ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

“Implementación de Minería de Datos Basada en Redes Bayesianas para la Toma de Decisiones en los Registros Académicos”

TESIS DE GRADO

Previo la obtención del Título de:

INGENIERO EN COMPUTACIÓN

Presentada por:

Alex Fernando Bonilla Gordillo
Miguel Angel Ojeda Schuldt

GUAYAQUIL – ECUADOR

Año: 2006

A G R A D E C I M I E N T O

A todas las personas que de una u otra manera colaboraron en la realización de este trabajo y en especial el Ing. Fabricio Echeverría, Director de Tesis.

D E D I C A T O R I A

DIOS

NUESTROS PADRES

NUESTROS HERMANOS

NUESTROS AMIGOS

TRIBUNAL DE GRADUACIÓN

_____	_____
Ing. Holger Cevallos SUB-DECANO DE LA FIEC PRESIDENTE	Ing. Fabricio Echeverría B. DIRECTOR DE TESIS
_____	_____
Ing. Ana Tapia R. VOCAL	Ing. Carlos Jordán V. VOCAL

DECLARACIÓN EXPRESA

"La responsabilidad del contenido de esta Tesis de Grado, nos corresponden exclusivamente; y el patrimonio intelectual de la misma a la ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL"

(Reglamento de Graduación de la ESPOL).

_____
Alex Fernando Bonilla Gordillo

_____
Miguel Angel Ojeda Schuldt

RESUMEN

Una de las decisiones a tomar dentro de la planificación académica en la Facultad de Ingeniería en Electricidad y Computación es la apertura de paralelos de las materias ofertadas a sus estudiantes.

Durante la toma de decisión sobre la apertura de paralelos se presenta el problema de ajustar el número de cupos disponibles, frente a la demanda de los estudiantes de la FIEC que cumplen con los requisitos en cada materia de sus respectivas carreras.

Como consecuencia a este problema, es posible que se presenten algunos inconvenientes:

- Apertura de paralelos en los cuales no se registra el número mínimo de estudiantes de la FIEC.
- Cupo agotado en paralelos cuando aún existen estudiantes de la FIEC que desean tomar la materia.

Aplicando las redes bayesianas como método de minería de datos, perseguimos encontrar un valor que lleve a la solución de la pregunta:

“¿Cuántos paralelos de las diferentes materias deben abrirse en el semestre?”

La minería de datos es una actividad de extracción cuyo objetivo es el descubrir hechos contenidos en las bases de datos. En la mayoría de los casos se refiere a un trabajo automatizado. Si hay alguna intervención humana a lo largo del proceso, no es considerado como minería de datos.

Las redes bayesianas, proveen una forma compacta de representar el conocimiento y métodos flexibles de razonamiento -basados en las teorías probabilísticas- capaces de predecir el valor de variables no observadas y explicar las observadas. Entre las características que poseen las redes bayesianas, se puede destacar que permiten aprender sobre relaciones de dependencia y causalidad, permiten combinar conocimiento con datos, evitan el sobre ajuste de los datos y pueden manejar bases de datos incompletas.

ÍNDICE GENERAL

Pág.

CAPÍTULO 4

ABREVIATURAS

CPU	Central Process Unit (Unidad Central de Procesamiento)
FIEC	Facultad Ingeniería Eléctrica y Computación
GPS	Global Positioning System (Sistema de Posicionamiento Global)
HC	Hill Climbing (Ascensión a la Colina)
IP	Internet Protocol (Protocolo Internet)
JDSK	Java Servlet Development Kit
KDD	Knowledge Discovery in Database (Descubrimiento de Conocimiento en Bases de Datos)
MAP	Maximun A Posteriori
NB	Naive Bayes
OLAP	On-line Analytical Processing (Procesamiento Analítico en Línea)
OLTP	On-line Transaction Processing (Procesamiento Transaccional en Línea)
PERL	Practical Extraction and Report Language
RBs	Redes Bayesianas
SO	Sistema Operativo
SQL	Structured Query Language (Lenguaje de Consulta Estructurado)
TCP	Transmission Control Protocol (Protocolo de Control de Transmisión)
URL	Uniform Resource Locator (Localizador Uniforme de Recurso)
POO	Programación Orientada a Objetos
DIO	Diagrama de Interacción a Objetos
CSV	Comma Separated Value
BD	Base de Datos

SIMBOLOGÍA

$P (A \mid B)$	Probabilidad que un evento A ocurra dado que ocurre un evento B
$P (A)$	Probabilidad de un evento A
$Pa (X_i)$	Padres de la variable X_i
G	Grafo
Θ	Conjunto de Distribución de Probabilidad
$\int$	Integral
$\prod$	Multiplicador
$\sum$	Sumatoria

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE FIGURAS

ÍNDICE DE FÓRMULAS

INTRODUCCIÓN

“Implementación de Minería de Datos Basada en Redes Bayesianas para la Toma de Decisiones en los Registros Académicos”, persigue desarrollar un modelo o conjunto de reglas que permita mejorar la Planificación Académica” en particular, el proceso de apertura de paralelos para las distintas materias, este trabajo tiene como objetivos:

- Determinar un método para encontrar las relaciones existentes en un conjunto de hechos.
- Determinar un método para resolver las proyecciones probabilísticas “que sí” de un almacén de datos (Data Warehouse)
- Sustentar el uso de la minería de datos para la toma de decisiones.

A continuación se detalla brevemente los capítulos de este trabajo:

Capítulo 1: describe conceptos básicos de la minería de datos y su importancia. También sus distintos métodos y una clasificación de los mismos según las tareas que resuelve cada uno. Se mencionan algunas aplicaciones y uso de la minería de datos.

Capítulo 2: introduce los métodos bayesianos para la minería de datos, la base teórica sobre la cual se fundamentan (Teorema de Bayes), se describe formalmente las redes Bayesianas y su utilidad para crear modelos probabilísticos.

Capítulo 3: estudia los antecedentes y el problema actual de los registros académicos, usando técnicas de POO. También se estudia las implicaciones de la toma de decisiones en la planificación del término académico.

Capítulo 4: detalla la metodología utilizada para el diseño de la solución, seguida de su implementación. Verifica la validez de los resultados obtenidos.

CAPÍTULO 1

1. CONCEPTOS BÁSICOS DE MINERÍA DE DATOS.

El aumento del volumen y variedad de información que se encuentra informatizada en bases de datos y otras fuentes ha crecido exponencialmente en las últimas décadas. Gran parte de esta información es histórica, es decir, representa transacciones o situaciones que se han producido. Además, ya que los datos pueden proceder de fuentes diversas y pertenecer a diferentes dominios, parece clara la inminente necesidad de analizarlos para la obtención de información útil para la organización.

En muchas situaciones, el método tradicional de convertir los datos en conocimiento consiste en un análisis e interpretación realizada manualmente. El especialista, quien es la persona que conoce el modelo del negocio y su comportamiento, analiza los datos y elabora un informe o

hipótesis que refleja las tendencias o pautas de los mismos. Este conocimiento, validado convenientemente, puede ser usado por los directivos de la organización para tomar decisiones importantes y significativas para la organización. Esta forma de actuar es altamente subjetiva y costosa en tiempo, dinero y productividad. De hecho, el análisis manual es impracticable en dominios donde el volumen de los datos crece exponencialmente: la abundancia de datos desborda la capacidad humana de comprenderlos sin la ayuda de herramientas potentes. Muchas decisiones importantes se realizan, no sobre la gran cantidad de datos disponibles, sino siguiendo la propia intuición del usuario al no disponer de las herramientas necesarias. Este es el principal cometido de la minería de datos: resolver problemas analizando los datos presentes en las bases de datos.

El análisis de los datos de una base de datos se realizaba mediante consultas efectuadas con lenguajes de consulta, como el SQL, y se producía sobre la base de datos operacional, es decir, junto al procesamiento transaccional en línea (OLTP) de las aplicaciones de gestión. No obstante, esta manera de actuar sólo permitía generar información

resumida de una manera previamente establecida, poco flexible y, sobretodo, poco escalable a grandes volúmenes de datos.

La tecnología de bases de datos ha respondido a este reto con una nueva arquitectura surgida recientemente: el almacén de datos (DataWarehouse). Se trata de un repositorio de fuentes heterogéneas de datos, integrados y organizados bajo un esquema unificado para facilitar su análisis y dar soporte a la toma de decisiones. Esta tecnología incluye operaciones de procesamiento analítico en línea (OLAP), es decir, técnicas de análisis como pueden ser el resumen, la consolidación o la agregación, así como la posibilidad de ver la información desde distintas perspectivas.

A pesar que las herramientas OLAP soportan cierto análisis descriptivo y de sumarización que permiten transformar los datos en otros datos agregados o cruzados de manera sofisticada, no generan reglas, patrones, pautas, es decir, conocimiento que pueda ser aplicado a otros datos. En muchos contextos, como los negocios, la medicina o la ciencia, los datos por sí solos tienen un valor relativo. Lo que de verdad es interesante es el conocimiento que puede inferirse a partir de los datos y, más aún, la capacidad de poder usar este conocimiento.

Existen otras herramientas analíticas que han sido empleadas para analizar los datos y que tienen su origen en la estadística, algo lógico teniendo en cuenta que la materia prima de esta disciplina son precisamente los datos. Aunque algunos paquetes estadísticos son capaces de inferir patrones a partir de los datos, el problema es que resultan muy confusos para los no estadísticos y no se integran bien con los sistemas de información. Sería injusto no reconocer que la estadística es, en cierto modo, la “madre” de la minería de datos.

Todos estos problemas y limitaciones de las aproximaciones clásicas han hecho surgir la necesidad de una nueva generación de herramientas y técnicas para soportar la extracción de conocimiento útil desde la información disponible, y que engloban bajo la denominación de minería de datos. El resultado de la minería de datos son reglas de asociación, ecuaciones, árboles de decisión, redes neuronales, redes bayesianas y otras representaciones menos usadas.

1.1 Definición de la Minería de Datos

Existen definiciones sobre la minería de datos según diferentes autores. La Minería de Datos “como el proceso de extraer conocimiento útil,

comprensible, previamente desconocido, de grandes cantidades de datos almacenados en distintos formatos” [Introducción a la Minería de Datos]. Su tarea fundamental es encontrar modelos inteligibles a partir de los datos, para ayudar a mejorar la toma de decisiones en los procesos fundamentales de un negocio, generado de forma automática o asistida.

- La minería de datos permite obtener valores a partir de la información que registran y manejan las empresas, lo que ayuda a dirigir esfuerzos de mejora respaldados en datos históricos de diversa índole.

- La minería de datos trabaja con enormes volúmenes de datos, que vienen de sistemas de información, con los problemas típicos que conlleva, tales como intratabilidad, volatilidad y ausencia de los datos. Usa técnicas adecuadas para analizar los mismos y extraer conocimiento novedoso y útil.

- La minería de datos es un campo multidisciplinar que se ha desarrollado en paralelo o como prolongación de otras

tecnologías. Por ello, la investigación y los avances en la minería de datos se nutren de los que se producen en estas áreas relacionadas.

Las disciplinas más influyentes son las siguientes:

La recuperación de información.- consiste en obtener información desde datos textuales, por lo que su desarrollo histórico se ha basado en el uso efectivo de bibliotecas y en la búsqueda por Internet. Una tarea común es encontrar documentos a partir de palabras determinas, lo cual puede verse como un proceso de clasificación de los documentos.

La estadística.- ha proporcionado muchos de los conceptos, algorítmicos y técnicas que se utilizan en minería de datos, como por ejemplo: la media, la varianza, las distribuciones, el análisis univariante y multivariante, la regresión lineal y no lineal, la teoría del muestreo, la validación cruzada y las técnicas bayesianas.

El aprendizaje automático.- la máquina aprende un modelo a partir de ejemplos y es usado para resolver el problema.

Los sistemas para la toma de decisiones.- son herramientas y sistemas informatizados que asisten a los directivos en la resolución de problemas y en la toma de decisiones. El objetivo es proporcionar la información necesaria para realizar decisiones efectivas en el ámbito directivo o en tareas de diagnóstico.

La visualización de datos.- permite al usuario descubrir, intuir o entender patrones que serían más difíciles de “visualizar” a partir de descripciones matemáticas o textuales de los resultados. A través de gráficas como por ejemplo: diagrama de barras, histogramas, tortas.

La computación paralela y distribuida.- actualmente, muchos sistemas de bases de datos comerciales incluyen tecnologías de procesamiento paralelo o distribuido. En estos sistemas el coste computacional de las tareas más complejas de minería de datos se reparte entre diferentes procesadores o computadores. Su éxito se debe en parte a la explosión de los almacenes de datos y de la minería de datos, en lo que las prestaciones de los algoritmos de consulta son críticas.

Otras disciplinas.- dependiendo del tipo de dato a ser minados o del tipo de aplicación, la minería de datos usa también técnicas de otras disciplinas como el lenguaje natural, el análisis de imágenes, el procesamiento de señales, los gráficos por computadora, entre otros.

La minería de datos es fundamental en la investigación científica y técnica, como herramienta de análisis y descubrimiento de conocimiento a partir de datos tomados de observaciones o resultados.

El objetivo de la minería de datos es convertir datos en conocimiento.

1.2 Tipos de Modelos

La minería de datos analizan los datos para extraer conocimiento, el cuál puede ser en forma de relaciones, patrones o reglas inferidos de los datos, o bien en forma de una descripción o un resumen de los mismos. Lo anterior, constituye el modelo de los datos analizados. Existen diversas formas de representar los modelos y cada una de ellas determina el tipo de técnica que puede usarse para inferirlos. En la práctica los modelos pueden ser de dos tipos: predictivos o descriptivos.

1.2.1 Modelos Predictivos

Los modelos predictivos pretenden estimar valores futuros o desconocidos de variables de interés, que denominamos *variables objetivo o dependientes*, usando otras variables o campos de la base de datos, llamadas *variables independientes o predictivas.* Un ejemplo sería un modelo predictivo que permita estimar la demanda de un nuevo producto en función del gasto de publicidad.

1.2.2 Modelos Descriptivos

Los modelos descriptivos, identifican patrones que explican o resumen los datos. Sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos. Por ejemplo, un agencia de viaje desea identificar grupos de personas con los mismos gustos, con el objeto de organizar diferentes ofertas para cada grupo y poder así remitirles esta información; para lograrlo se analizan los viajes que han realizado sus clientes e infiere un modelo descriptivo que caracteriza estos grupos.

Algunas tareas de minería de datos que producen modelos predictivos son la clasificación y la regresión, y las que dan lugar a modelos descriptivos son el agrupamiento, las reglas de asociación y el análisis correlacional.

1.3 La Minería de Datos y el proceso de Descubrimiento de Conocimiento en Bases de Datos.

Un término muy utilizado, y el más relacionado con la minería de datos, es la extracción o “descubrimiento de conocimiento en bases de datos”. De hecho, en muchas ocasiones ambos términos se han utilizado indistintamente, aunque existen claras diferencias entre los dos. Así, últimamente se ha usado el término KDD para referirse a un proceso que consta de una serie de fases, mientras que la minería de datos es sólo una de estas fases [Introducción a la Minería de Datos]. Las fases se detallan a continuación:

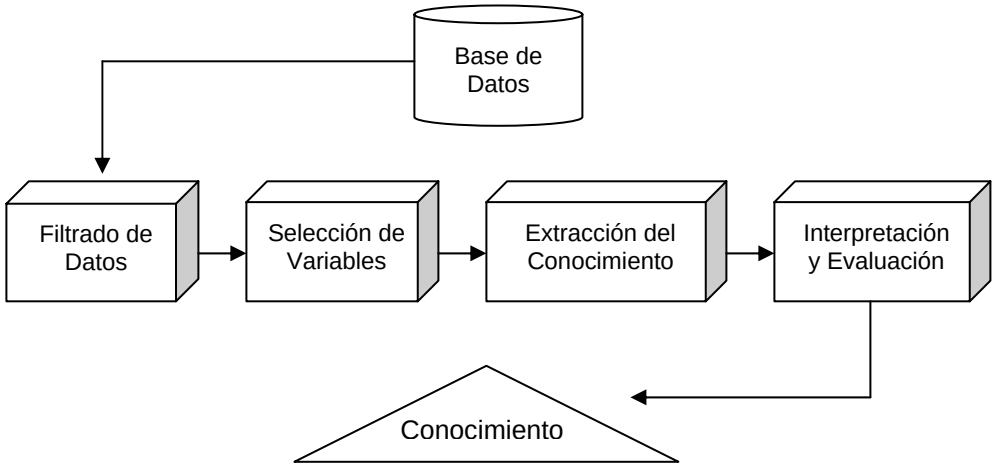


FIGURA 1.1 FASES DEL PROCESO DE DESCUBRIMIENTO DE CONOCIMIENTO EN BASE DE DATOS

Filtrado de Datos

El formato de los datos contenidos en la fuente de datos nunca es el idóneo, y la mayoría de las veces no es posible ni siquiera utilizar algún algoritmo de minería sobre la fuente de datos.

Mediante un preprocesado, se filtran los datos (de forma que se eliminan valores incorrectos, no válidos), se obtienen muestras de los mismos o reducen el número de valores posibles mediante redondeo, clustering (1).

Selección de Variables

A pesar de haber pasado por la filtración de datos, en la mayoría de los casos se tiene una cantidad grande de datos. La selección de variables reduce el tamaño de los datos eligiendo las variables más influyentes en el problema, sin sacrificar la calidad del modelo de conocimiento obtenido del proceso de minería. Los métodos para la selección de variables son básicamente dos:

---

(1) Técnica que consiste en agrupar los datos en “clusters” de similares características previo su procesamiento.

- Basados en la elección de los mejores atributos del problema.
- Aquellos que buscan variables independientes mediante pruebas de sensibilidad, algoritmos de distancia o heurísticos (2).

Extracción del Conocimiento

Mediante un método de minería de datos, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un preprocesado diferente de los datos.

Interpretación y Evaluación

Una vez obtenido el modelo, debe proceder a su validación, comprobando que las conclusiones que arroja son válidas.

---

(2) Pruebas donde las soluciones se descubren por la evaluación del progreso logrado en la búsqueda del resultado final.

En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, deben comparar los modelos en busca de uno que se ajuste mejor al problema. Según el texto, Introducción a la Minería de Datos, un conocimiento extraído para que sea deseable debe cumplir con las siguientes propiedades:

- **Válido:** hace referencia a que los patrones deben seguir siendo precisos para datos nuevos y no sólo para aquellos que han sido usados en su obtención.
- **Novedoso:** que aporte algo desconocido tanto para el sistema y preferiblemente para el usuario.
- **Potencialmente útil:** la información debe conducir a acciones que reporten algún tipo de beneficio para el usuario.
- **Comprensible:** la extracción de patrones no comprensibles dificulta o imposibilita su interpretación, revisión, validación y uso en la toma de decisiones.

