TRABAJO FINAL DE ESPECIALIDAD
EN
INGENIERÍA DE SISTEMAS EXPERTOS

ANÁLISIS DE CLASIFICADORES BAYESIANOS

Autor: Lic. Enrique José Fernández

Directora: M.Ing. Paola Britos

Diciembre 2004

Indice

1. Introducción

Una red bayesiana es un grafo acíclico dirigido en el que cada nodo representa una variable y cada arco una dependencia probabilística, son utilizadas para proveer: una forma compacta de representar el conocimiento, y métodos flexibles de razonamiento. El obtener una red, bayesiana a partir de datos, es un proceso de aprendizaje que se divide en dos etapas: el aprendizaje estructural y el aprendizaje paramétrico.

En este trabajo se describirá el funcionamiento de tres algoritmos de Clasificadores, Naïve Bayes, TAN y KDB. Se mostrará además como, a través del programa Elvira, se puede llegar a obtener una red Bayesiana con estos clasificadores. Dicha red variará dependiendo del algoritmo clasificador aplicado, y de la combinación de este con algún algoritmo de inducción de árboles de decisión. Por último se mostrará una comparación que permita analizar las diferencias entre los distintos clasificadores y la influencia que en ellos genera los algoritmos generadores de árboles de decisión.

1.2. Definición de conceptos

1.2.1. Aprendizaje automático

El aprendizaje puede ser definido como "cualquier proceso a través del cual un sistema mejora su eficiencia" [Felgaer, P. et al, 2003]. La habilidad de aprender es considerada como una característica central de los "sistemas inteligentes" [García-Martínez, 1997; García Martínez *et al*, 2003; Fritz *et al.*, 1989; García-Martínez, 1993; García-Martínez, 1995; García-Martínez & Borrajo, 2000], y es por esto que se ha invertido esfuerzo y dedicación en la investigación y el desarrollo de esta área. El desarrollo de los sistemas basados en conocimientos motivó la investigación en el área del aprendizaje con el fin de automatizar el proceso de adquisición de conocimientos el cual se considera uno de los problemas principales en la construcción de estos sistemas.
Un aspecto importante en el aprendizaje inductivo es el de obtener un modelo que represente el dominio de conocimiento y que sea accesible para el usuario, en particular, resulta importante obtener la información de dependencia entre las variables involucradas en el fenómeno, en los sistemas donde se desea predecir el comportamiento de algunas variables desconocidas basados en otras conocidas; una representación del conocimiento que es capaz de capturar esta información sobre las dependencias entre las variables son las redes bayesianas [Ramoni & Sebastiani, 1999; Felgaer, P. et al, 2003].

1.2.2. Minería de datos

Se denomina Minería de Datos [Servente, & García-Martínez, 2002; Perichinsky & García-Martínez, 2000; Perichinsky *et al.*, 2000; Perichinsky *et al.*, 2001; Perichinsky *et al.*, 2003; Felgaer, P. et al, 2003] al conjunto de técnicas y herramientas aplicadas al proceso no trivial de extraer y presentar conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de grandes conjuntos de datos, con objeto de predecir de forma automatizada tendencias y comportamientos; y describir de forma

automatizada modelos previamente desconocidos [Piatetski-Shapiro *et al.,* 1991; Chen *et al.,* 1996; Mannila, 1997; Felgaer, P. et al, 2003]. El término Minería de Datos Inteligente [Evangelos & Han, 1996; Michalskí *et al.,* 1998; Felgaer, P. et al, 2003] refiere específicamente a la aplicación de métodos de aprendizaje automático [Michalski *et al.,* 1983; Holsheimer & Siebes, 1991; Felgaer, P. et al, 2003], para descubrir y enumerar patrones presentes en los datos, para estos, se desarrollaron un gran número de métodos de análisis de datos basados en la estadística [Michalski *et aL,* 1982; Felgaer, P. et al, 2003]. En la medida en que se incrementaba la cantidad de información almacenada en las bases de datos, estos métodos empezaron a enfrentar problemas de eficiencia y escalabilidad y es aquí donde aparece el concepto de minería de datos. Una de las diferencias entre al análisis de datos tradicional y la minería de datos es que el primero supone que las hipótesis ya están construidas y validadas contra los datos, mientras que el segundo supone que los patrones e hipótesis son automáticamente extraídos de los datos [Hernández Orallo, 2000; Felgaer, P. et al, 2003]. Las tareas de la minería de datos se pueden clasificar en dos categorías: minería de datos descriptiva y minería de datos predicativa [Piatetski-Shapiro *et al.,* 1996; Han, 1999; Felgaer, P. et al, 2003].

1.2.3. Algoritmos de inducción.

Este tipo de algoritmos utiliza ejemplos como entradas para aplicar sobre ellos un proceso inductivo y así presentar la generalización de los mismos como resultado de salida. Existen dos tipos de ejemplo, los positivos y los negativos, Los primeros fuerzan la generalización, mientras que los segundos previenen para que no sea excesiva. Los ejemplos que se traten durante la fase de entrenamiento deben ser representativos de los conceptos que se está tratando de enseñar. En la actualidad existen numerosos enfoques de algoritmos de inducción y variedad en cada enfoque, el presente trabajo solo tratará los algoritmos orientados a generar árboles de decisión, particularmente los algoritmos ID3 y C4.5 que a continuación se detallan.

1.2.3.1 Introducción al Algoritmo ID3

El algoritmo ID3, diseñado en 1993 por J. Ross Quinlan [Quinlan, 1993a, Quinlan, 1993b; Servente et al, 2002], toma objetos de una clase conocida y los describe en términos de una colección fija de propiedades o de variables, produciendo un árbol de decisión sobre estas variables que clasifica correctamente todos los objetos [Quinlan, 1993b; Servente et al, 2002]. Hay ciertas cualidades que diferencian a este algoritmo de otros sistemas generales de inferencia. La primera se basa en la forma en que el esfuerzo requerido para realizar una tarea de inducción crece con la dificultad de la tarea. El ID3 fue diseñado específicamente para trabajar con masas de objetos, y el tiempo requerido para procesar los datos crece sólo linealmente con la dificultad, como producto de:

- La cantidad de objetos presentados como ejemplos.
- La cantidad de variables dadas para describir estos objetos.
- La complejidad del concepto a ser desarrollado (medido por la cantidad de nodos en el árbol de decisión).

