0.93	0.11	2.69	0.01	0.00	0.2	0.0	3.6	0.0	0.0	0.11	0.00	0.89	0.00	0.00
131	0.48	2.80	0.32	-0.77	1.43	-0.27	0.00	0.0	0.3	1.0	0.1	0.0	0.04	0.21	0.73	0.03	0.00
132	0.48	10.84	1.03	2.36	2.02	-0.35	0.00	0.3	2.6	2.0	0.2	0.0	0.10	0.52	0.37	0.01	0.00
133	0.48	0.41	0.03	-0.10	0.63	-0.01	0.00	0.0	0.0	0.2	0.0	0.0	0.00	0.02	0.97	0.00	0.00
134	0.48	3.98	1.00	1.54	-0.68	0.37	0.00	0.3	1.1	0.2	0.2	0.0	0.25	0.60	0.12	0.03	0.00
135	0.48	2.89	1.56	0.07	-0.53	-0.39	0.00	0.7	0.0	0.1	0.2	0.0	0.85	0.00	0.10	0.05	0.00
136	0.48	1.40	0.52	-0.90	-0.13	0.55	0.00	0.1	0.4	0.0	0.5	0.0	0.19	0.58	0.01	0.22	0.00
137	0.48	6.73	0.72	1.76	1.75	-0.23	0.00	0.1	1.4	1.5	0.1	0.0	0.08	0.46	0.46	0.01	0.00
138	0.48	0.50	0.06	-0.40	-0.53	-0.25	0.00	0.0	0.1	0.1	0.1	0.0	0.01	0.32	0.55	0.12	0.00
139	0.48	1.65	0.65	-0.76	-0.72	0.35	0.00	0.1	0.3	0.3	0.2	0.0	0.26	0.35	0.32	0.08	0.00
140	0.48	1.30	0.10	-0.37	-0.34	1.02	0.00	0.0	0.1	0.1	1.6	0.0	0.01	0.11	0.09	0.79	0.00
141	0.48	3.17	1.56	-0.67	-0.27	-0.47	0.00	0.7	0.2	0.0	0.3	0.0	0.77	0.14	0.02	0.07	0.00
142	0.48	13.20	-3.25	-0.34	1.38	0.77	0.00	3.0	0.1	0.9	0.9	0.0	0.80	0.01	0.14	0.05	0.00
143	0.48	2.86	1.00	0.21	1.27	0.46	0.00	0.3	0.0	0.8	0.3	0.0	0.35	0.02	0.56	0.07	0.00
144	0.48	2.19	0.56	1.15	-0.54	-0.51	0.00	0.1	0.6	0.1	0.4	0.0	0.14	0.60	0.13	0.12	0.00
145	0.48	7.83	0.14	1.88	2.03	0.40	0.00	0.0	1.6	2.0	0.3	0.0	0.00	0.45	0.53	0.02	0.00
146	0.48	0.68	0.36	0.60	-0.25	-0.36	0.00	0.0	0.2	0.0	0.2	0.0	0.20	0.53	0.09	0.19	0.00