Así, los sistemas de descubrimiento de conocimiento permiten la selección, limpieza, transformación y proyección de los datos; analizar los datos para extraer patrones y modelos adecuados; evaluar los

patrones para convertirlos en conocimiento; consolidar el conocimiento resolviendo posibles conflictos con conocimiento previamente extraído; y hacer el conocimiento disponible para su uso.

Podemos entonces decir, que el KDD es el proceso global de descubrir conocimiento útil desde las bases de datos mientras que la minería de datos se refiere a la aplicación de los métodos de aprendizaje y estadísticos para la obtención de patrones y modelos.

1.4 Métodos de Minería de Datos

Brevemente describiremos los diferentes métodos existentes para llevar a acabo un problema de minería de datos, los métodos son los siguientes:

- **Técnicas algebraicas y estadísticas:** se basan, generalmente, en expresar modelos y patrones mediante fórmulas algebraicas, funciones lineales, funciones no lineales, distribuciones o valores agregados estadísticos tales como medias, varianzas, correlaciones, entre otros. Frecuentemente, estas técnicas, cuando obtienen un patrón, lo hacen a partir de un modelo ya predeterminado del cual, se estiman unos coeficientes o parámetros.

- **Técnicas bayesianas:** se basan en estimar la probabilidad de pertenencia (a una clase o grupo), mediante la estimación de las probabilidades condicionales inversas o “a priori”, utilizando para ello el Teorema de Bayes. Algunos algoritmos muy populares son el clasificador bayesiano naive, los métodos basados en máxima verisimilitud y el algoritmo EM. Las redes bayesianas generalizan

las topologías de las interacciones probabilísticas entre variables y permiten representar gráficamente dichas interacciones.

- **Técnicas basadas en conteos de frecuencias y tablas de contingencia:** estas técnicas se basan en contar la frecuencia en la que dos o más sucesos se presenten conjuntamente. Cuando el conjunto de sucesos posibles es muy grande, existen algoritmos que van comenzando por pares de sucesos e incrementando los conjuntos sólo en aquellos casos que las frecuencias conjuntas superen un cierto umbral.

- **Técnicas basadas en árboles de decisión y sistemas de aprendizaje de reglas:** son técnicas que, además de su representación en forma de reglas, se basan en dos tipos de algoritmos: los llamados “divide y vencerás”, como el ID3 / C4.5 o el CART, y los denominados “separa y vencerás”, como el CN2.

- **Técnicas relacionales, declarativas y estructurales:** la característica principal de este conjunto de técnicas es que representan los modelos mediante lenguajes declarativos, como

los lenguajes lógicos, funcionales o lógico-funcionales. Las técnicas de programación lógica inductiva son las más representativas y las que han dado nombre a un conjunto de técnicas llamadas minería de datos relacional.

- **Técnicas basadas en redes neuronales artificiales:** se trata de técnicas que aprenden un modelo mediante el entrenamiento de los pesos que conectan un conjunto de nodos o neuronas. La topología de la red y los pesos de las conexiones determinan el patrón aprendido. Existen innumerables variantes de organización: perceptrón simple, redes multicapa, redes de base radial, el más conocido es el de retropropagación.

- **Técnicas estocásticas y difusas:** bajo este grupo se incluyen la mayoría de las técnicas que, junto a las redes neuronales, forman lo que se denomina computación flexible. Son técnicas en las que o bien los componentes aleatorios son fundamentales, como los métodos evolutivos y genéticos o bien al utilizar funciones de pertenencia difusas.

Este proyecto se enfocará en las técnicas bayesianas, en el siguiente capítulo se dará una amplia descripción.

1.5 Aplicaciones

La integración de las técnicas de minería de datos en las actividades del día a día se está convirtiendo en algo habitual. Los negocios de la distribución y la publicidad dirigida han sido tradicionalmente las áreas en las que más se han empleado los métodos de minería, ya que han permitido reducir costes o aumentar la receptividad de ofertas. Pero éstas no son las únicas áreas a las que pueden aplicar. Es más, podemos encontrar ejemplos en todo tipo de aplicaciones, tales como:

Aplicaciones financieras y banca:

- Obtención de patrones de uso fraudulento de tarjetas de crédito.
- Determinación del gasto en tarjeta de crédito por grupos.
- Análisis de riesgos en créditos.

Seguros y salud privada:

- Predicción de qué clientes contratan nuevas pólizas.
- Identificación de comportamiento fraudulento.
- Análisis de procedimientos médicos solicitados conjuntamente.

Educación:

- Selección o captación de estudiantes.
- Detección de abandonos y de fracaso.
- Estimación del tiempo de estancia en la institución.

Medicina:

- Identificación de patologías. Diagnóstico de enfermedades.
- Detección de pacientes con riesgo de sufrir una patología.
- Recomendación priorizada de fármacos para una patología.

Telecomunicaciones:

- Establecimiento de patrones de llamadas.
- Modelos de carga en redes.
- Detección de fraude.

Otras áreas:

- Correo electrónico y agendas personales: detección de correo spam, gestión de avisos.
- Recursos humanos: selección de empleados.

- Tráfico: modelo de tráfico a partir de fuentes diversas: cámaras, GPS.
- Web: análisis de comportamiento de los usuarios, detección de fraude en el comercio electrónico, análisis de logs de un servidor Web.
- Policiales: identificación de posibles terroristas en un aeropuerto.

CAPÍTULO 2

2. MÉTODO BAYESIANO PARA LA MINERÍA DE DATOS.

2.1 Introducción

La teoría de la probabilidad y los métodos bayesianos son las técnicas más utilizadas en problemas de inteligencia artificial y minería de datos. Una de las principales ventajas de los métodos bayesianos sobre las demás técnicas de minería de datos es la posibilidad de obtener conocimientos con incertidumbre; esto se debe a que los métodos bayesianos se basan en la teoría de la probabilidad.

Este modelo permite un uso tanto descriptivo como predictivo. Con respecto a su uso como modelo predictivo, los algoritmos de aprendizaje de redes bayesianas realizan tareas de descubrimiento de relaciones de dependencia o independencia entre las variables.

Posteriormente, se pueden realizar trabajos de inferencia y evaluar distintas hipótesis sobre estas relaciones.

Las redes bayesianas son utilizadas como clasificadores; para esta tarea se construyen redes mediante la adición de restricciones al proceso de aprendizaje, esto determina el nivel de complejidad de la red. Los métodos bayesianos son relevantes en el aprendizaje automático y la minería de datos por razones fundamentales:

- Es un método práctico para realizar las inferencias a partir de datos, induciendo modelos probabilísticos que después serán usados para razonar (formular hipótesis) sobre nuevos valores observados. Además, permite calcular de forma explícita la probabilidad asociada a cada una de las hipótesis posibles, lo que constituye una gran ventaja sobre las demás técnicas.

- Facilitan un marco de trabajo útil para la comprensión y análisis de numerosas técnicas de aprendizaje y minería de datos que no trabajan explícitamente con probabilidades.

Los métodos bayesianos, sin embargo, se encuentran con el inconveniente del coste computacional que requiere. Por ello, es habitual formular ciertas suposiciones que permiten restringir la complejidad de los modelos a utilizar. Se puede pensar que esto limita la capacidad expresiva de los modelos resultantes; sin embargo, se ha demostrado que a pesar de las simplificaciones realizadas por algunos de estos modelos (como el clasificador Naive Bayes) son realmente competitivos e incluso superan a otras técnicas complejas en determinadas aplicaciones.

2.2 Teorema de Bayes e Hipótesis MAP

El Teorema de Bayes es la regla básica para realizar inferencias. Nos permite evaluar la creencia que tenemos de un suceso o conjunto de sucesos a la luz de nuevos datos u observaciones, es decir, nos permite pasar de una probabilidad a priori P (suceso) a la probabilidad a P (suceso | observaciones).

La probabilidad a priori es aquella que fijamos inicialmente a un evento, sin conocer nada más. Por ejemplo, la probabilidad de que un estudiante de la FIEC se registre en un paralelo de una materia

determinada, podría ser P (Registro = sí) = 0,46. La probabilidad a posteriori es la que obtendríamos tras conocer cierta información, por tanto, puede verse como un refinamiento de nuestro conocimiento. Por ejemplo, si sabemos que el estudiante pertenece a la carrera Ingeniería en Computación, entonces la probabilidad a posteriori sería prácticamente P (Registro = sí | Carrera = Computación) = 0,90 suponiendo que la materia pertenece al pensúm de la carrera Ingeniería en Computación y que la misma tiene una alta demanda. El teorema de bayes viene representado por la siguiente expresión:

$$P(h|O) = \frac{P(O|h) \cdot P(h)}{P(O)}$$

FÓRMULA 2.1 TEOREMA DE BAYES

Donde, como podemos ver, lo que aparecen son la probabilidad a priori de la hipótesis (h) y de las observaciones (O) y las probabilidades condicionadas P (h|O) y P (O|h). A esta última se le conoce como verosimilitud de que la hipótesis h haya producido el conjunto de observaciones (O).

Centrándonos en el problema de la clasificación, con una variable clase C y un conjunto de variables predictoras $\{A_1,..., A_n\}$, el teorema de bayes tendría la siguiente forma:

$$P(C| A_1, A_n) = \frac{P(A_1, A_n | C) \cdot P(C)}{P(A_1, A_n)}$$

FÓRMULA 2.2 REDEFINICIÓN DEL TEOREMA DE BAYES

Evidentemente, si C tiene k valores posibles $\{c_1,..., c_k\}$, lo que nos interesa es identificar el más plausible y devolverlo como resultado de la clasificación. En el marco bayesiano, la hipótesis más plausible no es otra que aquella que tiene máxima probabilidad a posteriori dados los atributos, y es conocida como la hipótesis máxima a posteriori o hipótesis MAP (del inglés, maximum a posteriori). El teorema de bayes facilita un método sencillo para encontrar esta probabilidad.

Sin embargo, el método bayesiano tiene un problema, y es su altísima complejidad computacional, debido a que necesitamos trabajar con distribuciones de probabilidad que involucren muchas variables, haciéndolas en la mayoría de los casos inmanejables. Este problema puede ser superado haciendo uso de las (a veces supuestas) independencias entre las variables.

$$C_{MAP} = \arg_{c \in \Omega c} \max P\,(A_{i,}\, An|c)\, P(c)$$

FORMULA 2.3 EXPRESIÓN PARA OBTENER LA HIPÓTESIS MAP

2.3 Modelo Naive-Bayes de clasificación con Redes Bayesianas

Un clasificador es un mapeo realizado desde el espacio de un atributo X (discreto o continuo) hacia un conjunto discreto de etiquetas Y.

Naive Bayes es el modelo más simple de clasificación con redes bayesianas, ya que asume independencia entre todos los atributos dada una clase. NB es, por tanto, un modelo de atributos independientes. En este caso, la estructura de la red es fija y sólo es necesario aprender los parámetros (distribución de probabilidades). El fundamento principal del clasificador NB es la suposición de que todos los atributos son independientes conocido el valor de la variable clase.

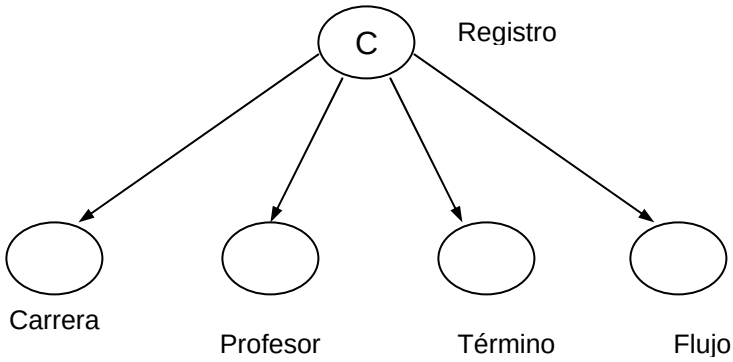


FIGURA 2.1 MODELO NB EN EL REGISTRO DE UNA MATERIA

A pesar de que asumir esta suposición en el clasificador NB es sin duda bastante fuerte y poco realista en la mayoría de los casos, se trata de

uno de los clasificadores más utilizados. Además, diversos estudios demuestran que sus resultados son competitivos con otras técnicas, particularmente en los casos en que existe completa independencia entre los atributos (igual que la suposición básica de NB), y también el caso de dependencias funcionales entre los atributos (lo que resulta un poco menos obvio).

Aplicando Naive Bayes al registro académico, la variable clase (Registro) se califica con los valores Si y No. Los atributos están conformados por el conjunto de datos académicos tales como:

- Profesor con quien se ofrece el paralelo.
- Carrera del estudiante de la FIEC que puede registrarse en el paralelo.
- Término en que se oferta el paralelo.

Al considerar que los atributos son independientes, es decir, el hecho que no existir una relación entre sí con las variables carrera, término y profesor, el problema se vuelve mucho más sencillo.

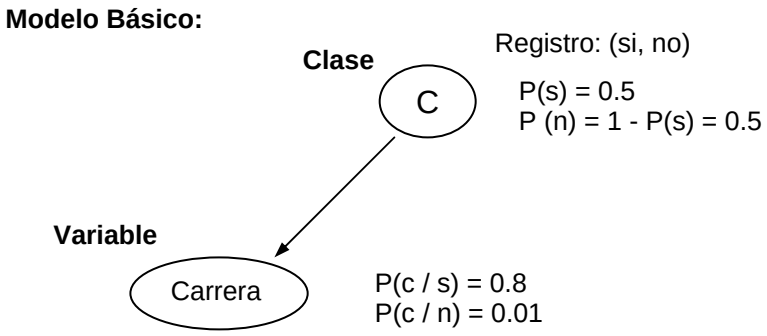


FIGURA 2.2 ILUSTRACIÓN DEL EJEMPLO: REGISTRO DE UN ESTUDIANTE DE LA FIEC EN UNA MATERIA

$$\text{Regla Bayes: } P(s|c) = \frac{P(c|s)\,P(s)}{P(c)} = \frac{P(c|s)\,.\,P(c)}{P(c|s)\,.\,P(c) + P(c|n)\,.\,P(n)}$$

FÓRMULA 2.4 TEOREMA DE BAYES CONSIDERANDO ATRIBUTOS INDEPENDIENTES

2.3.1 Estimación de los parámetros

La forma en que se debe proceder depende de que el atributo sea discreto o continuo, en nuestro proyecto se usan atributos discretos. Dentro de este tipo de atributo, la estimación de la probabilidad condicional se basa en las frecuencias de aparición que obtendremos en las base de datos.

Así, si llamamos n (x_i, Pa (x_i)) al número de registros de la base de datos en que la variable x_i toma el valor x_i y los padres de x_i

(Pa (x_i)) toman la configuración denotada por Pa (x_i), entonces la forma más simple de estimar P (x_i, Pa (x_i)) es:

$$P(x_i, Pa(x_i)) = \frac{n(x_i, Pa(x_i))}{n(Pa(x_i))}$$

FÓRMULA 2.5 ESTIMADOR USADO EN ATRIBUTOS DISCRETOS

Es decir el número de casos favorables dividido por el número de casos totales. Esta técnica se conoce como estimación por máxima verosimilitud y tiene como desventajas que necesita una muestra de gran tamaño y que sobre ajusta a los datos. Existen otros estimadores más complejos que palian estos problemas, entre ellos citaremos el estimador basado en la Ley de la Sucesión de Laplace:

$$P(x_i, Pa(x_i)) = \frac{n(x_i, Pa(x_i))+1}{n(Pa(x_i)) + |\Omega_{xi}|}$$

FÓRMULA 2.6 ESTIMADOR BASADO EN LEY DE LA SUCESIÓN DE LAPLACE

Es decir, el número de casos favorables más uno dividido por el número de casos totales más el número de valores posibles. Esta estimación pretende que todas las configuraciones posibles tengan cuando menos una mínima probabilidad.

Para entender mejor, decimos que la variable X_i es la variable “Profesor” y Pa (x_i) solo presenta a la variable “Registro” como padre de la variable “Profesor”, entonces para realizar P (x_i, Pa (x_i)), es decir, P (profesor, registro) utilizamos la fórmula 2.5 para calcular la distribución de probabilidad. Donde la variable “Profesor” toma cada uno de los valores posibles (nombre de los profesores), haciendo fijo el valor que toma la variable “Registro”.

2.4 Definición formal de las Redes Bayesianas

Las Redes Bayesianas (RBs) representan un modelo del conocimiento con incertidumbre. Ha demostrado ser una herramienta muy poderosa en diversos campos de investigación: la teoría de toma de decisiones, estadística e inteligencia artificial [Introducción a la Minería de Datos]. Entre sus principales aplicaciones, destacan la recuperación de información, la medicina, la visión artificial, la agricultura y con este trabajo queremos aportar novedades en el campo académico.

Una red bayesiana es un grafo dirigido y acíclico cuyos nodos representan variables y cuyos arcos representan relaciones de dependencia entre sus variables. Si existe un arco desde un nodo A hacia algún otro nodo B, entonces decimos que A es padre de B. Si un nodo posee un valor conocido, se dice que éste es un nodo evidencia.

Las preguntas respecto a la dependencia entre las variables pueden ser respondidas estudiando el grafo. Se puede demostrar que la noción llamada separación-d corresponde a la noción llamada independencia condicional: Si los nodos X y Y están separados-d (habiendo sido especificados los nodos evidencia), entonces las variables X y Y son

independientes dadas las variables evidencia. Es muy común trabajar con distribuciones Discretas o Gaussianas, ya que simplifican los cálculos.

Las RBs no sólo modelan de forma cualitativa el conocimiento sino que además expresan de forma numérica la “intensidad” de las relaciones entre sus variables. Esta parte cuantitativa del modelo suele expresarse en forma de distribuciones de probabilidad.

De manera formal, una red bayesiana es una tupla $B = (G,\Theta)$, donde G es el grafo y Θ es el conjunto de distribuciones de probabilidad $P(x_i, Pa(x_i))$ para cada variable desde $i = 1$ hasta n y $Pa(x_i)$ representa los padres de la variable X_i en el grafo G.

2.5 Aprendizaje de Redes Bayesianas

El aprendizaje de las redes bayesianas consiste en la construcción del modelo de la red. Es posible que un experto en el campo o negocio, construya el modelo de la red partiendo de su conocimiento del problema. Ya que el volumen de datos que se maneja durante la tarea de construcción del modelo de la red es grande, resulta de mucho interés proporcionar a estos expertos herramientas que adquieran este tipo de conocimiento automáticamente a partir de los datos del problema [Introducción a la Minería de Datos].

El aprendizaje se puede definir de la siguiente forma: Encontrar el grafo dirigido acíclico G que mejor represente el conjunto de dependencias / independencias presentes en un conjunto de datos D. Para solucionar este problema desde el punto de vista bayesiano, es necesario calcular la probabilidad a posteriori de una red bayesiana en concreto, dado el conjunto de datos conocidos, es P (G|D). Una vez que sabemos cómo calcular esta probabilidad, tendremos una medida de adecuación de cada grafo (red bayesiana) a los datos de partida y por consiguiente podremos comparar, para quedarnos con la mejor, entre distintas redes bayesianas.