Esta linealidad se consigue a costa del poder descriptivo ya que los conceptos desarrollados por el ID3 solo toman la forma de árboles de decisión basados en las

variables dadas, y este "lenguaje" es mucho más restrictivo que la 1ógica de primer orden o la lógica multivaluada, en la cual otros sistemas expresan sus conceptos [Quinlan, 1993b; Servente et al, 2002].
El ID3 fue presentado como descendiente del CLS creado por Hunt y, como contrapartida de su antecesor, es un mecanismo mucho más simple para el descubrimiento de una colección de objetos pertenecientes a dos o más clases. Cada objeto debe estar descripto en términos de un conjunto fijo de variables, cada una de las cuales cuenta con su conjunto de posibles valores. Por ejemplo, la variable humedad puede tener los valores {alta, baja} y la variable *clima*, {soleado, nublado, lluvioso}.

Una regla de clasificación en la forma de un árbol de decisión puede construirse para cualquier conjunto C de variables de esta forma [Quinlan, 1993b; Servente et al, 2002]:

- Si C está vacío, entonces se lo asocia arbitrariamente a cualquiera de las clases.
- Si C contiene los representantes de varias clases, se selecciona una variable y se particiona C en conjuntos disjuntos C_1,C_2,....., C_n, donde C_1 contiene aquellos miembros de C_i que tienen el valor i para la variable seleccionada. Cada una de estos subconjuntos se maneja con la misma estrategia.

El resultado es un árbol en el cual cada hoja contiene un nombre de clase y cada nodo interior especifica una variable para ser testeada con una rama correspondiente al valor de la variable.

1.2.3.1.1. Descripción del ID3

El objetivo del ID3 es crear una descripción eficiente de un conjunto de datos mediante la utilización de un árbol de decisión. Dados datos consistentes, es decir, sin contradicción entre ellos, el árbol resultante describirá el conjunto de entrada a la perfección. Además, el árbol puede ser utilizado para predecir los valores de nuevos datos, asumiendo siempre que el conjunto de datos sobre el cual se trabaja es representativo de la totalidad de los datos.

Dados:

- Un conjunto de datos.
- Un conjunto de descriptores de cada dato.
- Un clasificador/conjunto de clasificadores para cada objeto.

Se desea obtener un árbol de decisión simple basándose en la entropía, donde los nodos pueden ser:

- Nodos intermedios: en donde se encuentran los descriptores escogidos según el criterio de entropía, que determina cuál rama es la que debe tomarse.
- Hojas: estos nodos determinan el valor del clasificador.

Este procedimiento de formación de reglas Funcionará siempre, dado que no existen dos objetos pertenecientes a distintas clases pero con idéntico valor para cada una de

sus variables; si este caso llegara a presentarse, las variables son inadecuadas para el proceso de clasificación.

Hay dos conceptos importantes a tener en cuenta en el algoritmo ID3 [B1urock, 1996; Servente et al, 2002] la entropía y el árbol de decisión. La entropía se utiliza para encontrar el parámetro más significativo en la caracterización de un clasificador. El árbol de decisión es un medio eficiente e intuitivo para organizar los descriptores que pueden ser utilizados con funciones predictivas.

1.2.3.1.2. Pseudocódigo del Algoritmo ID3

A continuación, en la figura 1, se presenta el algoritmo del método ID3 para la construcción de árboles de decisión en función de un conjunto de datos previamente clasificados.

```
Función ID3
        (R: conjunto de atributos no clasificadores,
        C: atributo clasificador,
        S: conjunto de entrenamiento) devuelve un árbol de decisión;
Comienzo
        Si S está vacío,
            Devolver un único nodo con Valor Falla;
        Si todos los registros de S tienen el mismo valor para el atributo clasificador,
            Devolver un único nodo con dicho valor;
        Si R está vacío,
            Devolver un único nodo con el valor más frecuente del atributo
            clasificador en los registros de S [Nota: habrá errores, es decir, registros
            que no estarán bien clasificados en este caso];
        Si R no está vacío,
                D ← atributo con mayor Ganancia (D,S) entre los atributos de R;
                Sean {dj | j=1,2,...., m} los valores del atributo D;
                Sean {dj | j=1,2,...., m} los subconjuntos de S correspondientes a los
                valores de dj respectivamente;
                Devolver un árbol con la raíz nombrada como D y con los arcos
                nombrados dl, d2,...., dm que van respectivamente a los árboles
                ID3(R-{D}, C, Sl), ID3(R-{D}, C, S2), .., ID3(R-{D}, C, Sm);
Fin
```

Figura 1: Pseudocódigo del Algoritmo de ID3

1.2.3.1.3. Limitaciones de ID3

El ID3 puede aplicarse a cualquier conjunto de datos, siempre y cuando las variables sean discretas. Este sistema no cuenta con la facilidad de trabajar con variables continuas ya que analiza la entropía sobre cada uno de los valores de una variable, por lo tanto, tomaría cada valor de una variable continua individualmente en el cálculo de la entropía, lo cual no es útil en muchos de los dominios. Cuando se trabaja con variables continuas, generalmente se piensa en rangos de valores y no en valores particulares.

Existen varias maneras de solucionar este problema del ID3, como la agrupación de valores presentada en [Gallion et al.; Servente et al, 2002] o la discretización de los

mismos explicada en [B1urock, 1996; Quinlan, 1993d; Servente et al, 2002]. El C4.5 resolvió el problema de los atributos continuos mediante la discretización.

1.2.3.2. Introducción al Algoritmo C4.5

El C4.5 se basa en el ID3, por lo tanto, la estructura principal de ambos métodos es la misma. El C4.5 construye un árbol de decisión mediante el algoritmo "divide y reinarás" y evalúa la Información en cada caso utilizando los criterios de entropía y ganancia o proporción de ganancia, según sea el caso. A continuación, se explicarán las características particulares de este método que lo diferencian de su antecesor.