INDIVIDUS			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDENTIFICATEUR	P.REL	DISTO	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0
147	0.48	1.30	-0.76	0.31	-0.57	-0.61	0.00	0.2	0.0	0.2	0.6	0.0	0.42	0.07	0.24	0.27	0.00
148	0.48	5.62	-1.53	-0.18	-1.24	-1.30	0.00	0.7	0.0	0.8	2.6	0.0	0.42	0.01	0.27	0.30	0.00
149	0.48	1.83	-1.00	-0.13	0.88	-0.23	0.00	0.3	0.0	0.4	0.1	0.0	0.54	0.01	0.42	0.03	0.00
150	0.48	1.26	-0.72	-0.51	0.06	-0.69	0.00	0.1	0.1	0.0	0.7	0.0	0.41	0.21	0.00	0.38	0.00
151	0.48	1.79	-0.08	1.17	-0.64	0.11	0.00	0.0	0.6	0.2	0.0	0.0	0.00	0.76	0.23	0.01	0.00
152	0.48	1.42	-0.59	0.74	0.47	-0.56	0.00	0.1	0.3	0.1	0.5	0.0	0.24	0.38	0.15	0.22	0.00
153	0.48	7.39	0.91	0.11	-2.25	-1.22	0.00	0.2	0.0	2.5	2.3	0.0	0.11	0.00	0.69	0.20	0.00
154	0.48	3.68	-1.74	-0.67	0.29	0.34	0.00	0.9	0.2	0.0	0.2	0.0	0.82	0.12	0.02	0.03	0.00
155	0.48	5.75	-0.15	0.56	-2.32	-0.16	0.00	0.0	0.1	2.7	0.0	0.0	0.00	0.05	0.94	0.00	0.00
156	0.48	2.12	-0.48	-0.62	-1.22	0.14	0.00	0.1	0.2	0.7	0.0	0.0	0.11	0.18	0.70	0.01	0.00
157	0.48	10.91	-3.19	-0.49	0.41	0.55	0.00	2.9	0.1	0.1	0.5	0.0	0.93	0.02	0.02	0.03	0.00
158	0.48	1.12	-0.75	0.07	-0.39	-0.63	0.00	0.2	0.0	0.1	0.6	0.0	0.50	0.00	0.14	0.36	0.00
159	0.48	1.95	-1.22	-0.16	-0.62	-0.24	0.00	0.4	0.0	0.2	0.1	0.0	0.76	0.01	0.20	0.03	0.00
160	0.48	4.71	-2.03	0.00	-0.40	-0.66	0.00	1.2	0.0	0.1	0.7	0.0	0.87	0.00	0.03	0.09	0.00
161	0.48	2.39	1.09	0.37	-1.03	0.05	0.00	0.3	0.1	0.5	0.0	0.0	0.50	0.06	0.44	0.00	0.00
162	0.48	4.56	-1.83	-1.04	-0.18	0.30	0.00	0.9	0.5	0.0	0.1	0.0	0.73	0.24	0.01	0.02	0.00
163	0.48	1.80	-0.40	-1.18	-0.49	0.09	0.00	0.0	0.6	0.1	0.0	0.0	0.09	0.77	0.13	0.00	0.00
164	0.48	1.44	0.48	0.23	-0.85	0.66	0.00	0.1	0.0	0.4	0.7	0.0	0.16	0.04	0.50	0.30	0.00
165	0.48	5.81	-2.38	-0.22	0.12	-0.26	0.00	1.6	0.0	0.0	0.1	0.0	0.98	0.01	0.00	0.01	0.00
166	0.48	6.03	1.22	-1.08	0.78	-0.17	0.00	0.4	1.8	0.3	0.0	0.0	0.25	0.65	0.10	0.00	0.00