Debido a que en este proyecto de Tesis usamos el clasificador Naive Bayes, no necesitamos del aprendizaje ni de los algoritmos de búsqueda, porque el clasificador NB considera independiente a cada una de las variables del grafo. Sólo existe relación directa entre la variable padre y las demás variables del modelo, como lo muestra la figura 2.1.

2.5.1 Medidas Bayesianas

Sea G = (V,E) un grafo dirigido acíclico que representa una red bayesiana B = (G,Θ) donde Θ es el conjunto de parámetros definidos en la red, esto es, el conjunto de distribuciones de probabilidad condicional. Sean D los N casos de datos con la siguiente forma:

$$D = \begin{pmatrix} X_1[1] & X_2[1] & \dots & X_n[1] \\ . & . & . & . \\ . & . & . & . \\ X_1[1] & X_2[1] & \dots & X_n[1] \end{pmatrix}$$

FÓRMULA 2.7 CONJUNTO DE DISTRIBUCIONES DE PROBABILIDAD

El problema consiste en calcular la probabilidad P (B|D), es decir, la probabilidad de una red dados los datos. Aplicando la regla de Bayes [Probabilidad y Estadística para Ingenieros], obtenemos que:

$$P(G|D) = \frac{P(D|G) \cdot P(G)}{P(D)}$$

FÓRMULA 2.8 REGLA DE BAYES PARA LA RED DE DATOS

Como los datos son conocidos y constantes, entonces el término P (D) puede eliminarse de la expresión anterior. Además, una

red B posee dos componentes (parámetros y estructura o grafo) con lo que la expresión anterior se traduce en:

$$P(G|D) \propto P(G)\, L(D|G)$$

$$L(D|G) = \int_{\Theta} P(D|G,\Theta)\, P(\Theta|G)\, d\Theta$$

FÓRMULA 2.9 REDEFINICIÓN DE LA REGLA DE BAYES

Siendo L (D|G) la función denominada verosimilitud marginal. Para el caso discreto, asumiendo que las distribuciones son de la familia exponencial y en concreto para el caso multinomial y su conjugada, la fórmula anterior tiene una solución cerrada dando lugar a una medida bayesiana [Probabilidad y Estadística para Ingenieros], conocida como:

$$L(D|G) = \prod_{i=1}^{n} \prod_{i=1}^{q} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}$$

FÓRMULA 2.10 EXPRESIÓN PARA LA VEROSIMILITUD MARGINAL

Siendo Γ la función gamma y α_{ijk} los hiperparámetros de la distribución a priori. Si estos hiperparámetros se suponen uniformes, entonces los α_{ijk} pueden igualarse a una constante,

denominada tamaño muestral equivalente. Además, N_{ijk} es la frecuencia de aparición en los datos de la variable i-ésima con su K-ésimo valor (de los r posibles) y para la configuración j-ésima (de las q posibles) de sus padres en el grafo. Por último, tanto N_{ijk} como α_{ijk} es igual a su proyección sobre k, esto es: $N_{ij} = \sum_{k=1}^{r} N_{ijk}$.

2.5.2 Algoritmos de Búsqueda

Es posible plantear el problema del aprendizaje como un problema de optimización, esto es, determinar una métrica para calcular la adaptación de los datos a una red bayesiana y por tanto debemos encontrar la solución que maximice esta métrica. Ya que no es posible resolver el problema eficientemente de forma exacta, se han planteado algoritmos de búsqueda específicos o adaptación de algoritmos meta heurísticos para la resolución aproximada del aprendizaje estructural [Introducción a la Minería de Datos].

Algoritmo K2

Se puede considerar el primer algoritmo basado en búsqueda u optimización de una métrica bayesiana y muchos trabajos posteriores se han basado en él.

Este algoritmo utiliza un esquema voraz en su búsqueda de soluciones candidatas cada vez mejores y parte de que las variables de entrada están ordenadas, de forma que los posibles padres de una variable aparecen en el orden antes que ella misma. Esta restricción es bastante fuerte pero fue un estándar en el origen de los primeros trabajos sobre aprendizaje de redes bayesianas.

Algoritmo B

Al igual que el algoritmo K2, el algoritmo B se basa en un esquema voraz para la construcción de una solución aproximada a partir de la rede vacía de enlaces (cada nodo posee inicialmente un conjunto vacío de padres). A diferencia del algoritmo previo, este algoritmo no impone la restricción de

proporcionarle como entrada un orden específico entre las variables.

Algoritmo HC

Es un algoritmo local de ascensión de colinas (Hill Climbing) por el máximo gradiente basado en la definición de una vecindad. El algoritmo parte de una solución inicial, como puede ser la red de enlaces, u otra cualquiera (si se trata de una tarea de clasificación, se puede inicializar de una estructura NB). A partir de esta solución se calcula el nuevo valor de la métrica utilizada de todas las soluciones (grafos) vecinos a la solución actual y nos quedamos con el vecino que mejor valor de la métrica resulte.

Clasificadores como TAN, BAN necesitan obligatoriamente de los algoritmos de búsqueda porque realizan el aprendizaje de la red. La única característica en común que posee el clasificador NB frente a los demás es que todos parten de un modelo con las variables independientes entre sí, excepto con la variable padre.

CAPÍTULO 3

3. ANÁLISIS DEL PROBLEMA DE REGISTROS ACADÉMICOS.

3.1 Antecedentes

La Facultad de Ingeniería en Electricidad y Computación (FIEC) tiene como misión el formar profesionales de excelencia tanto a nivel de pregrado como de postgrado en las áreas relacionadas a Ingeniería Electrónica, Ingeniería en Electricidad, Computación, Telemática.

La FIEC a lo largo de su vida institucional ha buscado técnicas para mejorar la calidad de enseñanza, el control de profesores, el control de nuevos aspirantes, el proceso de graduación de los estudiantes y el control de su personal administrativo.

La Institución cuenta con sistemas de información confiables que colaboran con el desempeño de sus funciones prioritarias y la eficacia de los servicios institucionales. Son sistemas referidos al manejo de las carreras y programas, a las características de sus alumnos, al manejo de calificaciones, entre otros.

El proceso de planificación de apertura de paralelos está asociado a la gestión de los registros académicos que en la actualidad es un proceso fundamental de la Facultad e involucra un considerable manejo y circulación de información como parte de los procedimientos de los registros académicos.

3.2 Planteamiento del Problema

La Facultad ha hecho grandes esfuerzos para hacer coincidir la apertura de los paralelos con la demanda real, en varios semestres existían paralelos donde 8, 9 o 10 estudiantes se habían registrado, en algunos casos el paralelo no tenía estudiantes y permanecía abierto esperando llenarse.

En otro caso, se necesitaba abrir otro paralelo debido a la gran cantidad de estudiantes que deseaban inscribirse en el curso, todos estos son hechos controlables y sin mucha trascendencia. Sin embargo el peor caso que ha ocurrido es cuando el paralelo no cumple el mínimo de estudiantes inscritos y por ende se lo cierra; eso genera una desviación en el desarrollo académico del estudiante que se encontraba registrado en aquel paralelo, porque lo retrasa en el flujo de su carrera o porque descoordina el horario planificado para su semestre.

De acuerdo a lo mencionado, el problema se encuentra en la planificación de apertura de paralelos, ya que no existe una herramienta que permita equilibrar la predicción sobre datos pasados contra la demanda real.

3.3 Análisis del Modelo de Toma de Decisiones

Para el problema en la planificación de apertura de paralelos en un semestre académico, la toma de decisión sobre “¿Cuántos paralelos deben abrirse?” se verifica mediante datos de semestres anteriores, tales como:

- Alumnos que aprobaron la(s) materia(s) que es (son) requisito(s).
- Alumnos que cumplen con otros requisitos, como el nivel de la carrera u aprobación de cursos especiales.

Sin embargo, los resultados que aparecen después de este pequeño análisis de variables, no tienen un nivel de certeza alto. Esto se debe a la inconsistencia de los datos, campos en blanco o datos mal ingresados.

También se usan datos que ocurrieron en el mismo semestre pero en años pasados, para comparar y ver cual es la tendencia que los estudiantes tienen al registrarse en una materia.

Como alternativa pueden medirse otras variables y calcular a partir de ellas mediante un modelo adecuado la predicción deseada. No obstante, en muchas ocasiones, no se dispone de dicho modelo porque se conocen sólo las principales variables de entrada, pero no las relaciones existentes entre ellas. Para este tipo de problemas, la minería de datos ayuda a encontrar un modelo que represente una aproximación de las relaciones con un grado de probabilidad.

En todo proceso de minería de datos, el descubrimiento automático de hechos e hipótesis ocultas o no explícitas es un acontecimiento que busca una diferencia competitiva, junto con un aumento de la eficacia y productividad de la facultad. Pero el hecho de identificar qué tipo de conocimiento (tácito o explícito) se ha descubierto, cómo se ha generado, y de qué manera lo verificamos, interpretamos, interiorizamos y transmitimos, es lo que determina la magnitud de la minería de datos en la gestión del conocimiento y en el ciclo de conversión del conocimiento.

3.4 Casos de Uso

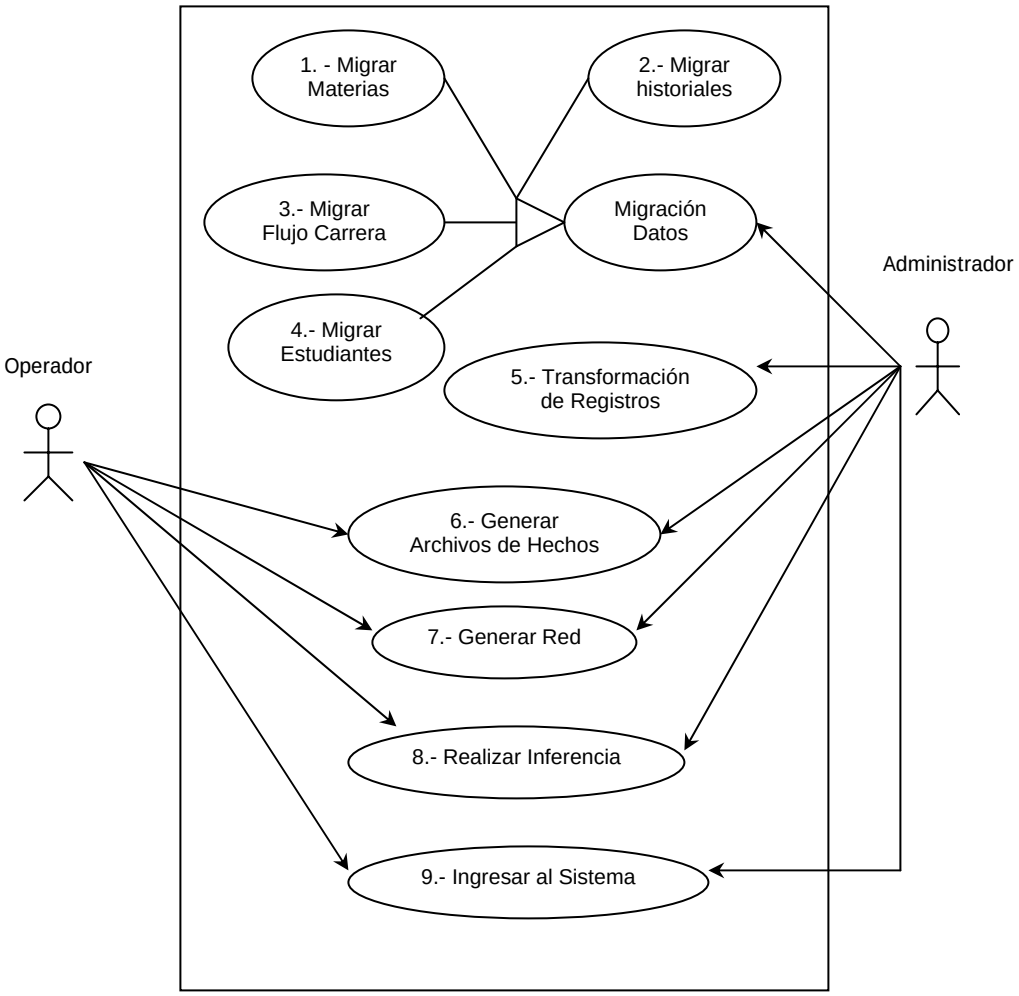


FIGURA 3.1 DIAGRAMA DE CASOS DE USO DEL SISTEMA

TABLA 1

CASO DE USO – MIGRAR MATERIAS

Actor: Administrador	
Descripción: permite migrar información a la base de datos sobre las materias que dicta la facultad, este caso se le permite acceder solo al administrador del sistema.	
Administrador	**Sistema**
Selecciona la opción Migrar datos	
Busca el archivo de materias para migrar	
Presiona el botón Ejecutar	Valida si el archivo es el correcto
	Divide el archivo en registros
	Inserta el registro en la tabla de materias
	Si existe error se hace rollback (3), caso contrario se hace commit (4)
	Se muestra un mensaje de retroalimentación al usuario

(3) Sentencia SQL que deshace la transacción actual.
(4) Sentencia SQL realiza la transacción actual.

TABLA 2

CASO DE USO – MIGRAR HISTORIALES

Actor: Administrador	
Descripción: permite migrar información a la base de datos sobre los historiales académicos de los estudiantes de la facultad, se debe notar que las materias dentro de los historiales, tienen que estar previamente ingresadas en la tabla de materias, este caso se le permite acceder solo al administrador del sistema.	
Administrador	**Sistema**
Selecciona la opción migrar datos	
Busca el archivo de historiales para migrar	
Presiona el botón Ejecutar	Valida si el archivo es el correcto
	Divide el archivo en registros
	Inserta el registro en la tabla de historiales
	Si existe error se hace rollback, caso contrario se hace commit
	Se muestra un mensaje de retroalimentación al usuario

TABLA 3

CASO DE USO – MIGRAR FLUJO DE CARRERA

Actor: Administrador	
Descripción: permite migrar información a la base de datos sobre los prerrequisitos y correquisitos de las diferentes materias que ofrece la facultad, se debe notar que las materias que se ingresen, tienen que estar previamente ingresadas en la tabla de materias, este caso se le permite acceder solo al administrador del sistema.	
Administrador	**Sistema**
Selecciona la opción migrar datos	
Busca el archivo de flujo de carrera para migrar	
Presiona el botón Ejecutar	Valida si el archivo es el correcto
	Divide el archivo en registros
	Inserta registros en la tabla flujo carrera
	Si existe error se hace rollback, caso contrario se hace commit
	Se muestra un mensaje de retroalimentación al usuario

TABLA 4

CASO DE USO – MIGRAR ESTUDIANTES

Actor: Administrador	
Descripción: permite migrar información a la base de datos sobre los estudiantes de la facultad, este caso se le permite acceder solo al administrador del sistema.	
Administrador	**Sistema**
Selecciona la opción migrar estudiantes	
Busca el archivo de estudiantes para migrar	
Presiona el botón Ejecutar	Valida si el archivo es el correcto
	Divide el archivo en registros
	Inserta registros en la tabla de estudiantes
	Si existe error se hace rollback, caso contrario se hace commit
	Se muestra un mensaje de retroalimentación al usuario

TABLA 5

CASO DE USO – TRANSFORMACIÓN DE REGISTROS

Actor: Administrador	
Descripción: encuentra conocimiento implícito (5) que existe en el historial académico y crea la tabla de hechos. Este caso colabora con el caso “Generar archivos de hechos”.	
Administrador	**Sistema**
Presiona el botón Ejecutar.	
	Muestra en pantalla: “Iniciando transformación”.
	Crea una tabla temporal a partir de una consulta.
	Borra la tabla de hechos existente.
	Renombra la tabla temporal con el nombre original.
	Muestra en pantalla: “Terminando transformación”.

(5) Información sobre estudiantes que pudieron registrarse en una materia en un año y término determinados.

TABLA 6

CASO DE USO – GENERAR ARCHIVOS DE HECHOS

Actor: Administrador / Operador	
Descripción: crea un conjunto de archivos que contiene: las variables que conformarán la red bayesiana y los valores que pueden tomar, un conjunto de casos que muestran la interacción entre las variables y un conjunto de casos de prueba para validar la red.	
Administrador / Operador	**Sistema**
Selecciona la materia.	
Presiona el botón Ejecutar.	
	Obtiene el conjunto de datos que resulta del proceso de transformación de registros.
	Construye los archivos .data, .name y .test a partir del conjunto de datos.
	Muestra retroalimentación al usuario.

TABLA 7

CASO DE USO – GENERAR RED

Actor: Administrador / Operador	
Descripción: construye una red bayesiana a partir de un conjunto de archivos de hecho.	
Administrador / Operador	**Sistema**
Selecciona la materia.	
Presiona el botón Ejecutar.	
	Construye la red.
	Genera el archivo .bif que contiene la red construida.
	Muestra retroalimentación al usuario.

TABLA 8

CASO DE USO – REALIZAR INFERENCIA

<table>
<tr><td colspan="2">Actor: Administrador / Operador</td></tr>
<tr><td colspan="2">Descripción: ayuda a la toma de decisiones sobre cuantos paralelos de una materia debe abrir la FIEC, mostrando la probabilidad de que los estudiantes se registren en la misma, este caso se le permite acceder tanto al administrador como al operador del sistema.</td></tr>
<tr><th>Administrador / Operador</th><th>Sistema</th></tr>
<tr><td>Selecciona la opción “Realizar Inferencia”.</td><td rowspan="3"></td></tr>
<tr><td>Selecciona la materia.</td></tr>
<tr><td>Presiona el botón Ejecutar.</td></tr>
<tr><td rowspan="3"></td><td>Ejecuta el editor de redes “JavaBayes” con el archivo de la materia seleccionada.</td></tr>
<tr><td>Muestra el gráfico de la red</td></tr>
<tr><td>Muestra las siguientes opciones que tiene el usuario</td></tr>
</table>

TABLA 9

CASO DE USO – INGRESAR AL SISTEMA

Actor: Administrador / Operador	
Descripción: permite al usuario ingresar al sistema y según el tipo de usuario se cargará el menú respectivo.	
Administrador / Operador	**Sistema**
Ingresa el nombre del servidor de Base de Datos	
Ingresa el nombre de usuario	
Ingresa la clave del usuario	
Presiona el botón Ingresar	Verifica si el servidor está en línea
	Verifica si el usuario existe
	Verifica si la clave le pertenece al usuario
	Carga el menú que compete al usuario

TABLA 10

ROLES DE LOS ACTORES DEL SISTEMA

Rol: Administrador
Descripción: tendrá habilitada todas las opciones del sistema, es la persona que le dará mantenimiento a la base de datos y podrá migrar nueva información a la misma y también realizar la transformación de registros.
Rol: Operador
Descripción: es la persona encargada de Generar archivos de hecho, Generar la red y Realizar la inferencia sobre cuantos paralelos debe la FIEC abrir.