1.2.3.2.1. Pseudocódigo del Algoritmo C4.5

El algoritmo del método C4.5, mostrado en la figura 2, para la construcción de árboles de decisión a grandes rasgos muy similar al del ID3. Varía en la manera en que realiza las pruebas sobre las variables.

```
Función C4.5
        (R: conjunto de atributos no clasificadores,
        C: atributo clasificador,
        S: conjunto de entrenamiento) devuelve un árbol de decisión;
Comienzo
        Si S está vacío,
                Devolver un único nodo con Valor Falla;
        Si todos los registros de S tienen el mismo valor para el atributo clasificador,
             Devolver un único nodo con dicho valor;
        Si R está vacío,
                Devolver un único nodo con el valor más frecuente del atributo
                clasificador en los registros de S [Nota: habrá errores, es decir,
                registros que no estarán bien clasificados en este caso];
        Si R no está vacío,
                D ← atributo con mayor Proporción de Ganancia(D,S) entre los
                atributos de R;
                Sean {dj | j=1,2,...., m} los valores del atributo D;
                Sean {dj | j=1,2,...., m} los subconjuntos de S correspondientes a los
                valores de dj
                respectivamente;
                Devolver un árbol con la raíz nombrada como D y con los arcos
                nombrados d1, d2,....,dm, que van respectivamente a los árboles
                C4.5(R-{D}, C, Sl), C4.5(R-{D}, C, S2), C4.5(R-{D}, C, Sm);
Fin
```

Figura 2: Pseudocódigo del Algoritmo de C4.5

1.2.3.2.2. Características particulares del C4.5

En cada nodo, el sistema debe decidir cuál prueba escoge para dividir los datos. Los tres tipos de pruebas posibles propuestas por el C4.5 son [Quinlan, 1993d; Servente et al, 2002]:

1. La prueba "estándar" para las variables discretas, con un resultado y una rama para cada valor posible de la variable.
2. Una prueba más compleja, basada en una variable discreta, en donde los valores posibles son asignados a un número variable de grupos con un resultado posible para cada grupo, en lugar de para cada valor.
3. Si una variable A tiene valores numéricos continuos, se realiza una prueba binaria con resultados A <= Z y A > Z, para lo cual debe determinarse el valor límite *Z.*

Todas estas pruebas se evalúan de la misma manera, mirando el resultado de la proporción de ganancia, o alternativamente, el de la ganancia resultante de la división que producen. Ha sido útil agregar una restricción adicional: para cualquier división, al menos dos de los subconjuntos T_i deben contener un número razonable de casos. Esta restricción, que evita las subdivisiones casi triviales, es tenida en cuenta solamente cuando el conjunto T es pequeño.

1.3. Redes Bayesianas

Una red bayesiana es un grafo acíclico dirigido en el que cada nodo representa una variable y cada arco una dependencia probabilística, en la cual se especifica la probabilidad condicional de cada variable dados sus padres, la variable a la que apunta el arco es dependiente (causa-efecto) de la que está en el origen de éste. La topología o estructura de la red nos da información sobre las dependencias probabilísticas entre las variables pero también sobre las independencias condicionales de una variable (o conjunto de variables) dada otra u otras variables, dichas independencias, simplifican la representación del conocimiento (menos parámetros) y el razonamiento (propagación de las probabilidades).
El obtener una red Bayesiana a partir de datos, es un proceso de aprendizaje que se divide en dos etapas: el aprendizaje estructural y el aprendizaje paramétrico [Pearl, 1988; Hernández O.J. et al, 2004]. La primera de ellas, consiste en obtener la estructura de la red bayesiana, es decir, las relaciones de dependencia e independencia entre las variables involucradas. La segunda etapa, tiene como finalidad obtener las probabilidades a priori y condicionales requeridas a partir de una estructura dada.
Estas redes [Pearl, 1988; Hernández O.J. et al, 2004] son utilizadas en diversas áreas de aplicación como por ejemplo en medicina [Beinlinch *et al.*, 1989; Hernández O.J. et al, 2004], ciencia [Breese & Blake, 1995; Hernández O.J. et al, 2004], y economía [Hernández O.J. et al, 2004]. Las mismas 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 [Heckerman *et al.*, 1995; Díaz & Corchado, 1999; Hernández O.J. et al, 2004] y pueden manejar bases de datos incompletas [Heckerman, 1995; Heckerman & Chickering, 1996; Ramoni & Sebastiani, 1996; Hernández O.J. et al, 2004].

A continuación se describirá el funcionamiento de los clasificadores naïve Bayes, TAN y KDB.

1.3.1 Descripción de los algoritmos de aprendizaje

1.3.1.1. Introducción al Clasificador naïve Bayes

El paradigma clasificatorio en el que se utiliza el teorema de Bayes en conjunción con la hipótesis de independencia condicional de las variables predictoras dada la clase que se conoce bajo diversos nombres que incluye los de idiota Bayes [Ohmann y Col, 1988; Larrañaga. P. et al, 2004], naïve Bayes [Kononenko, 1990; Larrañaga. P. et al, 2004], simple Bayes [Gammerman y Thatcher, 1991; Larrañaga. P. et al, 2004] y Bayes independiente [Todd y Stamper, 1994; Larrañaga. P. et al, 2004].

A pesar una larga tradición en la comunidad de reconocimiento de patrones [Duda y Hart, 1973; Larrañaga. P. et al, 2004] el clasificador Naïve Bayes aparece por primera vez en la literatura del aprendizaje automático a finales de los ochenta [Cestnik y Col, 1987; Larrañaga. P. et al, 2004] con el objetivo de comparar su capacidad predictiva con la de métodos mas sofisticados. De manera gradual los investigadores de esta comunidad de aprendizaje automático se han dado cuenta de su potencialidad y robustez en problemas de clasificación supervisada.

En esta sección se va a afectar una revisión del paradigma naïve *Bayes*, el cual debe su nombre a las hipótesis tan simplificadoras – independencia condicional de las variables predictoras dada la variable clase- sobre las que se construye dicho clasificador. Partiremos del paradigma clásico de diagnóstico para, una vez comprobado que necesita de la estimación de un número de parámetros ingente, ir simplificando paulatinamente las hipótesis sobre las que se construye hasta llegar al modelo naïve Bayes. Veremos a continuación un resultado teórico que nos servirá para entender mejor las características del clasificador naïve Bayes.