167	0.48	3.08	-0.69	-0.08	-1.36	-0.86	0.00	0.1	0.0	0.9	1.2	0.0	0.16	0.00	0.60	0.24	0.00
168	0.48	9.04	-2.93	0.16	0.50	0.40	0.00	2.4	0.0	0.1	0.3	0.0	0.95	0.00	0.03	0.02	0.00
169	0.48	2.59	-1.57	0.04	-0.28	0.19	0.00	0.7	0.0	0.0	0.1	0.0	0.96	0.00	0.03	0.01	0.00
170	0.48	7.25	0.49	-0.65	2.55	-0.26	0.00	0.1	0.2	3.2	0.1	0.0	0.03	0.06	0.90	0.01	0.00
171	0.48	14.18	-3.62	-0.16	0.20	0.99	0.00	3.7	0.0	0.0	1.5	0.0	0.93	0.00	0.00	0.07	0.00
172	0.48	3.16	-1.59	-0.77	-0.11	0.12	0.00	0.7	0.3	0.0	0.0	0.0	0.80	0.19	0.00	0.00	0.00
173	0.48	0.18	-0.22	-0.02	0.29	0.21	0.00	0.0	0.0	0.0	0.1	0.0	0.28	0.00	0.48	0.24	0.00
174	0.48	3.19	1.43	-0.08	-0.04	1.06	0.00	0.6	0.0	0.0	1.8	0.0	0.64	0.00	0.00	0.35	0.00
175	0.48	1.51	0.99	0.56	-0.39	0.28	0.00	0.3	0.1	0.1	0.1	0.0	0.64	0.21	0.10	0.05	0.00
176	0.48	5.38	2.15	-0.18	-0.82	0.22	0.00	1.3	0.0	0.3	0.1	0.0	0.86	0.01	0.12	0.01	0.00
177	0.48	0.15	-0.23	0.16	0.16	0.22	0.00	0.0	0.0	0.0	0.1	0.0	0.35	0.16	0.16	0.33	0.00
178	0.48	7.67	0.37	1.53	2.27	0.16	0.00	0.0	1.1	2.6	0.0	0.0	0.02	0.31	0.67	0.00	0.00
179	0.48	3.30	1.01	0.87	1.01	-0.71	0.00	0.3	0.4	0.5	0.8	0.0	0.31	0.23	0.31	0.15	0.00
180	0.48	5.50	-1.51	0.40	1.54	-0.81	0.00	0.6	0.1	1.2	1.0	0.0	0.42	0.03	0.43	0.12	0.00
181	0.48	11.91	2.25	-2.49	0.80	0.00	0.00	1.4	2.9	0.3	0.0	0.0	0.43	0.52	0.05	0.00	0.00
182	0.48	1.64	0.23	-0.23	-0.93	0.82	0.00	0.0	0.0	0.4	1.0	0.0	0.03	0.03	0.52	0.41	0.00
183	0.48	5.96	0.89	1.62	1.53	-0.46	0.00	0.2	1.2	1.2	0.3	0.0	0.13	0.44	0.39	0.03	0.00
184	0.48	6.22	-2.44	-0.33	-0.26	-0.27	0.00	1.7	0.0	0.0	0.1	0.0	0.96	0.02	0.01	0.01	0.00
185	0.48	7.32	2.03	1.46	-0.84	0.59	0.00	1.2	1.0	0.3	0.5	0.0	0.57	0.29	0.10	0.05	0.00
186	0.48	3.77	-1.81	-0.54	-0.26	0.35	0.00	0.9	0.1	0.0	0.2	0.0	0.87	0.08	0.02	0.03	0.00
187	0.48	4.29	1.44	0.06	1.49	0.03	0.00	0.6	0.0	1.1	0.0	0.0	0.48	0.00	0.52	0.00	0.00