3.5 Interacción de Objetos

Para analizar la interacción entre los objetos es necesario obtener los diferentes escenarios de los casos de uso del sistema a continuación se los detalla:

Escenario 1.1: migración de materias exitosa

Suposiciones

El administrador seleccionó la opción de Migrar Datos.

El administrador ingresó correctamente la ruta del archivo de materias.

El administrador pulsó el botón Ejecutar.

Resultados

El administrador recibe un mensaje de migración exitosa.

La información están debidamente almacenada en la base de datos.

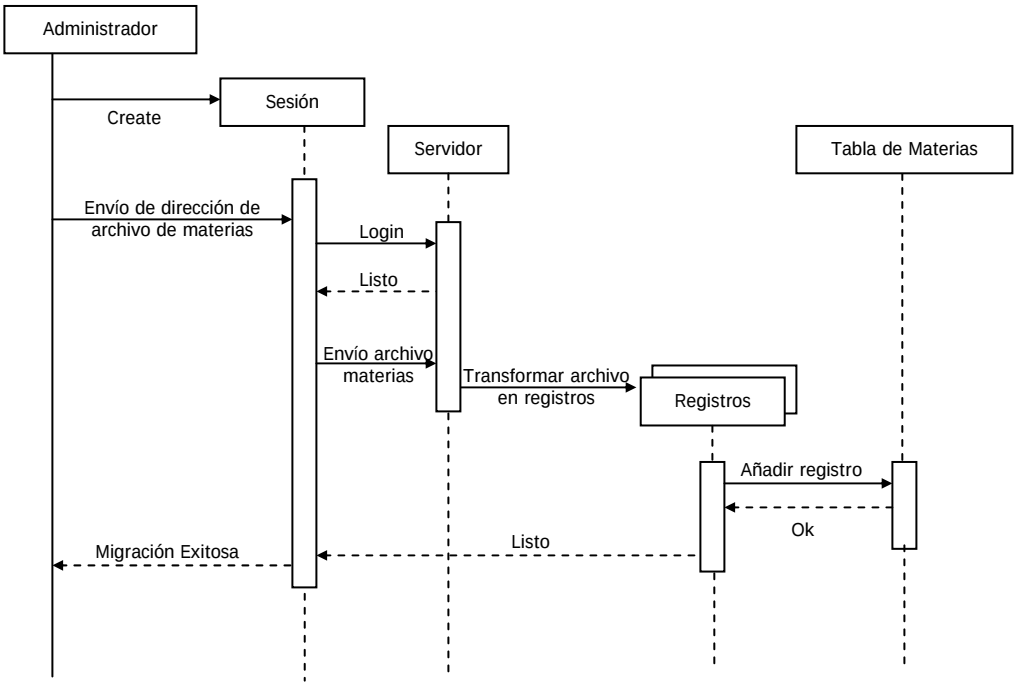


FIGURA 3.2 DIO PARA ESCENARIO 1.1

Escenario 1.2: migración de materias fallida, falla en la conexión a la base de datos.

Suposiciones

El administrador seleccionó la opción de Migrar Datos.

El administrador ingresó correctamente la ruta del archivo de materias.

La base de datos está en modo off - line.

El administrador pulsó el botón Ejecutar.

Resultados

El administrador recibe un mensaje de falla en la conexión con la base.

La información no se sube a la base de datos.

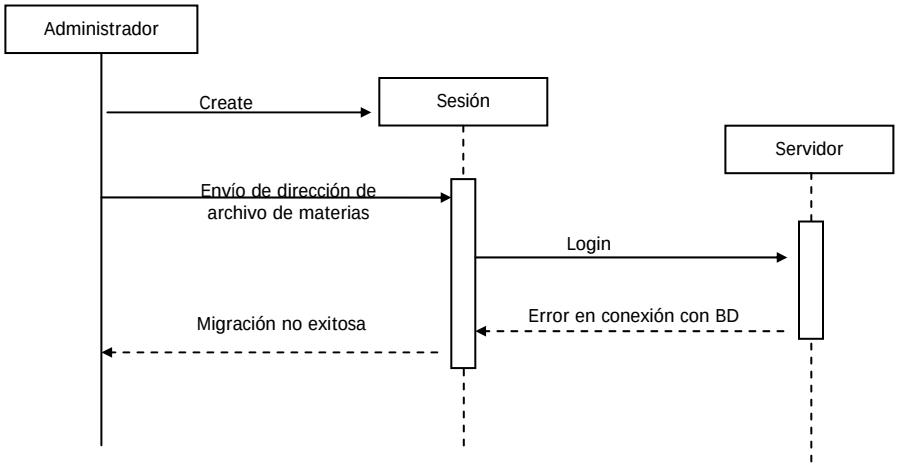


FIGURA 3.3 DIO PARA ESCENARIO 1.2

Escenario 1.3: migración de materias fallida, falla en inserción.

Suposiciones

El administrador seleccionó la opción de Migrar Datos.

El administrador ingresó correctamente la ruta del archivo de materias.

El archivo posee campos incorrectos.

El administrador pulsó el botón Ejecutar.

Resultados

El administrador recibe un mensaje de falla de inserción de datos.

La información no se sube a la base de datos.

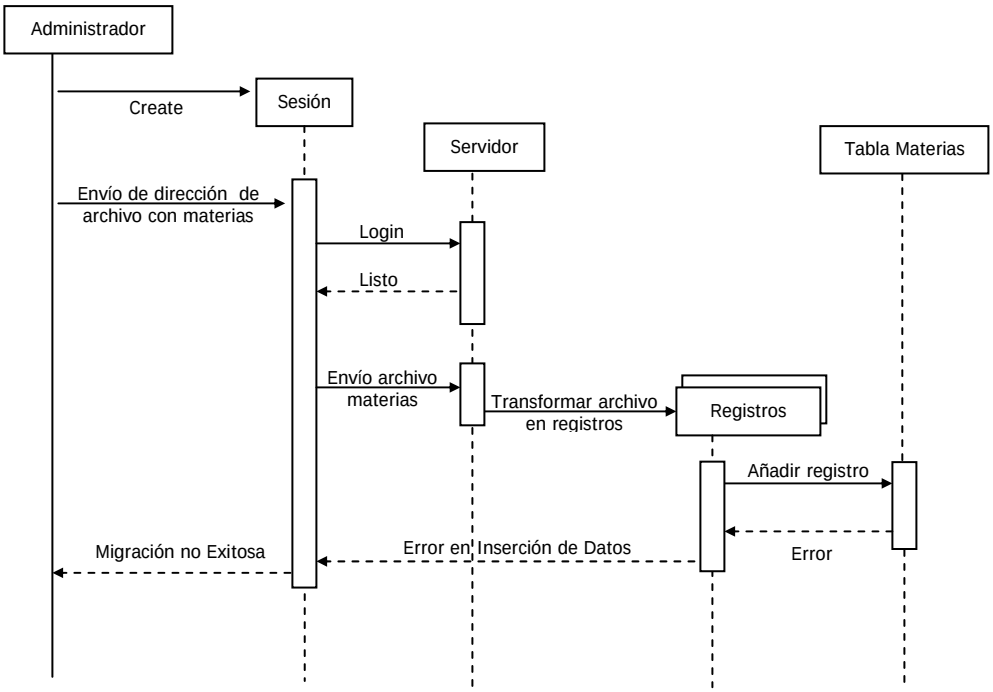


FIGURA 3.4 DIO PARA ESCENARIO 1.3

Escenario 2.1: migración de historiales exitosa

Suposiciones

El administrador seleccionó la opción de Migrar Datos.

El administrador ingresó la ruta del archivo de historiales.

El administrador pulsó el botón Ejecutar.

Resultados

El administrador recibe un mensaje de migración exitosa.

La información están debidamente almacenada en la base de datos.

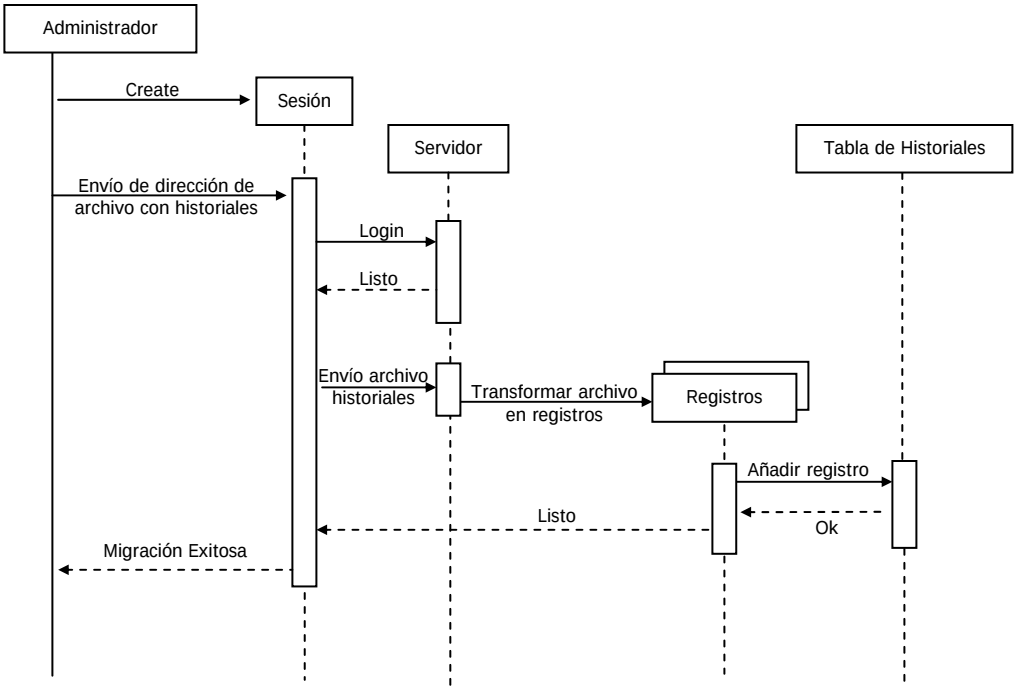


FIGURA 3.5 DIO PARA ESCENARIO 2.1

Escenario 2.2: migración de historiales fallida, falla en la conexión a la base de datos

Suposiciones

El administrador seleccionó la opción de Migrar Datos.

El administrador ingresó la ruta del archivo de historiales.

La base de datos está en modo off-line.

El administrador pulsó el botón Ejecutar.

Resultados

El administrador recibe un mensaje de falla en la conexión con la base.

La información no se sube a la base de datos.

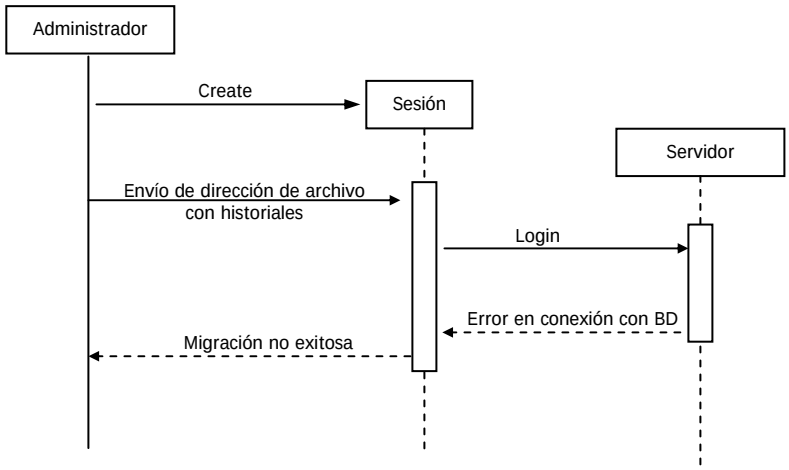


FIGURA 3.6 DIO PARA ESCENARIO 2.2

Escenario 2.3: migración de historiales fallida, falla en inserción.

Suposiciones

El administrador seleccionó la opción de Migrar Datos.

El administrador ingresó la ruta del archivo de historiales.

El archivo posee campos incorrectos.

El administrador pulsó el botón Ejecutar.

Resultados

El administrador recibe un mensaje de falla de inserción de datos.

La información no se sube a la base de datos.

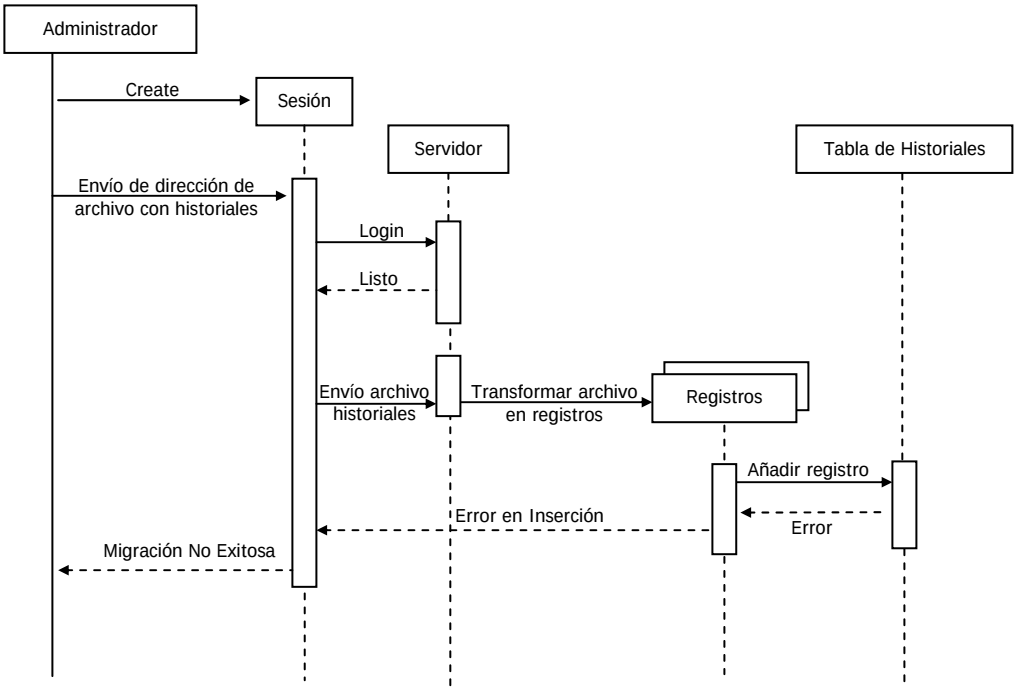


FIGURA 3.7 DIO PARA ESCENARIO 2.3

Escenario 3.1: migración de flujo de carrera exitosa

Suposiciones

El administrador seleccionó la opción de Migrar Datos.

El administrador ingresó la ruta del archivo de flujo de carrera.

El administrador pulsó el botón Ejecutar.

Resultados

El administrador recibe un mensaje de migración exitosa.

La información están debidamente almacenada en la base de datos.

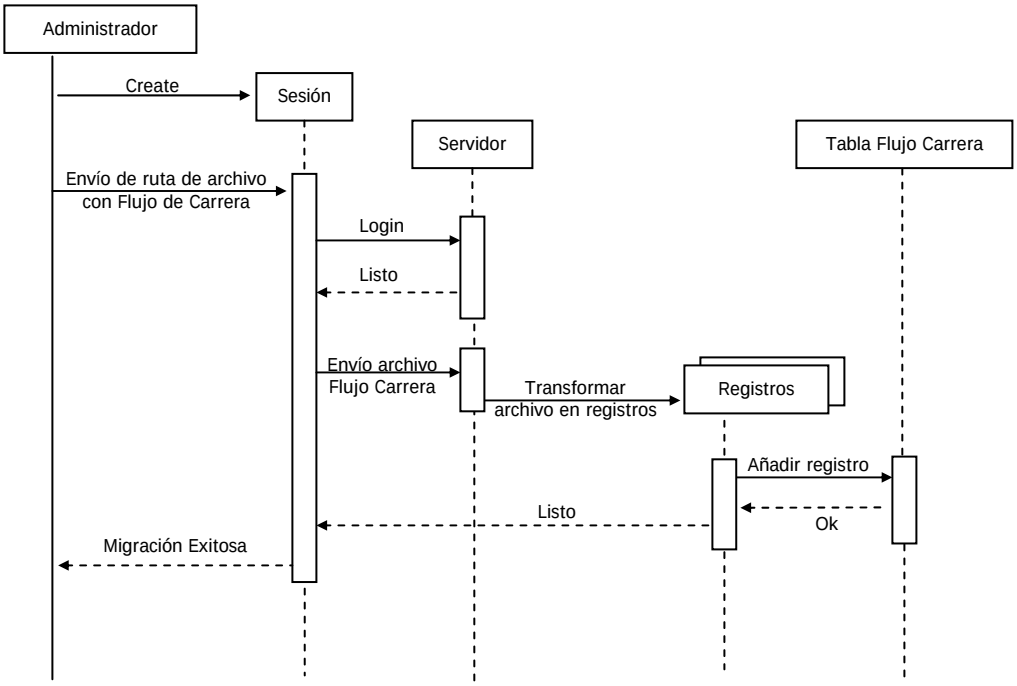


FIGURA 3.8 DIO PARA ESCENARIO 3.1

Escenario 3.2: migración de flujo de carrera fallida, falla en la conexión a la base

Suposiciones

El administrador seleccionó la opción de Migrar Datos.

El administrador ingresó la ruta del archivo de flujo de carrera.

La base de datos está en modo off-line.

El administrador pulsó el botón Ejecutar.

Resultados

El administrador recibe un mensaje de falla en la conexión con la base.

La información no se sube a la base de datos.

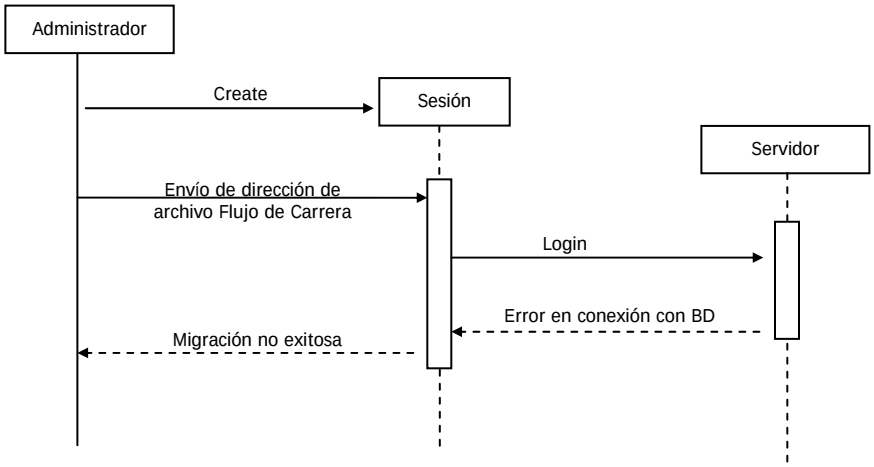


FIGURA 3.9 DIO PARA ESCENARIO 3.2

Escenario 3.3: migración de flujo de carrera, falla en inserción

Suposiciones

El administrador seleccionó la opción de Migrar Datos.