1.3.1.2. Del Paradigma Clásico de Diagnóstico al Clasificador Naïve Bayes

Vamos y comenzar recordando el teorema de Bayes con una formulación de sucesos, para posteriormente formularlo en términos de variables aleatorias. Una vez visto el teorema de Bayes, se presenta el paradigma clásico de diagnóstico. Viéndose 1a necesidad de ir simplificando las premisas sobre las que se construye en aras de obtener paradigmas que puedan ser de aplicación para la resolución de problemas reales. El contenido de este apartado resulta ser una adaptación del material que Díez y Nell (1998) dedican al mismo.

Teorema (Bayes, 1764) Sean A y B dos sucesos aleatorios cuyas probabilidades se denotan por p(A) y p(B) respectivamente, verificándose que p(B) > 0. Supongamos conocidas las probabilidades a priori de los sucesos A y B, es decir, p(A) y p(B), así como la probabilidad condicionada del suceso B dado el suceso A, es decir p(B|A). La probabilidad a posteriori del suceso A conocido que se verifica el suceso B, es decir p(A|B), puede calcularse a partir de la siguiente fórmula:

$$p(A|B) = \frac{p(A, B)}{p(B)} = \frac{p(A)\, p(B|A)}{p(B)} = \frac{p(A)\, p(B|A)}{\Sigma_{A'}\, p(A')\, p(B|A')}$$

La formulación del teorema de Bayes puede efectuarse también para variables aleatorias, tanto unidimensionales como multidimensionales.

Comenzando por la formulación para dos variables aleatorias unidimensionales que denotamos por X e Y, tenemos que:

$$P(Y = y \mid X = x) = \frac{P(Y = y)\, p(X = x \mid Y = y)}{\Sigma_{Y'}\, p(Y = y')\, p(X = x \mid Y = y')}$$

El teorema de Bayes también puede ser expresado por medio de una notación que usa el número de componentes de cada una de las variables multidimensionales anteriores X e Y, de la siguiente manera:

$$P(Y = y \mid X = x) = p(Y_1 = y_1, \ldots\ldots, Y_m = y_m \mid X_1 = x_1, \ldots\ldots, X_m = x_m)$$

$$= \frac{p(Y_1 = y_1, \ldots\ldots, Y_m = y_m)\; p(X_1 = x_1, \ldots\ldots, X_m = x_m \mid Y_1 = y_1, \ldots\ldots, Y_m = y_m)}{\Sigma_{y'1 \ldots y'n}\, p(X_1 = x_1, \ldots, X_m = x_m \mid Y_1 = y'_1, \ldots\ldots, Y_m = y'_m)\; p(Y_1 = y'_1, \ldots, Y_m = y'_m)}$$

En el problema de clasificación supervisada reflejado en la tabla 1, tenemos que Y = C es una variable unidimensional, mientras que X = ($X_1 \ldots X_n$) es una variable n-dimensional.

	X_1	.....	X_n	Y
$(x^{(1)}, y^{(1)})$	$x_1^{(1)}$	.....	$x_n^{(1)}$	$y^{(1)}$
$(x^{(2)}, y^{(2)})$	$x_1^{(2)}$	.....	$x_n^{(2)}$	$y^{(2)}$
.....		.....		
$(x^{(n)}, y^{(n)})$	$x_1^{(n)}$	.....	$x_n^{(n)}$	$y^{(n)}$

Tabla 1: Problema de clasificación supervisada.

Vamos a plantear la formulación clásica de un problema de diagnóstico utilizando una terminología habitual en medicina. Es evidente que la terminología puede trasladarse a otras ramas de la ciencia y de la técnica, en particular a la ingeniería. La terminología a usar incluye términos como:

Hallazgo, con el cual nos referiremos a la determinación del valor de una variable predictora X_r. Así por ejemplo x_r (Valor de la variable X_r) puede estar representando la existencia de vómitos en un determinado enfermo.

Evidencia, denota el conjunto de todos los hallazgos para un determinado individuo. Es decir x ($x_1, \ldots x_n$) puede estar denotando (si n = 4) que el individuo en cuestión es joven, hombre, presenta vómitos y además no tiene antecedentes familiares.

Diagnóstico, denota el valor que toman las m variables aleatorias Y_1, ... Y_m, cada una de las cuales se refiere a una enfermedad.

Probabilidad a priori del diagnóstico, p(y) o $p(Y_1 = y_1, \ldots\ldots, Y_m = y_m)$, se refiere a la probabilidad de un diagnóstico concreto, cuando no se conoce nada acerca de los hallazgos, es decir, cuando se carece de evidencia.

Probabilidad a posteriori de un diagnóstico, p(ylx) o $p(Y_1 = y_1, \ldots\ldots, Y_m = y_m \mid X_1 = x_1, \ldots\ldots, X_m = x_m)$, es decir, la probabilidad de un diagnóstico concreto cuando se conocen n hallazgos (evidencia).

En el planteamiento clásico del diagnóstico (véase Tabla 2) se supone que los m diagnósticos posibles son no excluyentes, es decir, pueden ocurrir a la vez, siendo cada uno de ellos dicotómico. Para fijar ideas en relación con el ámbito médico, podemos pensar que cada uno de los m posibles diagnósticos no excluyentes se relaciona con una enfermedad, pudiendo tomar dos valores: 0 (no existencia) y 1 (existencia). Por lo que se refiere a los n hallazgos o síntomas, se representarán por medio de 1as n variables aleatorias X_1,.....X_n y también asumiremos que cada variable predictora es dicotómica, con valores 0 y 1. El valor 0 en la variable X_i indica la ausencia del i-ésimo hallazgo o síntoma mientras que el valor 1 indica la presencia del hallazgo o síntoma correspondiente.