188	0.48	7.50	-1.56	2.08	0.56	-0.66	0.00	0.7	2.0	0.2	0.7	0.0	0.32	0.58	0.04	0.06	0.00
189	0.48	10.04	-1.98	1.24	-2.08	0.51	0.00	1.1	0.7	2.2	0.4	0.0	0.39	0.15	0.43	0.03	0.00
190	0.48	11.17	-3.23	-0.64	0.22	0.53	0.00	2.9	0.2	0.0	0.4	0.0	0.93	0.04	0.00	0.03	0.00
191	0.48	2.43	-0.32	-1.00	0.08	-1.15	0.00	0.0	0.5	0.0	2.1	0.0	0.04	0.41	0.00	0.54	0.00
192	0.48	3.19	1.22	0.51	-0.27	-1.17	0.00	0.4	0.1	0.0	2.1	0.0	0.47	0.08	0.02	0.43	0.00
193	0.48	5.10	-1.93	-0.92	0.50	0.52	0.00	1.0	0.4	0.1	0.4	0.0	0.73	0.17	0.05	0.05	0.00
194	0.48	7.84	2.40	1.02	-1.00	0.13	0.00	1.6	0.5	0.5	0.0	0.0	0.74	0.13	0.13	0.00	0.00
195	0.48	2.57	0.37	-0.12	1.54	-0.21	0.00	0.0	0.0	1.2	0.1	0.0	0.05	0.01	0.92	0.02	0.00
196	0.48	20.28	-3.24	2.12	1.95	1.23	0.00	2.9	2.1	1.9	2.3	0.0	0.52	0.22	0.19	0.07	0.00
197	0.48	3.71	-1.15	-1.46	0.37	-0.36	0.00	0.4	1.0	0.1	0.2	0.0	0.35	0.57	0.04	0.04	0.00
198	0.48	5.02	-2.18	-0.22	0.09	-0.47	0.00	1.3	0.0	0.0	0.3	0.0	0.95	0.01	0.00	0.04	0.00
199	0.48	3.18	-0.92	1.37	-0.61	-0.30	0.00	0.2	0.9	0.2	0.1	0.0	0.26	0.59	0.12	0.03	0.00
200	0.48	1.73	-1.19	0.10	-0.52	-0.22	0.00	0.4	0.0	0.1	0.1	0.0	0.81	0.01	0.16	0.03	0.00
201	0.48	5.35	2.26	-0.47	0.01	0.20	0.00	1.4	0.1	0.0	0.1	0.0	0.95	0.04	0.00	0.01	0.00
202	0.48	0.75	0.28	0.37	0.71	-0.17	0.00	0.0	0.1	0.3	0.0	0.0	0.11	0.18	0.68	0.04	0.00
203	0.48	14.02	-2.99	1.33	-1.79	0.30	0.00	2.5	0.8	1.6	0.1	0.0	0.64	0.13	0.23	0.01	0.00
204	0.48	3.23	1.16	-0.45	-0.30	-1.27	0.00	0.4	0.1	0.0	2.5	0.0	0.42	0.06	0.03	0.49	0.00
205	0.48	1.61	-0.57	0.81	0.56	-0.56	0.00	0.1	0.3	0.2	0.5	0.0	0.20	0.41	0.20	0.19	0.00
206	0.48	1.38	-0.76	0.31	-0.57	-0.61	0.00	0.2	0.0	0.2	0.6	0.0	0.42	0.07	0.24	0.27	0.00
207	0.48	1.61	-0.57	0.55	-0.88	0.47	0.00	0.1	0.1	0.4	0.3	0.0	0.20	0.19	0.48	0.13	0.00
208	0.48	4.86	1.19	1.73	0.54	0.40	0.00	0.4	1.4	0.1	0.2	0.0	0.29	0.62	0.06	0.03	0.00
209	0.48	0.79	-0.26	0.75	-0.29	0.28	0.00	0.0	0.3	0.0	0.1	0.0	0.09	0.70	0.11	0.10	0.00