El administrador ingresó la ruta del archivo de flujo de carrera.

El archivo posee campos incorrectos.

El administrador pulsó el botón Ejecutar.

Resultados

El administrador recibe un mensaje de falla de inserción de datos.

La información no se sube a la base de datos.

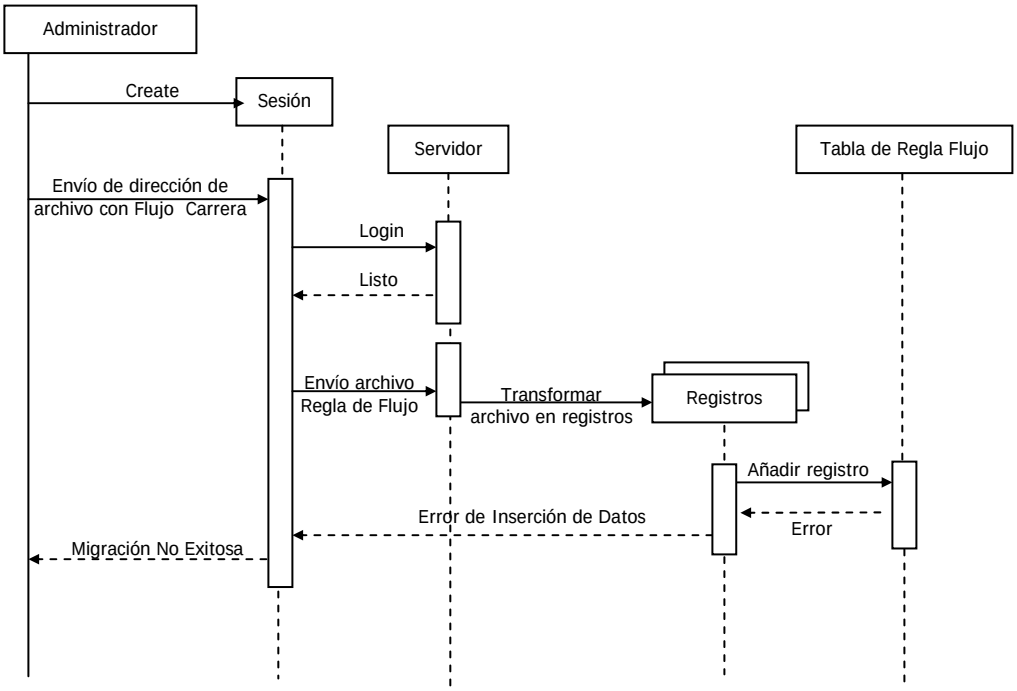


FIGURA 3.10 DIO PARA ESCENARIO 3.3

Escenario 4.1: migración de estudiantes exitoso

Suposiciones

El administrador seleccionó la opción de Migrar Datos.

El administrador ingresó la ruta del archivo de estudiantes.

El administrador pulsó el botón Ejecutar.

Resultados

El administrador recibe un mensaje de migración exitosa.

La información están debidamente almacenada en la base de datos.

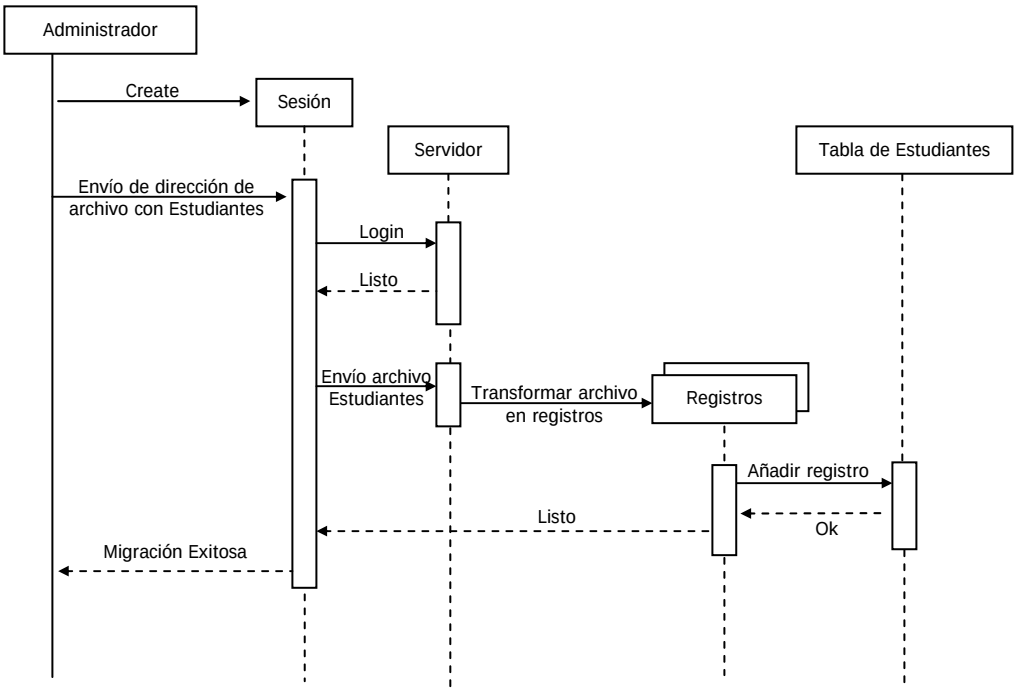


FIGURA 3.11 DIO PARA ESCENARIO 4.1

Escenario 4.2: migración de estudiantes fallido, error en conexión a la base

Suposiciones

El administrador seleccionó la opción de Migrar Datos.

El administrador ingresó la ruta del archivo de estudiantes.

La base de datos está en modo off-line.

El administrador pulsó el botón Ejecutar.

Resultados

El administrador recibe un mensaje de falla en la conexión con la base.

La información no se sube a la base de datos.

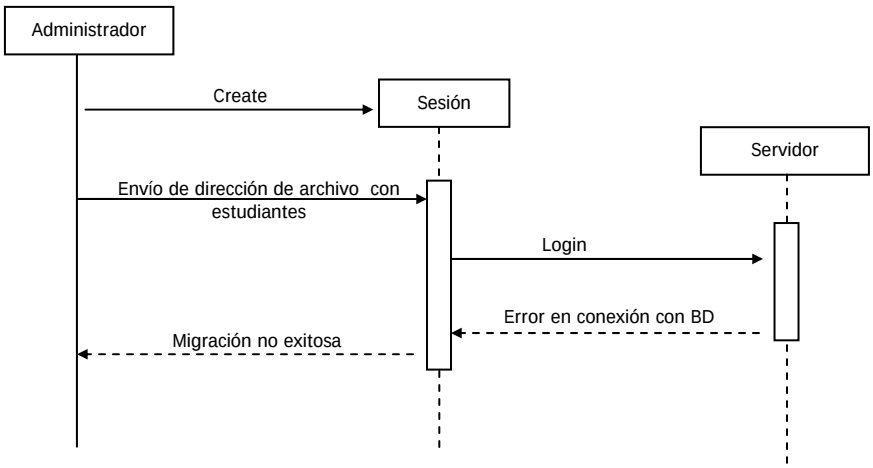


FIGURA 3.12 DIO PARA ESCENARIO 4.2

Escenario 4.3: migración de estudiantes, falla en inserción

Suposiciones

El administrador seleccionó la opción de Migrar Datos.

El administrador ingresó la ruta del archivo de estudiantes.

El archivo posee campos incorrectos.

El administrador pulsó el botón Ejecutar.

Resultados

El administrador recibe un mensaje de falla de inserción de datos.

La información no se sube a la base de datos.

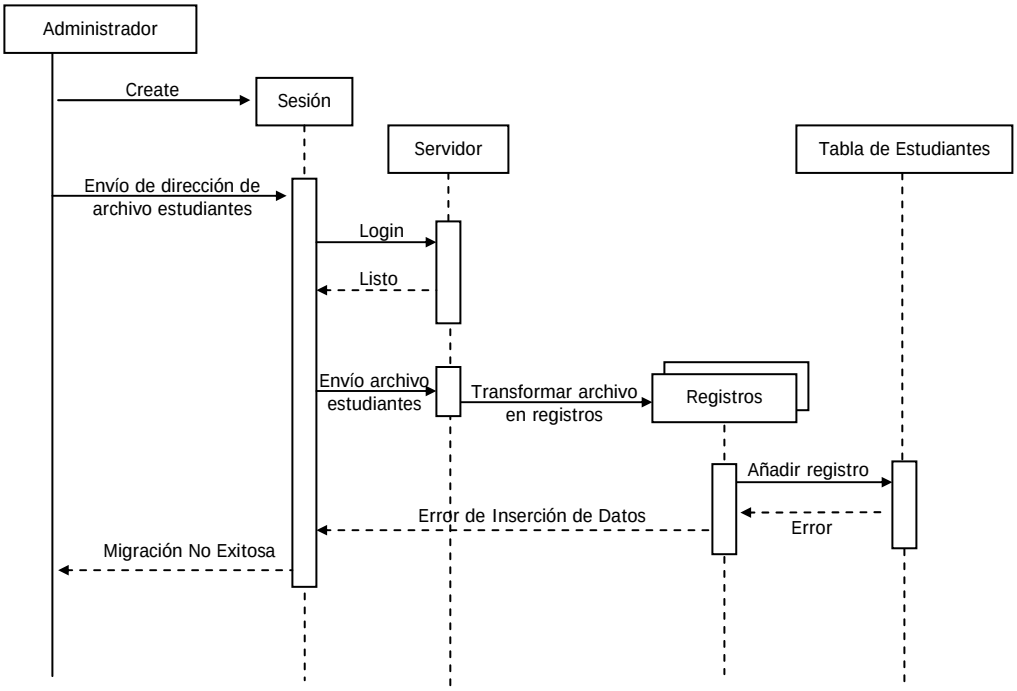


FIGURA 3.13 DIO PARA ESCENARIO 4.3

Escenario 5.1: transformación de registros exitosa

Suposiciones

El administrador seleccionó la opción Transformación de Registros.

Resultados

Se almacena los registros encontrados en la tabla de hechos.

El administrador recibe mensaje de transformación exitosa.

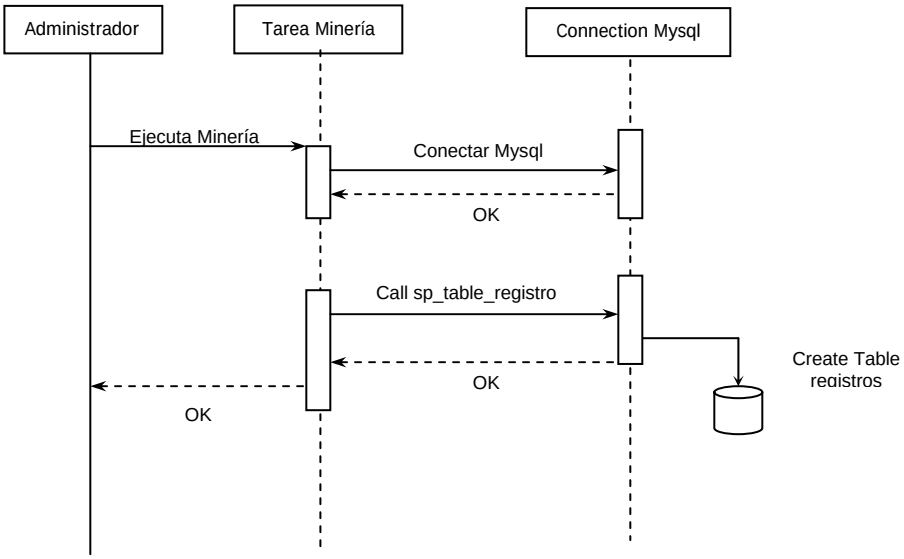


FIGURA 3.14 DIO PARA ESCENARIO 5.1

Escenario 6.1: generación de archivos de hechos exitosa

Suposiciones

El administrador / operador ingresó al sistema correctamente.

El administrador / operador selecciona la materia de la cual se desean los archivos de hechos.

El administrador / operador presiona el botón Ejecutar.

La tabla de registros académicos contiene datos.

Resultados

El sistema genera el archivo .name, .data y .test

El administrador / operador recibe mensaje de archivos creados.

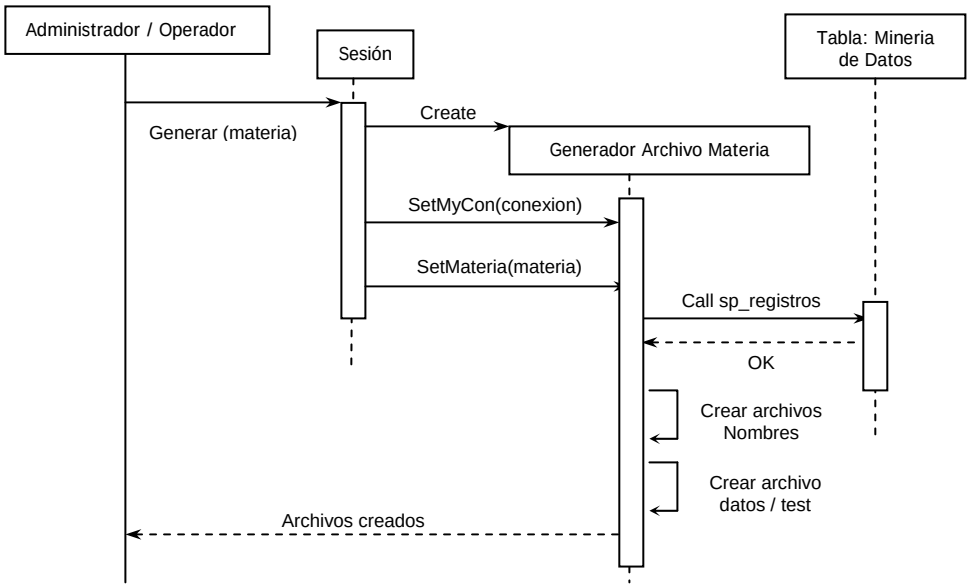


FIGURA 3.15 DIO PARA ESCENARIO 6.1

Escenario 7.1: generación de red exitosa

Suposiciones

El administrador / operador ingresó al sistema correctamente.

El administrador / operador selecciona la materia de la cual se desea crear la red.

El administrador / operador presiona el botón Ejecutar.

Los archivos .name, .data y .test existen.

Resultados

El sistema genera el archivo de red bayesiana .bif

El administrador / operador recibe mensaje de red creada con éxito.

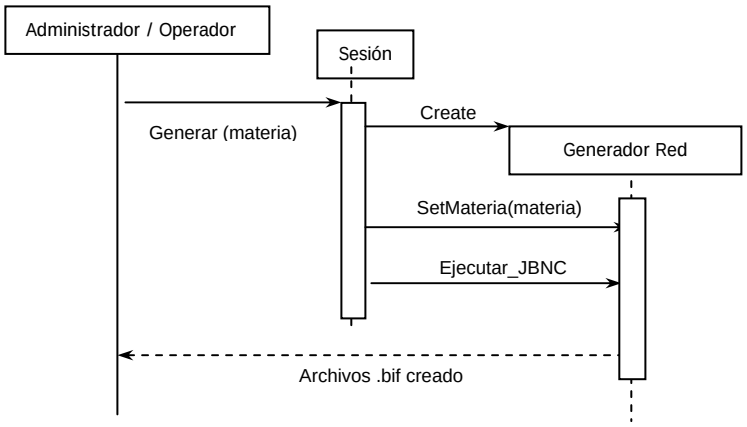


FIGURA 3.16 DIO PARA ESCENARIO 7.1

Escenario 8.1: inferencia sobre una materia exitosa

Suposiciones

El administrador / operador ingresó al sistema correctamente.

El administrador / operador seleccionó la opción Realizar Inferencia.

El administrador / operador seleccionó la materia.

El administrador / operador pulsó el botón “Ejecutar”.

Resultado

El sistema abre una ventana con el diagrama de Bayes y con las opciones para manipular el diagrama.

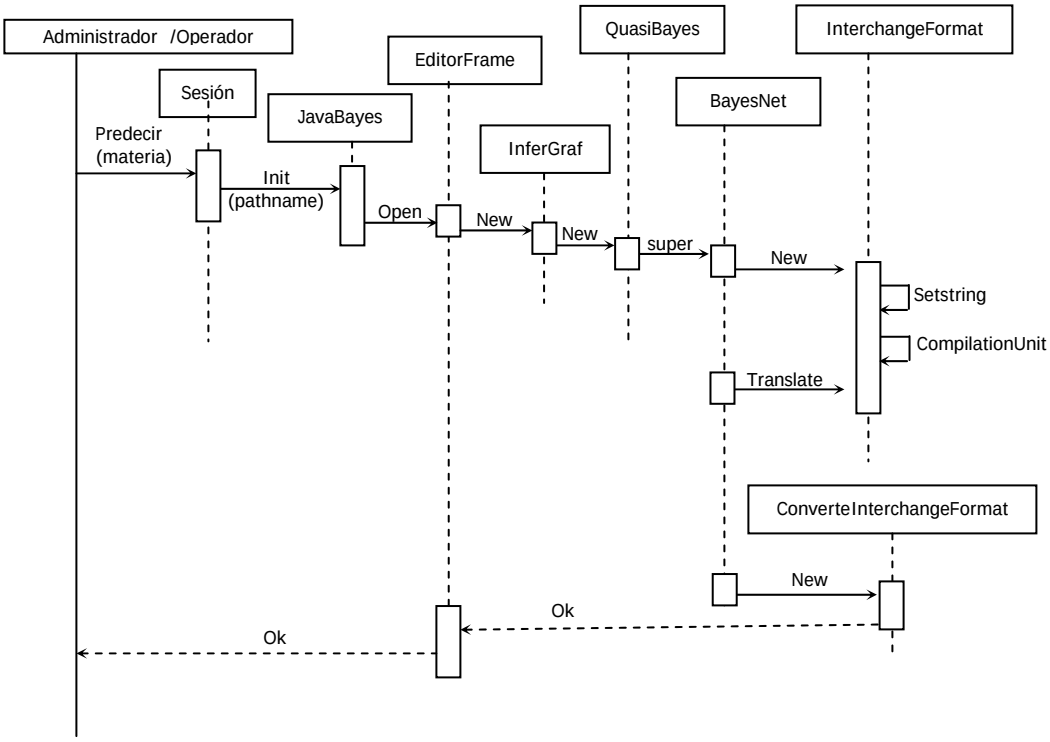


FIGURA 3.17 DIO PARA ESCENARIO 8.1

Escenario 8.2: inferencia sobre una materia, fallo en el archivo de red.

Suposiciones

El administrador / operador ingresó al sistema correctamente.

El administrador / operador seleccionó la opción Realizar Inferencia.

El administrador / operador seleccionó la materia.

El archivo de la red bayesiana de la materia escogida no existe.

El administrador / operador pulsó el botón Ejecutar.

Resultados

El administrador / operador recibe mensaje de error y pide que se cree el archivo de la red.

El sistema no procesa absolutamente nada.