	X_1	.....	X_n	Y_1	.....	Y_m
$(x^{(1)}, y^{(1)})$	$x_1^{(1)}$	.....	$x_n^{(1)}$	$y_1^{(1)}$	.....	$y_m^{(1)}$
$(x^{(2)}, y^{(2)})$	$x_1^{(2)}$	.....	$x_n^{(2)}$	$y_1^{(2)}$	.....	$y_m^{(2)}$
.....		.....			.....	
$(x^{(N)}, y^{(N)})$	$x_1^{(N)}$	.....	$x_n^{(N)}$	$y_1^{(N)}$	.....	$y_m^{(N)}$

Tabla 2: Problema clásico de diagnóstico.

El problema del diagnóstico consiste en encontrar el diagnóstico más probable a posteriori, una vez conocido el valor de la evidencia. En notación matemática el diagnóstico óptimo, $(y_1^*, ..., y_m^*)$ será aquel que verifique:

$$(y_1^*,....,y_m^*) = \arg \max_{(y1,...yn)} p(Y_1 = y_1,....., Y_m = y_m \mid X_1 = x_1,....., X_m = x_m)$$

Aplicando el teorema de Bayes para calcular $p(Y_1 = y_1,..., Y_m = y_m \mid X_1 = x_1,.., X_m = x_m)$ obtenemos:

$$p(Y_1 = y_1 ,.....,Y_m = y_m \mid X_1 = x_1 ,.....,X_m = x_m)$$

$$= \frac{p(Y_1 = y_1 ,.....,Y_m = y_m)\ p(X_1 = x_1 ,.....,X_m = x_m \mid Y_1 = y_1 ,.....,Y_m = y_m)}{\Sigma_{y'1....y'n}\ p(Y_1 = y'_1 ,....,Y_m = y'_m)\ p(X_1 = x_1 ,....,X_m = x_m \mid Y_1 = y'_1 ,....,Y_m = y'_m)}$$

Veamos a Continuación el número de parámetros que se deben estimar para poder especificar el paradigma anterior y de esa forma obtener e1 valor de $(y_1^*, ..., y_m^*)$. Es importante tener en cuenta que la estimación de cada uno de los parámetros anteriores se deberá efectuar a partir del archivo de N casos, reflejado en la Tabla 4.

Para estimar $p(Y_1 = y_1 ,.....,Y_m = y_m)$, y teniendo en cuenta que cada variable Y_1 es dicotómica, necesitaremos un total de 2^m - 1 parámetros. De igual forma, por cada una de las distribuciones de probabilidad condicionada, $p(X_1 = x_1 ,.....,X_m = x_m \mid Y_1 = y'_1 ,.....,Y_m = y'_m)$, se necesitan estimar 2^n - 1 parámetros. Al tener un total de 2^m de tales distribuciones de probabilidad condicionadas, debemos estimar $(2^n - 1)\ 2^m$ parámetros. Es decir, que el número total de parámetros necesarios para determinar un modelo concreto del paradigma clásico de diagnóstico es: $2^n - 1 + 2^m\ (2^n - l)$. Para hacernos una idea del número de parámetros a estimar podemos consultar la Tabla 3, en la cual vemos de manera aproximada el número de parámetros a estimar para distintos valores de m (número de enfermedades) y n (número de hallazgos).

M	n		Parámetros
3	10	≈	8.10^3
5	20	≈	33.10^6
10	50	≈	11.10^{17}

Tabla 3: Número de parámetros a estimar, en función de m (número de enfermedades) y n (número de síntomas), en el paradigma clásico de diagnóstico.

Ante la imposibilidad de poder estimar el elevado número de parámetros que se necesitan en el paradigma clásico de diagnóstico, en lo que sigue se simplificaran las premisas sobre las que se ha construido dicho paradigma.

En primer Lugar vamos a considerar que los diagnósticos son excluyentes, es decir, que dos diagnósticos no pueden darse al unísono. Esto trae como consecuencia que en lugar de considerar el diagnóstico como una variable aleatoria m-dimensional, este caso puede verse como una única variable aleatoria unidimensional siguiendo una distribución polinomial con m valores posibles.

Vamos a denotar por $X_1,....X_n$ a las n variables predictoras. Supongamos que todas ellas sean binarias. Denotamos por C la variable de diagnóstico, que suponemos puede tomar m posibles valores. La tabla 4 refleja la situación anterior. La búsqueda del diagnostico mas probable a posteriori, c^*, una vez conocidos los síntomas de un determinado paciente, $x = (x_1,....,x_n)$, puede plantearse como la búsqueda del estado de la variable C con mayor probabilidad a posteriori. Es decir

$$c^* = \arg \max_c \; p(C = c \mid X_1 = x_1,....., X_m = x_m)$$

El cálculo de $p(C = c \mid X_1 = x_1,....., X_m = x_m)$ puede llevarse a cabo utilizando el teorema de Bayes, y ya que el objetivo es calcular el estado de C, c^*, con mayor probabilidad a posteriori, no es necesario calcular el denominador del teorema de Bayes. Es decir,

$$p(C = c \mid X_1 = x_1,....., X_m = x_m) \propto p(C=c)\, p(X_1 = x_1,....., X_m = x_m \mid C = c)$$

Por tanto, en el paradigma en el que los distintos diagnósticos son excluyentes, y considerando que el número de posibles diagnósticos es m, y que cada variable predictora X_i es dicotómica, tenemos que el número de parámetros a estimar es $(m - 1) + m (2^n - 1)$, de los cuales:

- m - 1 se refiere a las probabilidades a priori de las variable C;
- m (2n - 1) se relacionan con las probabilidades Condicionadas de cada posible combinación de las variables predictoras dado cada posible valor de la variable.

La Tabla 4 nos da una idea del número de parámetros a estimar para distintos valores de m y n.

M	n		Parámetros
3	10	≈	3.10^3
5	20	≈	5.10^6
10	50	≈	11.10^{15}

Tabla4: Número de parámetros a estimar, en función de m (número de enfermedades) y n (número de síntomas), en el paradigma clásico de diagnóstico con diagnósticos excluyentes.

Vemos de nuevo que el número de parámetros a estimar sigue siendo elevado, de hay que necesitamos imponer suposiciones más restrictivas para que los paradigmas puedan convertirse en modelos implementables.