COORDONNEES ET VALEURS-TEST DES MODALITES
AXES 1 A 4

MODALITES			VALEURS-TEST					COORDONNEES					
IDEN - LIBELLE	EFF.	P.ABS	1	2	3	4	0	1	2	3	4	0	DISTO.
5 . TICOL													
TIC1 - Ofic	33	33.00	1.1	1.2	1.1	0.5	0.0	0.24	0.20	0.18	0.05	0.00	0.13
TIC2 - Priv	176	176.00	-1.1	-1.2	-1.1	-0.5	0.0	-0.04	-0.04	-0.03	-0.01	0.00	0.00
6 . BACHT													
BAC1 - Acad	151	151.00	-1.3	1.2	0.8	-2.0	0.0	-0.07	0.05	0.03	-0.05	0.00	0.01
BAC2 - Tech	57	57.00	1.3	-1.1	-0.8	2.0	0.0	0.19	-0.13	-0.09	0.13	0.00	0.08
6_ - reponse manquante	1	1.00	0.1	-0.8	-0.1	-0.5	0.0	0.10	-0.80	-0.07	-0.29	0.00	0.73
7 . PROGS													
PR1 - 1	13	13.00	-1.2	0.9	-1.1	-0.2	0.0	-0.42	0.25	-0.28	-0.03	0.00	0.32
PR2 - 2	35	35.00	-3.3	1.8	0.6	1.3	0.0	-0.66	0.28	0.08	0.11	0.00	0.53
PR3 - 3	4	4.00	-0.4	0.0	-1.0	-0.8	0.0	-0.23	-0.02	-0.49	-0.22	0.00	0.35
PR4 - 4	25	25.00	-2.6	-2.0	0.8	-0.3	0.0	-0.63	-0.38	0.15	-0.03	0.00	0.57
PR5 - 5	2	2.00	1.4	-0.5	-0.7	1.3	0.0	1.24	-0.35	-0.46	0.51	0.00	2.14
PR6 - 6	17	17.00	-1.9	1.6	-0.9	1.3	0.0	-0.57	0.38	-0.21	0.17	0.00	0.54
PR7 - 7	14	14.00	-1.4	-0.2	0.3	-0.8	0.0	-0.48	-0.06	0.09	-0.11	0.00	0.26
PR8 - 8	1	1.00	0.2	1.5	-1.0	-0.5	0.0	0.31	1.54	-0.97	-0.27	0.00	3.47
PR9 - 9	6	6.00	1.4	2.0	1.0	0.1	0.0	0.76	0.83	0.41	0.02	0.00	1.43
PR10 - 10	11	11.00	1.7	-3.6	-1.3	1.4	0.0	0.67	-1.07	-0.37	0.23	0.00	1.77
PR11 - 11	1	1.00	1.6	1.4	-0.9	1.1	0.0	2.03	1.46	-0.84	0.59	0.00	7.32
PR12 - 12	6	6.00	1.0	-0.2	2.3	-0.9	0.0	0.50	-0.09	0.92	-0.20	0.00	1.14
PR13 - 13	17	17.00	2.3	2.6	-1.5	0.5	0.0	0.70	0.61	-0.34	0.06	0.00	0.98
PR14 - 14	34	34.00	2.8	-2.4	0.3	-2.4	0.0	0.57	-0.38	0.05	-0.21	0.00	0.52
PR15 - 15	2	2.00	1.1	-0.3	-0.6	1.9	0.0	1.06	-0.18	-0.38	0.73	0.00	1.83
PR16 - 16	5	5.00	-0.1	-1.9	-0.2	-0.5	0.0	-0.07	-0.84	-0.08	-0.12	0.00	0.74
PR17 - 17	3	3.00	0.7	-0.5	0.7	1.0	0.0	0.49	-0.28	0.40	0.33	0.00	0.59
PR18 - 18	3	3.00	1.5	-0.1	0.3	-0.8	0.0	1.10	-0.08	0.18	-0.26	0.00	1.33
PR19 - 19	10	10.00	1.1	1.3	1.2	-0.6	0.0	0.43	0.42	0.36	-0.10	0.00	0.49


Facteur 2
1.5
1.0
0.5
0
-0.5
-1.0
-0.5
0
0.5
1.0
1.5
2.0
Facteur 1
8
11
9
13
Micfes
19
6
2
1
ARITMÉTI
Ofic
Acad
7
3
NOTAPD
Priv
Tecn
12
18
15
17
4
QUAOMAT
14
5
16
reponse manquante
10


Facteur 2
Facteur 1
Micfes
ARITMÉTI
NOTAPD
CUAOMAT
Acad
Ofic
Teen
Priv
SR
3
2
1
0
-1
-2
-3.0
-1.5
0
1.5

CONCLUSIONES

1. En el análisis Estadístico basados en los métodos de Regresión Lineal Múltiple y Análisis de Componentes Principales no se encontró estructuras relacionales importantes en e sistema de información.

2. No se encontró ninguna explicación de la nota obtenida en el curso de Matemáticas a partir de la nota obtenida en la prueba diagnóstica y la nota del área de Matemáticas en el Icfes.

HIPÓTESIS

1. Es probable que las respuestas a las pruebas icfes y diagnóstica obedezcan más a comportamientos aleatorios a la hora de responder. Se recomienda en la aplicación de futuras pruebas diagnósticas realizar pruebas de aletoriedad mediante test apareados

2. La deserción o cancelación del curso de Matemáticas 1 o su aprobación, pueden estar ligados a otras variables no medidas tales como situación económica ambiente familiar, expectativas no satisfechas, tiempos de desplazamiento entre el sitio de trabajo o el hogar y la Universidad, tiempo de dedicación al estudio, relación del contenido con el resto de cursos, entre otras

3. Puede existir una influencia importante en la transición entre la secundaria y la universidad

4. Las pruebas icfes, diagnóstica y las evaluaciones del curso de Matemáticas 1, tienen objetivos diferentes y son realizadas en momentos diferentes; lo cual puede conllevar a una ausencia de relación entre ellas.