Se permite al administrador / operador volver a ingresar una materia.

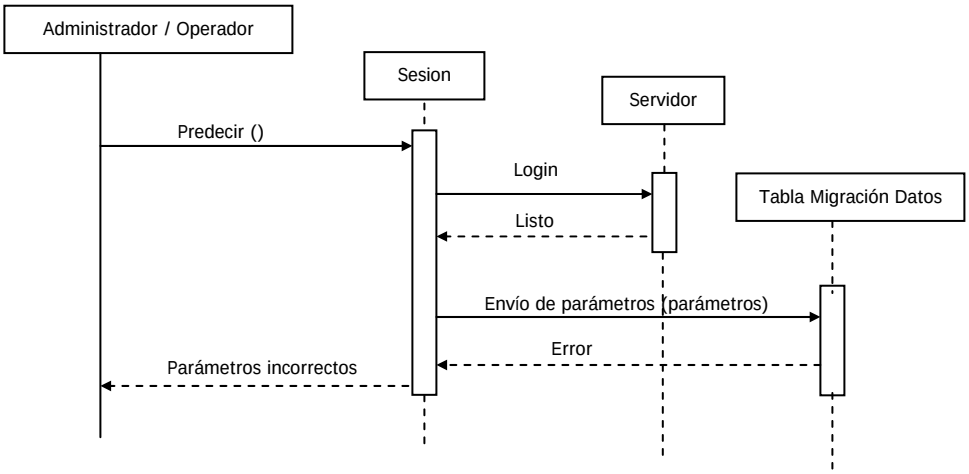


FIGURA 3.18 DIO PARA ESCENARIO 8.2

Escenario 8.3: inferencia sobre una materia, falla datos en el archivo de red

Suposiciones

El administrador / operador ingresó al sistema correctamente.

El administrador / operador seleccionó la opción Realizar Inferencia.

El administrador / operador seleccionó la materia.

El archivo de la red bayesiana de la materia escogida existe.

El archivo de la red bayesiana contiene incompatibilidad en sus datos.

El administrador / operador pulsó el botón Ejecutar.

Resultados

El administrador / operador recibe un mensaje de fallo formato del archivo de red bayesiana.

El sistema no procesa absolutamente nada.

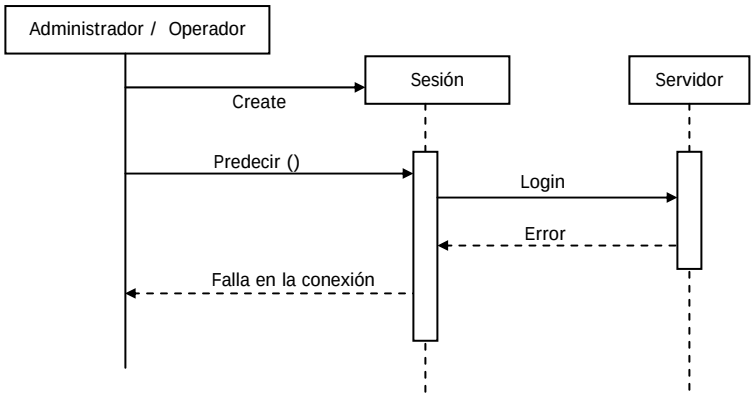


FIGURA 3.19 DIO PARA ESCENARIO 8.3

Escenario 9.1: ingreso exitoso al sistema

Suposiciones

El usuario ingresa el nombre del servidor de base de datos correcto.

El usuario ingresa el nombre de usuario correcto.

El usuario ingresa la clave correcta.

Resultado

Se muestra la ventana principal de la aplicación con el menú de acuerdo al tipo de usuario.

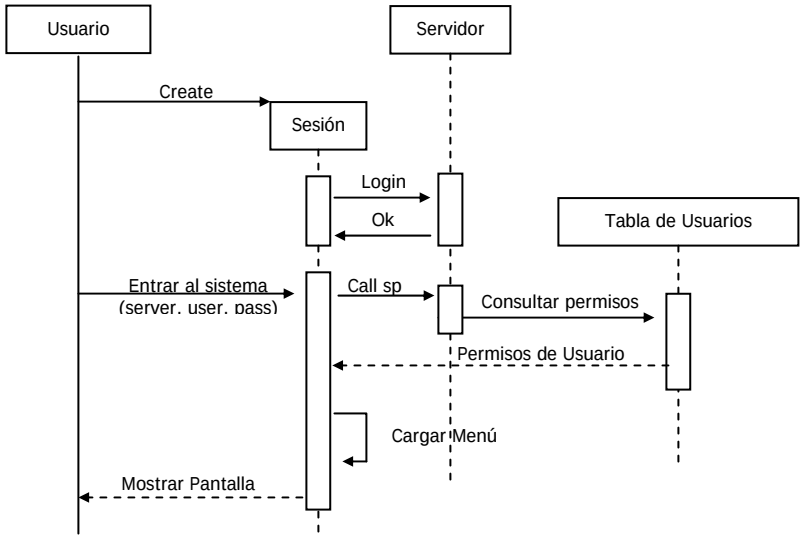


FIGURA 3.20 DIO PARA ESCENARIO 9.1

Escenario 9.2: ingreso fallido al sistema, falla en conexión a BD.

Suposiciones

El usuario ingresa el nombre del servidor de base de datos incorrecto.

El usuario ingresa el nombre de usuario correcto.

El usuario ingresa una clave correcta.

Resultado

El usuario recibe un mensaje de error en ingreso al sistema por falla en la conexión a la base de datos.

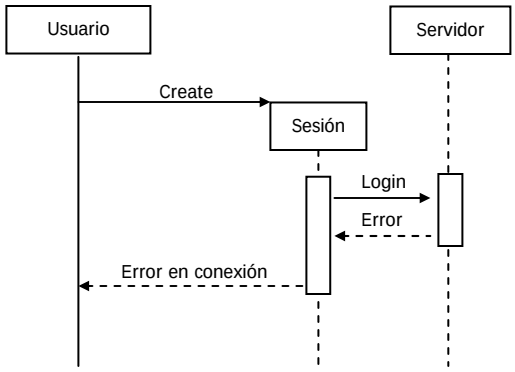


FIGURA 3.21 DIO PARA ESCENARIO 9.2

Escenario 9.3: ingreso fallido al sistema, nombre de usuario correcto pero la clave es fallida.

Suposiciones

El usuario ingresa el nombre del servidor de base de datos correcto.

El usuario ingresa el nombre de usuario correcto.

El usuario ingresa una clave incorrecta.

Resultado

El usuario recibe un mensaje de error en ingreso al sistema.

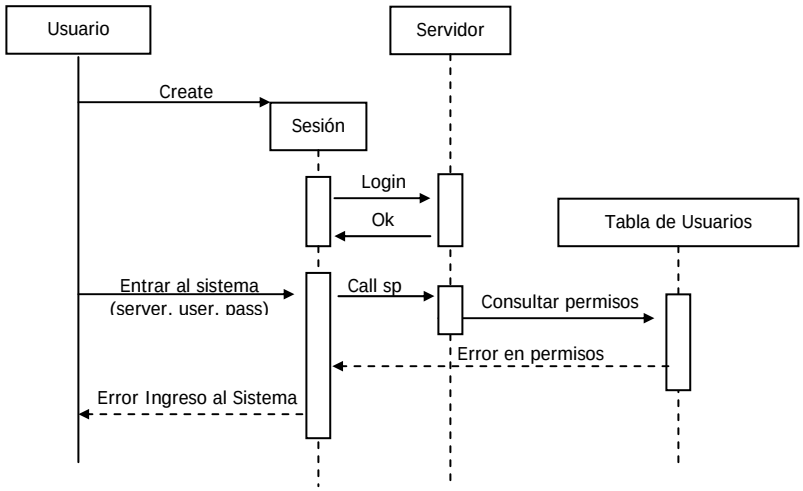


FIGURA 3.22 DIO PARA ESCENARIO 9.3

La arquitectura que se va a implementar es un Cliente / Servidor, el cual nos ofrece las siguientes características para el proyecto:

1. El cliente y el servidor pueden actuar como una sola entidad y también pueden actuar como entidades separadas, realizando actividades o tareas independientes.
2. Las funciones del cliente y servidor pueden estar en plataformas separadas, o en la misma plataforma.
3. Un servidor da servicio a múltiples clientes en forma concurrente.

4. Cada plataforma puede ser escalable independientemente. Los cambios realizados en las plataformas de los clientes o de los Servidores, ya sean por actualización o por reemplazo tecnológico, se realizan de una manera transparente para el usuario final.
5. Un sistema de servidores realiza múltiples funciones al mismo tiempo que presenta una imagen de un solo sistema a las estaciones clientes**.** Esto se logra combinando los recursos de cómputo que se encuentran físicamente separados en un solo sistema lógico, proporcionando de esta manera el servicio más efectivo para el usuario final.

También es importante hacer notar que las funciones cliente / servidor pueden ser dinámicas. Ejemplo, un servidor puede convertirse en cliente cuando realiza la solicitud de servicios a otras plataformas dentro de la red.

Esta arquitectura puede incluir múltiples plataformas, bases de datos, redes y sistemas operativos. Por lo tanto, en nuestra implantación de la

solución nos involucraremos con el estándar TCP/IP, funcionando sobre una la plataforma Windows, con un framework JSDK.

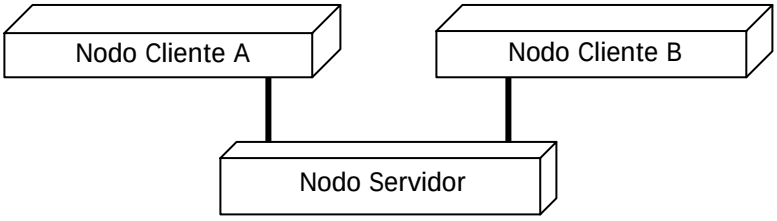


FIGURA 3.23 ARQUITECTURA DE 2 CAPAS DEL SISTEMA

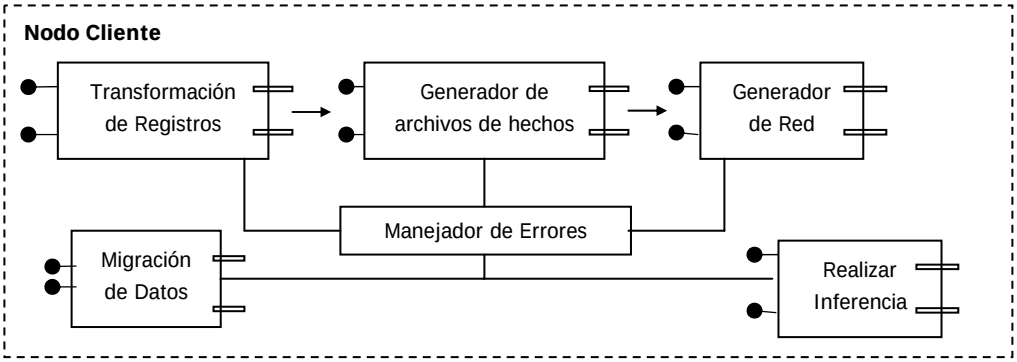


FIGURA 3.24 DETALLE DEL NODO CLIENTE

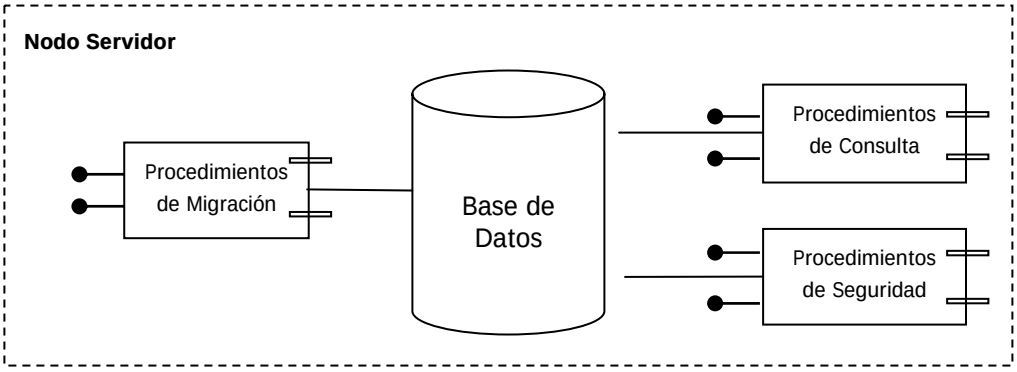


FIGURA 3.25 DETALLE DEL NODO SERVIDOR

CAPÍTULO 4

4. DISEÑO E IMPLEMENTACIÓN DE LA SOLUCIÓN.

4.1 Metodología

4.1.1 Integración y Recopilación de Datos

Los procesos de análisis, diseño e implementación de la solución se fundamentan en el uso de datos de tipo académico que permitirán obtener conocimiento acerca del comportamiento de los estudiantes de la FIEC durante los registros académicos, brindando una base para la toma de decisiones en la planificación de la apertura de paralelos. El proyecto se basará en la información de tipo académico referente a los estudiantes de la FIEC; la misma esta constituida por:

- Listado de materias actuales de la FIEC.
- Listado de carreras que ofrece la FIEC.
- Flujo de materias en cada carrera en la FIEC.
- Registros académicos del presente semestre en la FIEC.
- Historial académico de los estudiantes de la FIEC cuyas matrículas inicien con el número 2000 o un número mayor.
- Listado de Profesores de la FIEC que han sido asignados a los distintos paralelos de cada materia.

A medida que aumenta el número de datos, el conocimiento obtenido tendrá una tasa de certeza más alta, debido a que el método bayesiano tiene como base la Teoría de la Probabilidad.

4.1.2 Fase de Minería de Datos

Para poder resolver el problema planteado en este proyecto de tesis sobre: “¿Cuántos paralelos de una materia determinada se deben abrir?”, necesitamos determinar: “¿Cuál es la probabilidad de que un estudiante se registre en aquella materia?”. Este resultado está directamente relacionado con la solución.

Para poder encontrar este valor, necesitamos determinar el número de estudiantes que se registraron en la materia del total de estudiantes que podían registrarse. A este evento lo llamamos “registro” y toma los valores “Si” o “No”. Reconocemos que el evento “registro" es la variable bayesiana de interés, la variable clase, con dos valores posibles. Sin embargo, ésta variable no se encuentra almacenada en la base de datos. Es necesario realizar una transformación sobre los datos de historia académica, que presentan la información deseada de forma implícita.

A través de este proceso perseguimos conocer, qué estudiante se podría registrar en una materia determinada, dentro de un año y término específico, sometido a un conjunto de reglas de flujo de materias correspondientes a su carrera. Para este proceso es necesario conocer el año y término en que cada estudiante tuvo aprobados los prerrequisitos de la materia analizada.

Para obtener la información del evento de registro, hacemos uso de un procedimiento almacenado en la base de datos llamado: sp_crear_tabla_registros.

Básicamente a través de este procedimiento almacenado, preguntamos tres cosas:

a. Si cada estudiante cumple con los prerrequisitos de una materia en particular en un año y término determinados haciendo uso de una versión de las reglas de flujo de materias para su carrera.
b. Si cada estudiante que cumple con los prerrequisitos de una materia se registró o no en la misma, en un año y término determinados.
c. Si el estudiante ha aprobado la materia en un año y término anterior al año y término en observación.

Este procedimiento se ejecuta cada vez que se posee nuevos datos almacenados en la base respecto a estudiantes, paralelos e historia académica. No recibe ningún parámetro y como

resultado genera la tabla registros. Ésta tabla contiene la información sobre las circunstancias que determinaron el evento del registro, el mismo que puede tener dos valores: “si” o “no”.

La tabla registros contiene los siguientes campos:

TABLA 11

ESTRUCTURA DE TABLA REGISTROS

Secuencia	Identificador del término al que corresponde la muestra o registro.
Versionflujo	Versión del flujo que rige a la muestra o registro.
Anio	Año de observación de la muestra o registro.
Termino	Término de observación de la muestra o registro.
Materia	Código y nombre de materia.
cod_materia	Código de la materia.
Matricula	Matrícula del estudiante.
Sexo	Sexo del estudiante.
Promedio	Promedio del estudiante.
Factor_p	Factor socio-económico del estudiante.
Carrera	Carrera del estudiante (determina el flujo utilizado)
Term_ap_req	Término en el que el estudiante aprobó el último prerrequisito de la materia en cuestión.
numero_paralelo	Número de paralelo abierto en el término correspondiente.
Profesor	Nombre del profesor asignado al paralelo.
Registro	Especifica si el estudiante se registró.

4.1.3 Aplicación del Método Bayesiano en la Solución

Una vez realizada la transformación de datos se ha obtenido su resultado (que estudiante podría registrarse en cuál materia y en que momento) y se encuentra almacenado de manera explícita en la base de datos (tabla registros).

El método bayesiano persigue relacionar las variables influyentes sobre un evento. Esta característica es determinante para optar por éste método en la solución del problema de los registros académicos. A través de esta característica, es posible construir varias redes bayesianas que representan las situaciones deseadas para la evaluación de la variable clase, el evento del registro en la materia.

Inicialmente, probamos plantear una pregunta sin condiciones, tan solo especificando la materia: ¿Cuál es la probabilidad de que alguien se registre en una materia en particular?. Ahora, podríamos realizar la misma pregunta agregando una condición: ¿Cuál es la probabilidad de que alguien se registre en la materia

en particular dado que este alguien pertenece a la carrera de Ingeniería en Computación?

De esta forma, podemos agregar algunas otras condiciones. A medida que agregamos más condiciones nuestra pregunta se hace más específica. Es así que podemos “filtrar” las circunstancias bajo las cuales la variable “registro” muestra comportamientos diferentes.