Vamos finalmente a introducir el paradigma naïve Bayes: diagnósticos excluyentes y hallazgos condicionalmente independientes dado el diagnóstico. El paradigma naïve Bayes se basa en dos premisas establecidas sobre las variables predictoras (hallazgos, síntomas) y la variaba a predecir (diagnósticos). Dichas premisas son:

Los diagnósticos son excluyentes, es decir, la variable C a predecir toma uno de sus m posibles valores: c_l,.....,c_m;

Los hallazgos son coincidentalmente independientes dado el diagnóstico, es decir, que si una conoce, el valor de 1a variable diagnóstico, el conocimiento del valor de cualquiera de los hallazgos es irrelevante para el resto de los hallazgos. Esta condición se expresa matemáticamente por medio de la fórmula:

$$p(X_1 = x_1,....., X_m = x_m \mid C = c) = \prod_{i=1}^{n} p(X_i = x_i \mid C = c)$$

ya que por medio de la regla de la cadena se obtiene:

$$\begin{aligned} p(X_1 = x_1,....., X_m = x_m \mid C = c) = \quad & p(X_1 = x_1 \mid X_2 = x_2,....., X_n = x_n, C = c) \\ & p(X_2 = x_2 \mid X_3 = x_3,....., X_n = x_n, C = c) \\ & .....p(X_n = x_n \mid C = c) \end{aligned}$$

Por otra parte teniendo en cuenta la independencia condicional entre las variables predictoras dada la variable clase, se tiene que:

$p(X_i = x_i \mid X_{i+1} = x_{i+1},....., X_n = x_n, C = c) = p(X_i = x_i \mid C = c)$

Por todo I = 1,….,n. De ahí que se verifique la ecuación.

Por tanto, en el paradigma naïve Bayes, la búsqueda del diagnóstico más probable, c^*, una vez conocidos los síntomas (x_l,.....x_n) determinado paciente, se reduce a:

$$c^* = \arg \max_{c} \; p(C = c \mid X_1 = x_1,....., X_m = x_m)$$

$$= c^* = \arg \max_{c} \; p(C = c) \prod_{i=1}^{n} p(X_i = x_i \mid C = c)$$

Suponiendo que todas las variables predictoras son dicotómicas, el número de parámetros necesarios para especificar un modelo naïve Bayes resulta ser (m – 1) + m n, ya que: Se necesitan (m - 1) parámetros para especificar la probabilidad a priori de la variable C. Para cada variable predictora X_i se necesitan m parámetros para determinar las distribuciones de probabilidad condicionadas.

Con los números reflejados en la tabla 5 nos podemos hacer una idea del número de parámetros necesarios en función del número de posibles diagnósticos y del número de síntomas necesarios para especificar el paradigma naïve Bayes.

M	n	parámetros
3	10	32
5	20	104
10	50	509

Tabla 5: Número de parámetros a estimar en el paradigma naïve Bayes en función del número de diagnósticos posibles (m) y del número de síntomas (n).

En e1 caso de que las n variables predictoras $X_l,.....,X_n$ sean continuas, se tiene que el paradigma naïve Bayes se convierte en buscar el valor de la variable C, que denotamos por c^*, que maximiza la probabilidad a posteriori de la variable C, dada la evidencia expresada como una instanciación de las variables $X_1,....,X_n$, esto es, $x = (x_1,....,x_n)$.

Es decir, el paradigma naïve Bayes con variables continuas trata de encontrar c^* verificando:

$$c^* = \arg \max_c \; p(C = c \mid X_1 = x_1,....., X_n = x_n)$$

$$= c^* = \arg \max_c \; p(C = c) \prod_{i=1}^{n} f X_i \mid C = c(x_i \mid c)$$

donde $f X_i \mid C = c(x_i \mid c)$ denota, para todo i = 1,.....,n, la función de densidad de la variable X_i condicionada a que el valor del diagnóstico sea c.

Suele ser habitual utilizar una variable aleatoria normal (para cada valor de C) para modelar el comportamiento de la variable X_i. Es decir, para todo c, y para todo $i \in \{1,.....,n\}$, asumimos

$$f X_i \mid C = c(x_i \mid c) \sim N(x_i ; \mu_i^c \; ; \; (\sigma_i^c)^2)$$

En tal caso el paradigma naïve Bayes obtiene c^* como:

$$c^* = \arg \max_c \; p(C = c) \prod_{i=1}^{n} \left| \frac{1}{\sqrt{2\pi i \sigma_i^c}} \; e^{-1/2 ((xi\, \mu\, ic) / \sigma ic)2} \right|$$

En este caso el número de parámetros a estimar es (m – 1) + 2mn:

- m - 1 en relación con las probabilidades a priori p(C = c)
- 2nm en relación con las funciones de densidad condicionadas

Finalmente puede ocurrir que algunos de los hallazgos se recojan en variables discretas mientras que otros hallazgos sean continuos. En tal caso hablaremos del paradigma naïve Bayes con predictoras continuas y discretas.

Supongamos que de las n variables predictoras, n_1 de ellas, $X_1,....,X_n$, sean discretas, mientras que el resto $n – n_1 = n_2$, $Y_1,.....Y_{n2}$, sean continuas. En principio al aplicar directamente la formula del paradigma naïve Bayes correspondiente a esta situación se obtiene:

$$p(c \mid x_1,.....,x_{m1},y_1,.....,y_{m2}) \propto p(c) \prod_{i=1}^{n_1} p(x_i \mid c) \prod_{j=1}^{n_2} f(y_i \mid c)$$

Esta expresión puede propiciar el conceder una mayor importancia a las variables continuas, ya que mientras que $p(x_i \mid c)$ verifica $0 <= p(x_i \mid c) >= 1$, puede ocurrir que $f(y_i \mid c) > 1$. Con objeto de evitar esta situación, proponemos la normalización de la aportación de las variables continuas dividiendo cada uno de los factores correspondientes por el $máx_y f(y_j \mid c)$. Obtenemos por tanto:

$$p(c \mid x_1,.....,x_{m1},y_1,.....,y_{m2}) \propto p(c) \prod_{i=1}^{n_1} p(x_i \mid c) \prod_{j=1}^{n_2} \frac{f(y_i \mid c)}{máx_{xy} f(y_j \mid c)}$$

En el caso en que las funciones de densidad de las variables continuas condicionadas a cada posible valor de la variable clase sigan distribuciones normales, es decir si $Y_j \mid C = c \sim N(y_j ; \mu_c^j, (\sigma_c^j)^2)$, se tiene que la formula se expresa de la siguiente manera:

$$p(c \mid x_1,.....,x_{n1},y_1,.....,y_{n2}) \propto p(c) \prod_{i=1}^{n_1} p(x_i \mid c) \prod_{j=1}^{n_2} e^{-1/2 (y_j - \mu_c^j / \sigma_c^j)^2}$$

La Figura 3: refleja la estructura gráfica de un modelo naïve Bayes.