Volviendo al ejemplo anterior, declaramos que A es el evento de que un estudiante se registre y B el evento de que el estudiante sea de la carrera de Ingeniería Computación. Usando la fórmula 2.1, correspondiente al Teorema de Bayes, lo acoplamos al ejemplo:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

FORMULA 4.1 TEOREMA DE BAYES ACOPLADO A LOS REGISTROS ACADÉMICOS

Lo que se leería: “la probabilidad de que un estudiante se registre en la materia X dado que es estudiante de Ingeniería en Computación, es igual a la probabilidad de que un estudiante sea de Ingeniería computación dado que se registró, por la probabilidad de que alguien se registre, dividido para la probabilidad de que un estudiante sea de Ingeniería en Computación”. Gráficamente y colocando valores arbitrarios, tenemos:

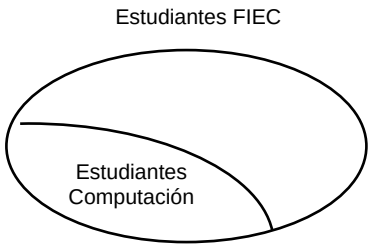


FIGURA 4.1 REPRESENTACIÓN DE ESTUDIANTES

Si tenemos 1100 estudiantes de la FIEC y 230 son de Computación P (B) = 0.2090. Ahora gráficamente tenemos:

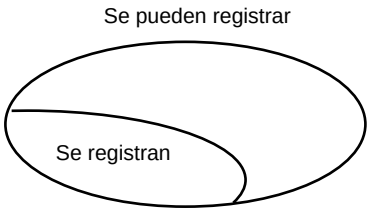


FIGURA 4.2 REPRESENTACIÓN DE LOS REGISTROS

Si 120 estudiantes se pueden registrar y sólo 30 se registran, tenemos P(A) = 0.25

FIGURA 4.3 ESTUDIANTES DE LA CARRERA INGENIERÍA EN COMPUTACIÓN QUE SE REGISTRAN

Ahora, si de los 30 estudiantes que se registran, 7 son de Computación, P (B|A) = 0.2333 entonces:

$$P(A|B) = \frac{(0.2333) \times (0.25)}{(0.0090)}$$

$$P(A|B) = 0.2790$$

FORMULA 4.2 VALORES EN FORMULA DE BAYES

4.1.4 Diseño de Base de Datos

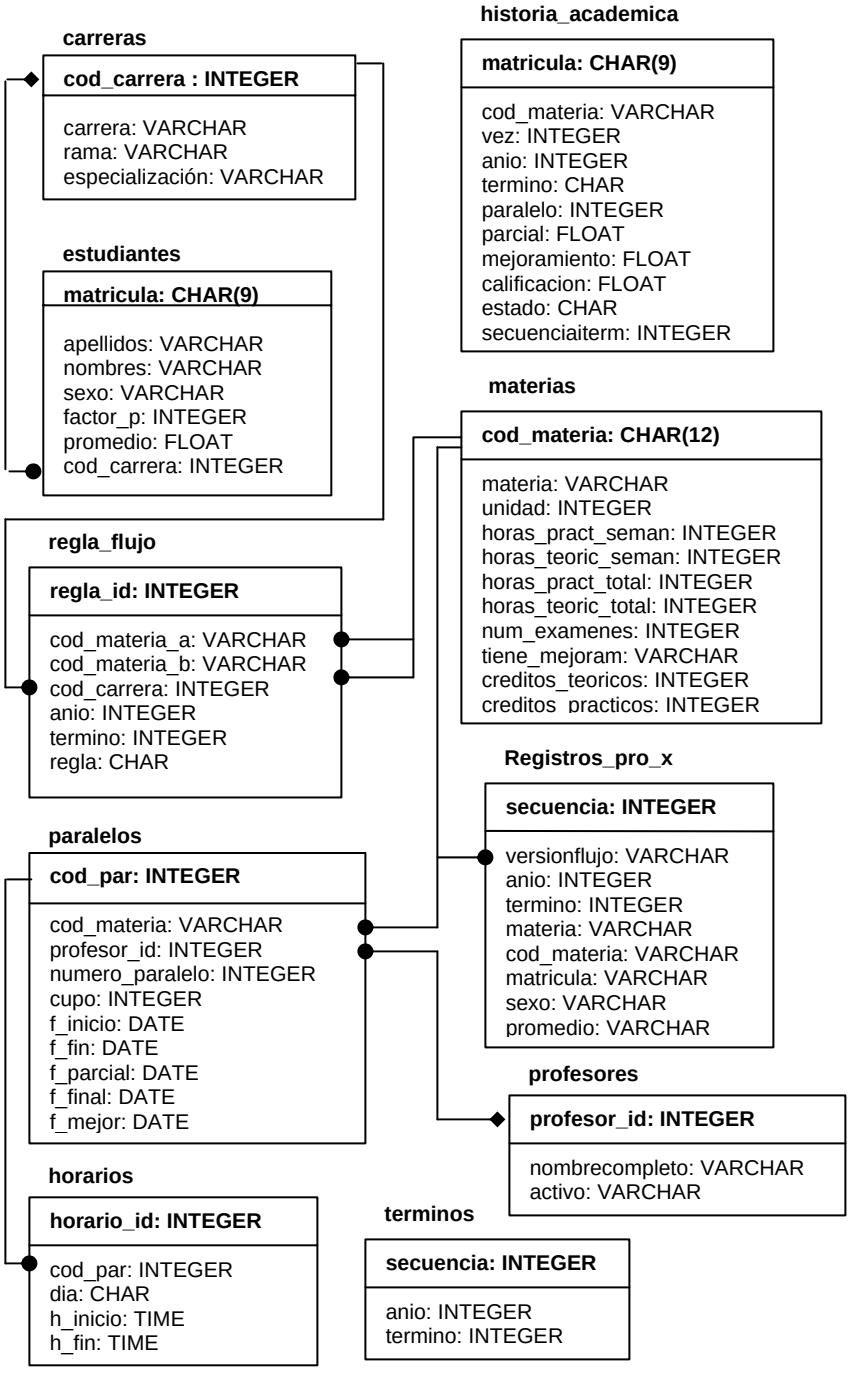


FIGURA 4.4 MODELO FISICO DE LA BASE DE DATOS

4.1.5 Diseño de Interacción con el Usuario

Para una mejor interacción con el usuario el sistema cumple con:

- Ser flexible, posee mensajes fáciles de entender y muestra resultados que cumplen con el objetivo.
- Tener formas de escape efectivas.
- Proporcionar ayuda al usuario, a través de mensajes sencillos, proveyendo respuestas efectivas a los problemas que se susciten.
- Proporcionar información necesaria, sin caer en el exceso de información ya que puede distraer y confundir.
- Usar palabras familiares y no difíciles, para que fácilmente pueda reconocer un comando del sistema. Los menús y etiquetas de botones tienen palabras claves del proceso.
- Retroalimentar al usuario, mantenerlo alerta e informado de lo que pasa en el sistema.

La aplicación cumple con ser consistente, ya que profundiza en aspectos que están catalogados de la siguiente manera:

- **Interpretación del comportamiento del usuario:** la interfaz del usuario comprende el significado que le atribuye un usuario a cada requerimiento. Ejemplo: mantiene el significado de los comandos abreviados (atajos) definidos por el usuario.

- **Pequeñas estructuras visibles:** se ha establecido un conjunto de objetos visibles capaces de ser controlados por el usuario, éstos permiten ahorrar tiempo en la ejecución de tareas específicas. Ejemplo: icono botón para predicción.

- **Una sola aplicación o servicio:** la interfaz permite visualizar a la aplicación o servicio utilizado como un componente único. Ejemplo: La interfaz despliega un menú, pudiendo además acceder al mismo mediante comandos abreviados.

- **Consistencia de la plataforma:** la interfaz es concordante con la plataforma, ya que tiene un esquema basado en

ventanas, el cual es acorde al manejo del sistema operativo Windows.

La inconsistencia en el comportamiento de componentes de la interfaz debe ser fácil de visualizar. Los objetos son consistentes con su comportamiento. Desde luego, la única forma de verificar si la interfaz satisface las expectativas del usuario es mediante el testeo.

Este proyecto considera la productividad del usuario antes que la productividad de la máquina. Si el usuario debe esperar la respuesta del sistema por un período prolongado, estas pérdidas de tiempo se pueden convertir en pérdidas de eficiencia en la Facultad.

4.1.6 Justificación de las Herramientas Seleccionadas

Para la implementación de la solución, se ha elegido el lenguaje de programación JAVA, porque posee características muy poderosas, mencionando algunas tenemos:

Arquitectura neutral.- es decir, Java permite que su aplicaciones se interpreten de manera independiente a la plataforma donde se esté trabajando. Esto significa que las aplicaciones o los applets se comportan de la misma forma sin importar que se ejecute desde Windows, Linux o Macintosh. Para hacer independientes las aplicaciones escritas en Java, se compila su código en un archivo objeto (byte codes) que no depende de la arquitectura de la máquina en que será ejecutado. Cualquier máquina que tenga el sistema “run-time” puede ejecutar ese código. Para la realización del proyecto usaremos como run-time, el J2SDK 1.4.0

Seguridad.- el código en Java pasa por una serie de pruebas antes de “ejecutarse” en una computadora. Dicho código se pasa a través de un verificador de “byte codes” que comprueba el formato de los fragmentos de código y aplica métodos para detectar fragmentos de código ilegal, como violación a derechos de acceso sobre objetos o cambiar el tipo o clase de un objeto.

Portabilidad.- implementa estándares para facilitar el desarrollo, tal es el caso de los tipos de datos o interfaces gráficas de usuario que mantienen la independencia de la plataforma.

Multihilos.- Java fue diseñado para satisfacer los requerimientos del mundo real, en cuanto a crear programas interactivos en un ambiente de red. Para ello, proporciona la programación multihilo la cual permite la escritura de programas que realicen varias tareas simultáneamente.

Java es un lenguaje dinámico, debido a que las clases son cargadas en el momento en que son necesitadas, ya sea del sistema de archivos local o desde algún sitio de la red mediante un protocolo URL.

Java utiliza un modelo de memoria conocido como "administración automática del almacenamiento" (automatic storage management), en el que el sistema en tiempo de ejecución de Java mantiene un seguimiento de los objetos. En el

momento que no están siendo referenciados por alguien, automáticamente se libera la memoria asociada con ellos.

Una característica muy interesante de Java y útil para el mejor desarrollo de nuestro proyecto, es la administración de errores y excepciones. Las excepciones son la manera como Java indica que ha ocurrido algo “extraño” durante la ejecución de un programa en Java. Comúnmente las excepciones son generadas y lanzadas por el sistema, cuando uno de estos eventos ocurre.

Cuando se genera una excepción, el sistema de tiempo de ejecución de Java, y en particular el manejador (handler) de errores y excepciones, busca un manejador para esa excepción, comenzando por el método que la originó y después hacia abajo en la pila de llamadas.

Cuando se encuentra un manejador, éste atrapa la excepción y se ejecuta el código asociado con dicho manejador. En el caso que no se encuentre un manejador para alguna excepción previamente lanzada, se ejecuta el manejador del sistema, cuya

acción típica es imprimir un mensaje de error y terminar la ejecución del programa.

El modelo de base de datos, se lo implementará con ayuda del motor MySQL 5.01 ya que es una base muy popular, confiable, multiplataforma, compacta y poderosa. Su gran ventaja es que se puede utilizar de manera gratis y su código fuente siempre está disponible. Sus principales características son:

- Velocidad y robustez.
- Los clientes pueden ser creados en lenguajes como C, C++, Java, Perl, Visual Basic.
- Multiproceso, es decir puede usar varias CPU si éstas están disponibles.
- Puede trabajar en distintas plataformas y SO distintos.
- Sistema de contraseñas y privilegios muy flexible y segura.
- Registros de longitud fija y variable.
- El servidor soporta mensajes de error en distintas lenguas.

- Diversos tipos de columnas como enteros de 1, 2, 3, 4, y 8 bytes, coma flotante, doble precisión, carácter, fechas, enumerados.

4.1.7 Plan de Pruebas

Para garantizar la eficacia del sistema se expone a continuación el plan de pruebas:

TABLA 12

PRUEBA DEL INICIO DE SESIÓN DEL ADMINISTRADOR

Propósito:	Probar que el administrador puede iniciar sesión con el nombre de usuario apropiado y su contraseña.
Prerrequisitos:	El administrador no ha iniciado sesión. El nombre de usuario abonilla existe y la cuenta tiene permisos de administrador.
Datos de Prueba:	server = {localhost:3306} username = {abonilla} password = {abonilla}
Pasos:	1. Entrar a la ventana de login 2. Teclear el server 3. Teclear el username 4. Teclear el password 5. Hacer “click” en Entrar 6. Ver: menú de inicio del administrador

TABLA 13

PRUEBA DEL INICIO DE SESIÓN DEL OPERADOR

Propósito:	Probar que el operador puede iniciar sesión con el nombre de usuario apropiado y su contraseña.
Prerrequisitos:	El operador no ha iniciado sesión todavía. El nombre de usuario mojeda existe y la cuenta tiene permisos de operador
Datos de Prueba:	server = {localhost:3306} username = {mojeda} password = {mojeda}
Pasos:	1. Entrar a la ventana de login 2. Teclear el server 3. Teclear el username 4. Teclear el password 5. Hacer “click” en Entrar 6. Ver: menú de inicio del operador

TABLA 14

PRUEBA DE MIGRACIÓN DE MATERIAS

Propósito:	Probar que el administrador puede migrar un archivo con datos de las materias que dictan en la FIEC
Prerrequisitos:	El usuario ha iniciado como administrador. El archivo contiene información sobre las materias de la facultad.
Datos de Prueba:	archivo = {materias.csv}
Pasos:	1. Seleccionar la opción Migrar Datos de la barra de menú 2. Teclear o buscar la dirección del archivo tipo “csv” que contenga datos de las materias. 3. Hacer “click” en Migrar 4. Ver: estado de la migración de materias 5. Ver: mensaje de migración exitosa, previa terminación del proceso de migrar datos. 6. Hacer en “click” Aceptar 7. Ver: ventana principal con la barra de menú.
Notas:	Se asume que el archivo tiene sus datos bien organizados y de acuerdo a los campos de la tabla.

TABLA 15

PRUEBA DE MIGRACIÓN DE FLUJO DE CARRERA

Propósito:	Probar que el administrador puede migrar un archivo con datos de los flujos de carreras que ofrece la FIEC.
Prerrequisitos:	El usuario ha iniciado como administrador. El archivo contiene información sobre los flujos de carrera.
Datos de Prueba:	Archivo = {flujos.csv}
Pasos:	1. Seleccionar la opción Migrar Datos de la barra de menú 2. Teclear o buscar la dirección del archivo tipo “csv” que contenga datos del flujo de carrera. 3. Hacer “click” en Migrar 4. Ver: estado de la migración del flujo 5. Ver: mensaje de migración exitosa, previa terminación del proceso de migrar datos. 6. Hacer en “click” Aceptar 7. Ver: ventana principal con la barra de menú.
Notas:	Se asume que el archivo tiene sus datos bien organizados y de acuerdo a los campos de la tabla.

TABLA 16

PRUEBA DE MIGRACIÓN DE HISTORIALES

Propósito:	Probar que el administrador puede migrar de un archivo con datos de historiales académicos de estudiantes de la FIEC.
Prerrequisitos:	El usuario ha iniciado como administrador. El archivo contiene historiales académicos de los estudiantes.
Datos de Prueba:	archivo = {historiales.csv}
Pasos:	1. Seleccionar la opción Migrar Datos de la barra de menú 2. Teclear o buscar la dirección del archivo tipo “csv” que contenga datos del historial académico. 3. Hacer “click” en Migrar 4. Ver: estado de la migración de historiales 5. Ver: mensaje de migración exitosa, previa terminación del proceso de migrar datos. 6. Hacer en “click” Aceptar 7. Ver: ventana principal con la barra de menú.
Notas:	Se asume que el archivo tiene sus datos bien organizados y de acuerdo a los campos de la tabla.

TABLA 17

PRUEBA DE MIGRACIÓN DE ESTUDIANTES

Propósito:	Probar que el administrador puede migrar de un archivo con datos de estudiantes de la FIEC.
Prerrequisitos:	El usuario ha iniciado como administrador. El archivo contiene estudiantes de la facultad.
Datos de Prueba:	archivo = {estudiantes.csv}
Pasos:	1. Seleccionar la opción Migrar Datos de la barra de menú 2. Teclear o buscar la dirección del archivo tipo “csv” que contenga datos de los estudiantes. 3. Hacer “click” en Migrar 4. Ver: estado de la migración de estudiantes 5. Ver: mensaje de migración exitosa, previa terminación del proceso de migrar datos. 6. Hacer en “click” Aceptar 7. Ver: ventana principal con la barra de menú.
Notas:	Se asume que el archivo tiene sus datos bien organizados y de acuerdo a los campos de la tabla.

TABLA 18

PRUEBA DE INFERENCIA DE UNA MATERIA POR PARTE DEL ADMINISTRADOR

Propósito:	Probar que el administrador puede hacer inferencia sobre una materia.
Prerrequisitos:	El usuario ha iniciado como administrador El administrador selecciona la materia El archivo .bif de la materia existe
Datos de Prueba:	Código materia = {FIEC02097}
Pasos:	1. Seleccionar la opción Realizar Inferencia de la barra de menú 2. Teclear el código de materia FIEC02097 3. Hacer “click” en Ejecutar 4. Ver: barra nueva ventana 5. Ver: diagrama de la red bayesiana 6. Ver: opciones para manipular la red

TABLA 19

PRUEBA DE INFERENCIA DE UNA MATERIA POR PARTE DEL OPERADOR

Propósito:	Probar que el operador puede hacer inferencia sobre una materia.
Prerrequisitos:	El usuario ha iniciado como operador El operador selecciona la materia El archivo .bif de la materia existe
Datos de Prueba:	Código materia = {FIEC02097}
Pasos:	1. Seleccionar la opción Realizar Inferencia de la barra de menú 2. Teclear el código de materia FIEC02097 3. Hacer “click” en Ejecutar 4. Ver: barra nueva ventana 5. Ver: diagrama de la red bayesiana 6. Ver: opciones para manipular la red

4.2 Resultados Esperados

Como resultado de este proyecto de tesis, se ha obtenido una herramienta que permite:

a. Obtener de la historia académica y del conjunto de estudiantes, información acerca de qué estudiante pudo registrarse en una materia específica en un año y término determinado, basado en una regla de flujo que pertenezca a la carrera cada estudiante. Esta información se encuentra de manera implícita en los registros de la historia académica; por ello, a este proceso lo hemos denominado 'transformación'.

   Estos datos que nos dicen en qué se registró cada estudiante y cuándo, también nos pueden decir "*cuándo un estudiante pudo registrarse en una materia"*, de acuerdo a las reglas de su flujo de carrera, y observar si efectivamente se registró o no. Ésta información es el punto de partida para poder realizar observaciones sobre las circunstancias o en términos estadísticos, un conjunto de eventos que aumentan o disminuyen la probabilidad de un registro.

b. Generar el conjunto de archivos que contenga la información específica para la materia que deseemos analizar. El conjunto está compuesto por tres archivos:

- El archivo de nombres contiene los nombres de las variables que proponemos para creación de la red.
- El archivo de datos es el conjunto de hechos a partir del cual realizaremos el aprendizaje de la red bayesiana: las relaciones de sus variables y las distribuciones de probabilidad.
- El archivo de pruebas, similar al archivo de datos, contiene casos con su respectivo valor de variable clase; luego de aprender la red a partir del archivo de datos, el archivo de pruebas sirve para intentar predecir los casos que éste contiene y comparar el valor inferido contra el valor real y medir así el error de predicción.

La herramienta también permitirá aprender la red de relaciones entre las variables, y establecer las distribuciones de probabilidad entre los valores que toma cada variable en cada relación, para finalmente inferir la probabilidad de registro en una materia realizando observaciones sobre las demás variables.

CONCLUSIONES Y RECOMENDACIONES

Nosotros podemos concluir que las redes bayesianas tienen un gran potencial en muchos campos y el campo académico no es la excepción, al proveer información importante y necesaria sobre la apertura de paralelos de una o mas materias.

Las redes bayesianas son una herramienta para la minería de datos, que nos permite hallar relaciones en un conjunto de hechos. En este proyecto ha sido aplicada en los registros académicos, encontrando así las relaciones existentes en las variables que determinan un evento de registro.

Los métodos bayesianos resuelven proyecciones probabilísticas, realizando observaciones sobre las variables que participan en la red bayesiana, permitiendo responder preguntas “que si” sobre un almacén de datos.

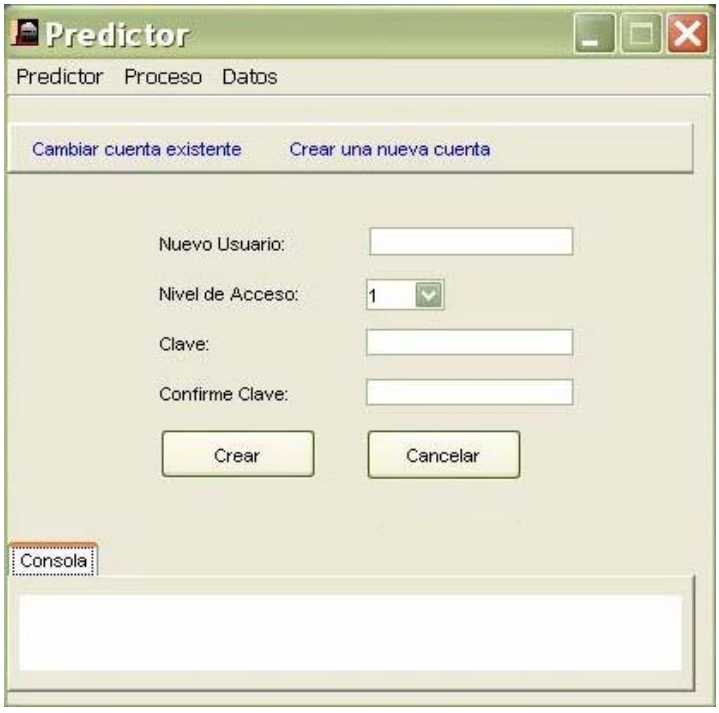
La aplicación “Predictor”, producto del análisis y diseño de este proyecto de tesis, puede dar soporte para la toma de decisión de cuántos paralelos abrir y así, reducir el problema del registro de un estudiante en un paralelo.

Nosotros recomendamos, luego de realizar este proyecto de tesis, lo siguiente:

- Mantener la base de datos actualizada con datos correctos y con la menor cantidad valores en blanco o nulos.
- Realizar la transformación de los datos almacenados en la base en información explícita, cada vez que se migre nueva información a la base de datos.
- Crear una aplicación para manipular las redes bayesianas de manera gráfica, permitiendo una mayor interacción entre el usuario y el sistema; mejorando así el entendimiento de la red bayesiana.

APÉNDICES

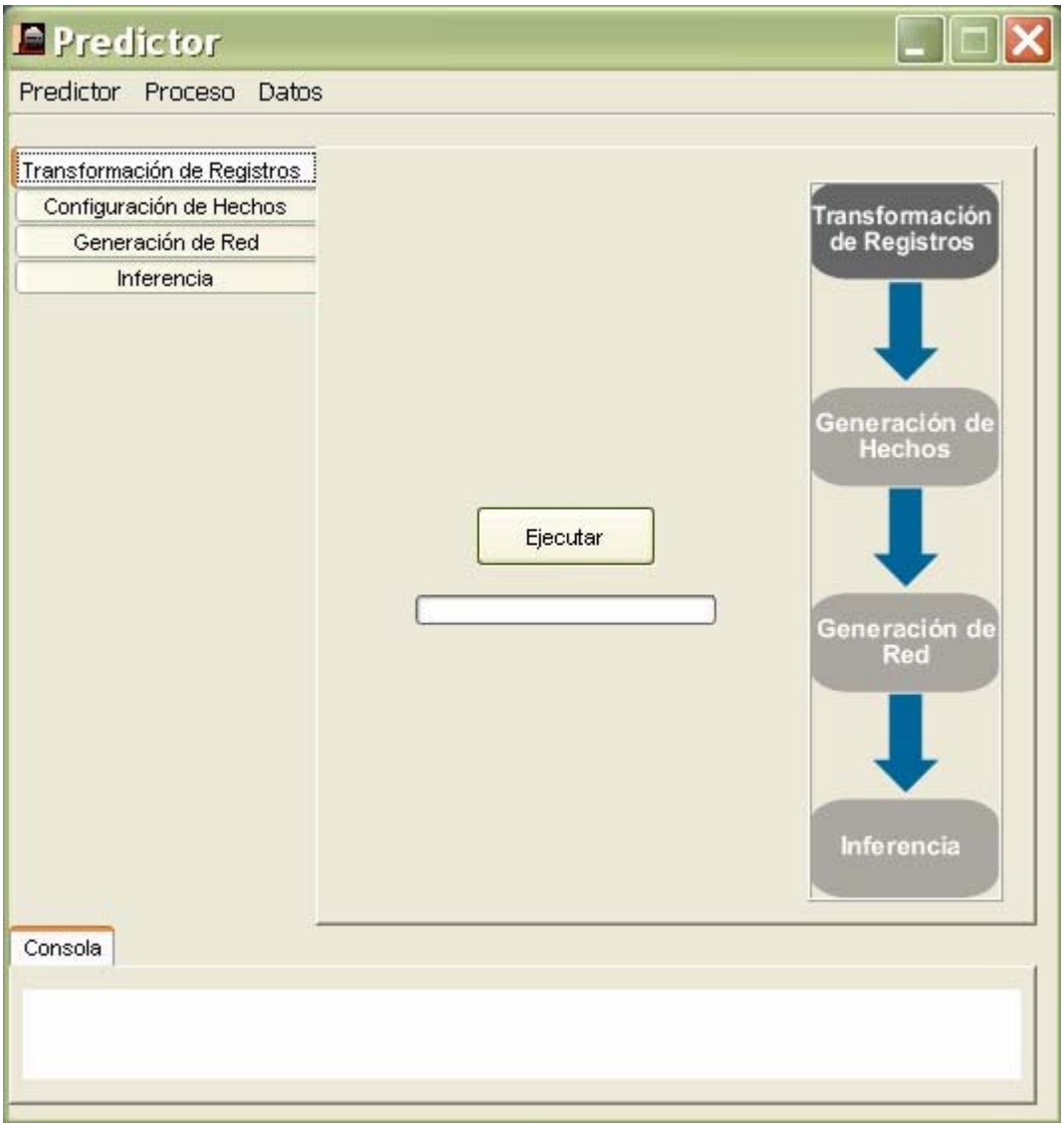
VENTANA DE INGRESO AL SISTEMA



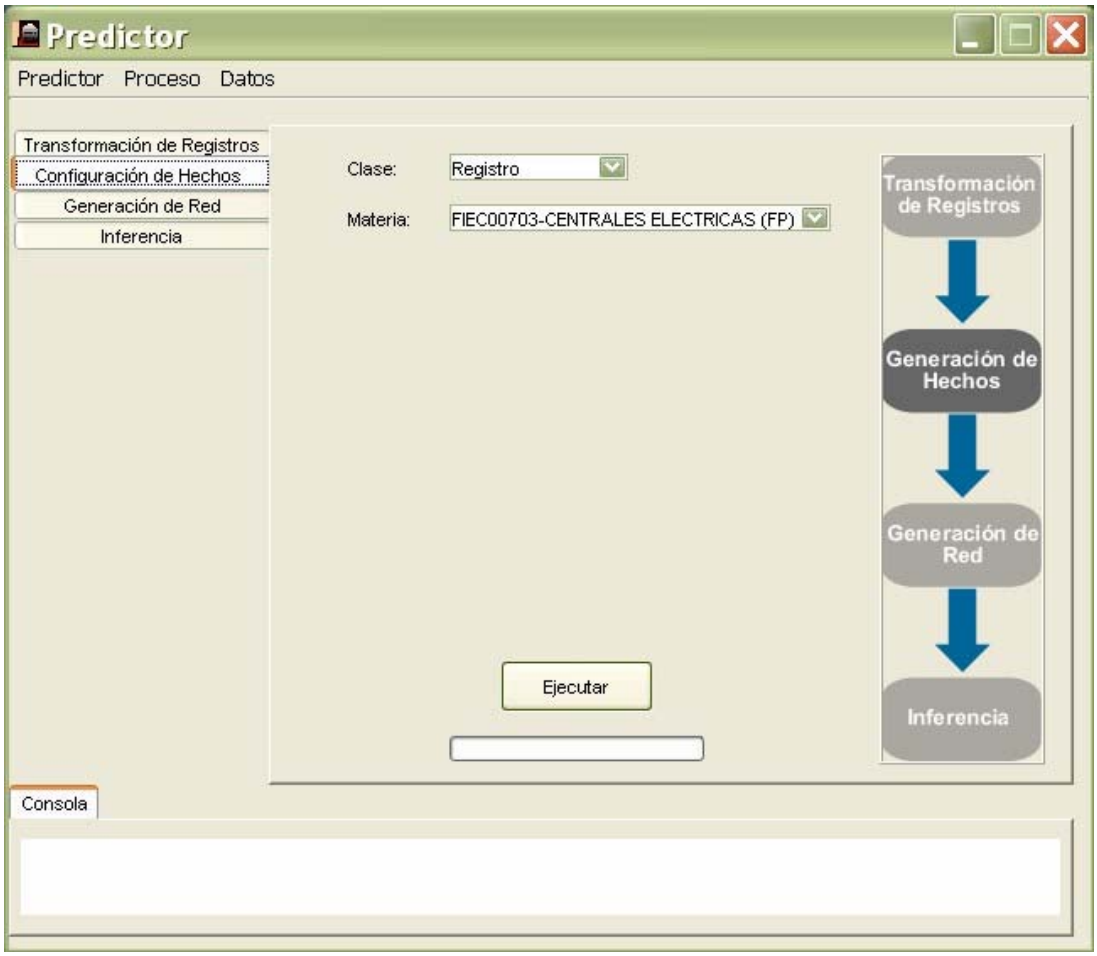
VENTANA DE CREACIÓN DE CUENTA

VENTANA DE MODIFICACIÓN DE CUENTA

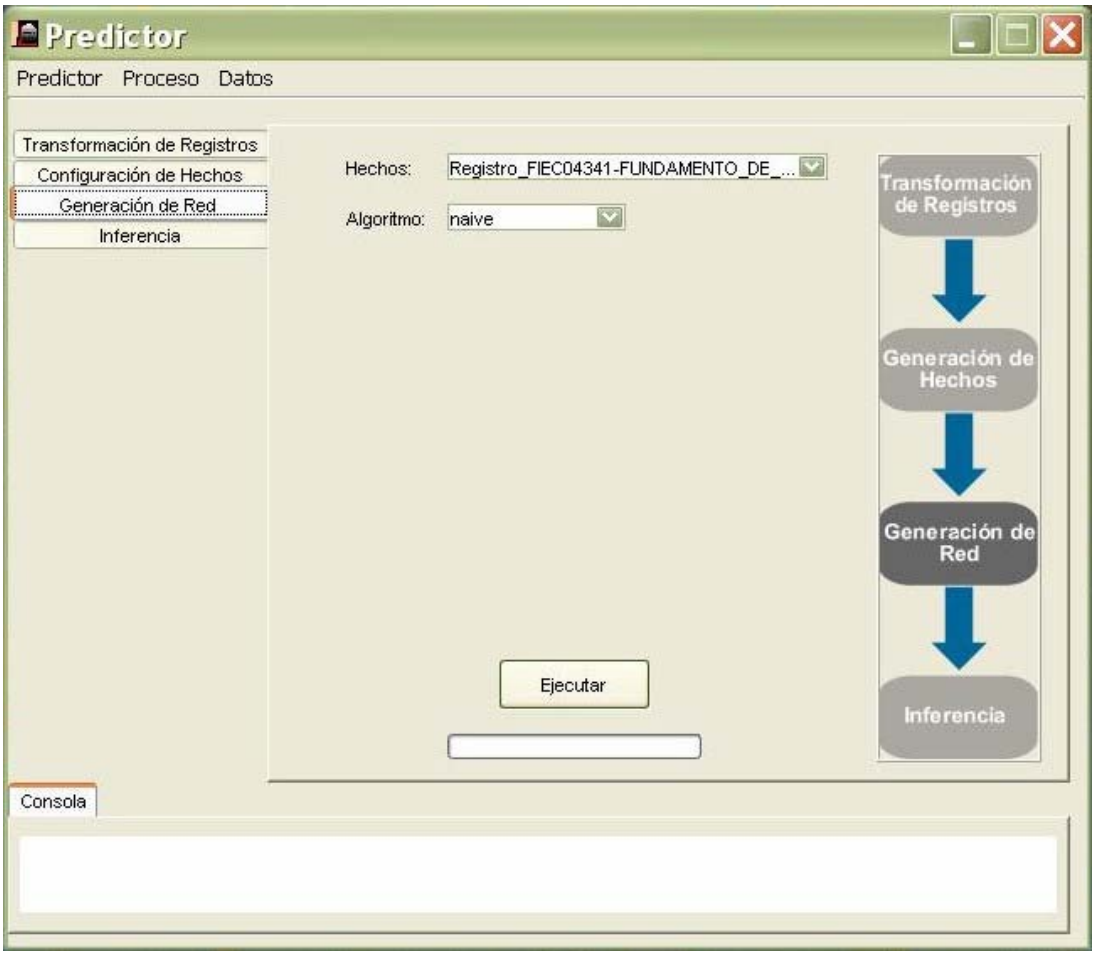
VENTANA DE MIGRACION DE DATOS



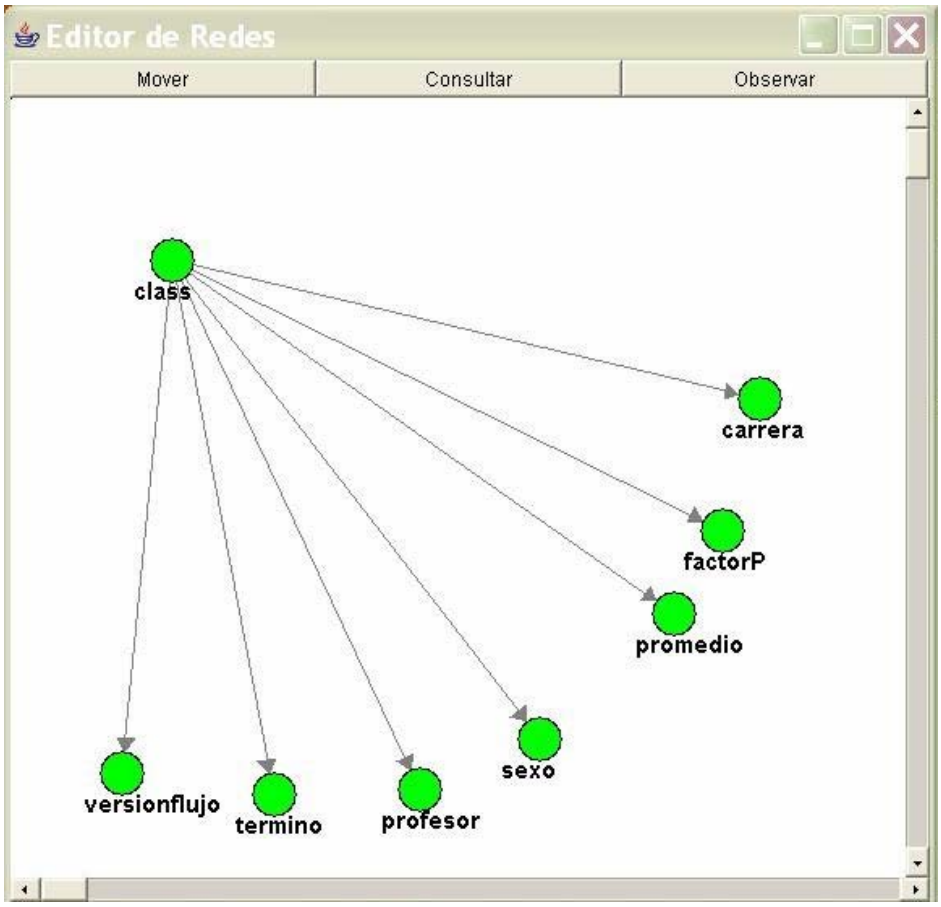
VENTANA DE TRANSFORMACIÓN DE REGISTROS

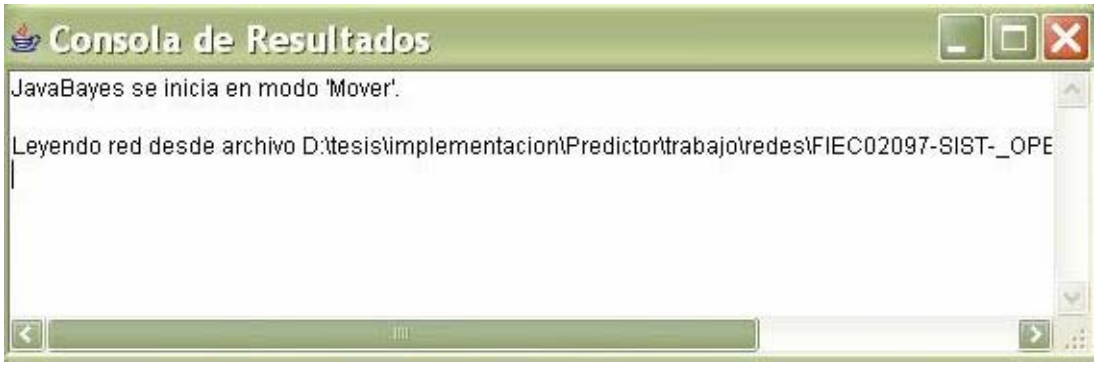


VENTANA DE GENERACIÓN DE ARCHIVOS DE HECHOS



VENTANA DE GENERACIÓN DE RED





VENTANA DE INFERENCIA Y EDICION DE REDES

BIBLIOGRAFÍA

1. HERNÁNDEZ JOSE, RAMÍREZ MA. JOSÉ, FERRI CESAR, Introducción a la Minería de Datos, Valencia – España, Prentice Hall, 257p.

2. JONSON RICHARD, Probabilidad y estadística para ingenieros de Miller y Freund, Wisconsin – USA, 6ª. Edición, Prentice Hall, 76p.

3. Deitel & Deitel, Java How to Program, New Jersey – USA, 2ª. Edición, Prentice Hall, 122p.

4. http://www.programacion.net/bbdd/tutorial/jdbc/ Tutorías de JAVA

5. http://www.mor.itesm.mx/~rdec/node153.html Redes Bayesianas

6. http://es.wikipedia.org/wiki/ Enciclopedia Online

7. http://jbnc.sourceforge.net Código abierto JBNC