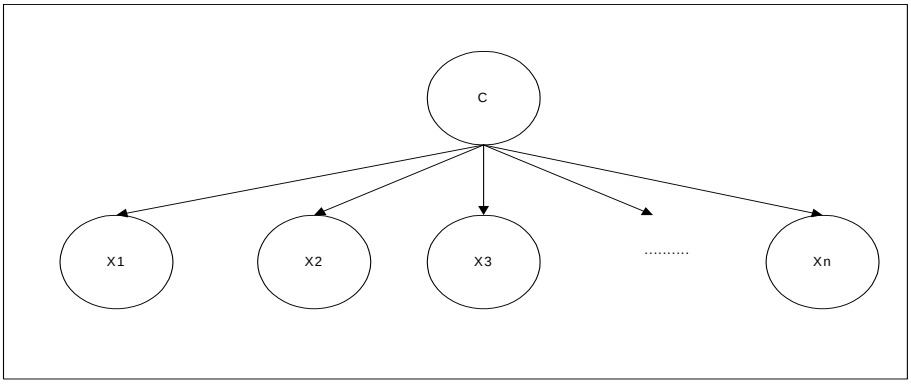

Figura 3: Estructura gráfica de un modelo naïve Bayes.

1.3.1.3. Naïve Bayes Aumentado a Árbol (TAN)

En esta sección vamos a presentar algunos trabajos que construyen clasificadores con estructura naïve Bayes aumentada a árbol (Tree Augmented Network (TAN)). Para obtener este tipo de estructura se comienza por una estructura de árbol con las variables predictoras, para posteriormente conectar la variable clase con cada una de las variables predictoras. La Figura 4 ilustra un ejemplo de estructura naïve Bayes aumentada a árbol.

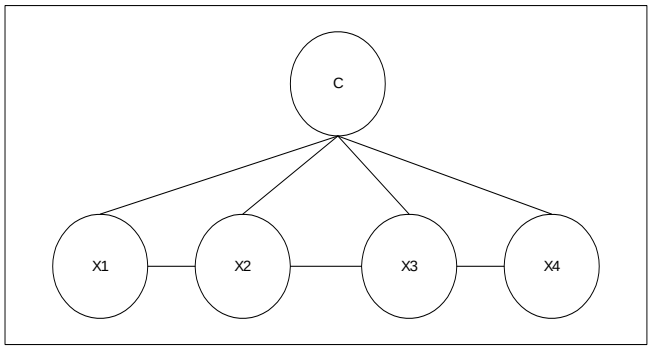

Figura 4: Ejemplo de estructura naïve Bayes aumentada a árbol (Tree Augmented Network (TAN))

Fridman y Col. (1997) presentan un algoritmo denominado Tree Augmented Network (TAN) el cual consiste básica en una adaptación del algoritmo de Chow-Liu (1968). En dicho algoritmo se tiene en cuenta la cantidad de información mutua condicionada a la variable clase, en lugar de la cantidad de información mutua en la que se basa el algoritmo de Chow-Liu. La cantidad de información mutua entre las variables discretas X e Y condicionada a la variable C se define como:

$$I(X, Y \mid C) = \sum_{i=1}^{} \sum_{j=1}^{} \sum_{r=1} p(x_i, y_j, c_r) \log \frac{p(x_i, y_j, c_r)}{p(x_i \mid c_r)\, p(y_j \mid c_r)}$$

Tal y como puede verse en el pseudocódigo de la Figura 5, TAN consta de cinco pasos. En el primer paso se calculan las cantidades de información mutua para cada par de variables (X_i, X_j) condicionadas por la variable C. A Continuación se debe construir un grafo no dirigido completo con n nodos, uno por cada una de las variables predictoras, en el cual el peso de cada arista viene dado por la cantidad de información mutua entre las dos variables unidas por la arista condicionada a la variable clase. El algoritmo de Kruskall parte de los n (n – 1) / 2 pesos obtenidos en el paso anterior para construir el árbol expandido de máximo peso de la siguiente manera:

1. Asignar las dos aristas de mayor peso al árbol a construir.
2. Examinar la siguiente arista de mayor peso, y añadirla al árbol a no ser que forme un ciclo, en cuyo caso se descarta y se examina la siguiente arista de mayor peso.
3. Repetir el paso 2 hasta que se hayan seleccionado n - 1 aristas.

Las propiedades teóricas de este algoritmo de construcción de TAN son análogas a las del algoritmo de Chow-Liu (1968). Es decir, si los datos han sido generados por una estructura Tree Augmented Network, el algoritmo TAN es asintóticamente correcto, en el sentido de que si la muestra de casos es suficientemente grande, recuperará la estructura que generó el archivo de casos. En la Figura 5a se muestra un ejemplo de aplicación del algoritmo.

Paso 1. Calcular $I(X_i, X_j \mid C)$ con $i < j, i, j = 1, \ldots, n$
Paso 2. Construir un grafo no dirigido completo cuyos nodos correspondan a las variables predictoras $X_1, \ldots, X_n$. Asignar a cada arista conectando las variables X_i y X_j un peso dado por $I(X_i, X_j \mid C)$
Paso 3. A partir del grafo completo anterior y siguiendo el algoritmo de Kruskal construir un árbol expandido de máximo peso
Paso 4. transformar el árbol no dirigido resultante en uno dirigido, escojiendo una variable como raíz, para a continuación direccionar el resto de las aristas
Paso 5. Construir un modelo TAN añadiendo un nodo etiquetado como C y posteriormente un arco desde C a cada variable predictora X_i

Figura 5: Pseudocódigo del algoritmo TAN (Friedman y Col, 1997).

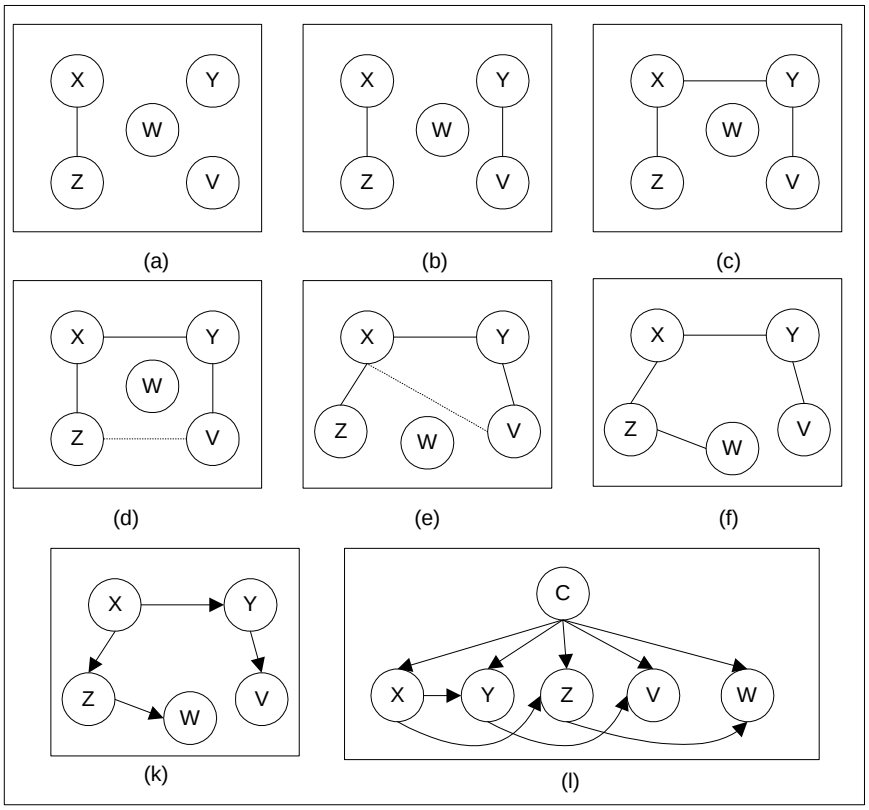

Figura 5a: Ilustración del algoritmo TAN con cinco variables predictoras X, Y, Z, V y W. Se supone que el orden de las cantidades de información mutuas condicionadas ha sido:
I(X, Z | C) > I(Y,V | C) > I(X,Y | C) > I(Z,v | C) > I(X,V | C) > I(Z,W | C) > I(X,Y | C) > I(X,W | C) > I(Y, Z | C) > I(Y, W | C) > I(V, W | C). Las subfiguras (a) a (f) corresponden a la aplicación del algoritmo de Kruskall. La subfigura (g) corresponde al Paso 4 del algoritmo TAN y finalmente en la subfigura (h) se realiza el Paso 5 de TAN. El modelo clasificatorio obtenido es: $P(C \mid x, y, z, v, w) \alpha\ p(c)\ p(x \mid c)\ p(y \mid x,c)\ p(z \mid x,c)\ p(v \mid y,c)\ p(w \mid z,c)$

Keogh y Pazzani (1999) proponen un algoritmo voraz que va añadiendo arcos a una estructura naïve Bayes. En cada paso se añade el arco que, mantiene la condición de que en la estructura final cada variable no tenga más de un padre, mejore en mayor medida el porcentaje de bien clasificados obtenido mediante el mismo.

1.3.1.4. Clasificadores Bayesianos k Dependientes (KDB)

Sahami (1996) presenta un algoritmo denominado k Dependence Bayesian classifier (KDB) el cual posibilita atravesar el amplio espectro de dependencias disponibles entre el modelo naïve Bayes y el modelo correspondiente a una red Bayesiana completa. El algoritmo se fundamenta en el concepto de clasificador Bayesiano K-

dependiente, el cual contiene la estructura del clasificador naïve Bayes y permite a cada variable predictora tener un máximo de K variables padres sin contar a la variable clase. De esta manera, el modelo naïve Bayes se corresponde con un clasificador Bayesiano 0-dependiente, el modelo TAN sería un clasificador Bayesiano 1-dependiente y el clasificador Bayesiano completo (en la estructura no se refleja ninguna independencia) correspondería a un clasificador Bayesiano (n - 1)- dependiente. El pseudocódigo del algoritmo KDB puede consultarse en la Fig. 6.

Paso 1. Para cada variable predictora X_i i = 1,....,n calcular la cantidad de información mutua con respecto a la clase C, I (X_i ,C)
Paso 2. Para cada par de variables predictoras calcular la cantidad de información mutua condicionada a la clase, I (X_i , X_j | C), cin i ≠ j, i, j = 1,...., n
Paso 3. Inicializar a vació la lista de variables usando *N*
Paso 4. Inicializar la red Bayesiana a construir, BN, con un único nodo, el correspondiente a la variable C
Paso 5. Repetir hasta que *N* incluya a todas las variables del dominio:
Paso 5.1. Seleccionar de entre las variables que no están en *N*, aquella X_{max} con mayor cantidad de información mutua respecto a C, $I(X_{max}, C) = \max_{X \notin N} I(X, C)$
Paso 5.2. Añadir un nodo a BN, $X \notin N$ representando X_{max}
Paso 5.3. Añadir un arco C a X_{max} en BN
Paso 5.4. Añadir m = mín(|*N*|, k) arcos de las m variables distintas X_j en que tengan los mayores valores $I(X_{max}, X_j |C)$
Paso 5.5. Añadir X_{max} a *N*
Paso 6. Computar las probabilidades condicionadas necesarias para especificar la red Bayesiana BN

Figura 6: Pseudocódigo del algoritmo KDB (Sahami, 1996).

La idea básica del algoritmo consiste en generalizar el algoritmo propuesto por Fridman y Col (1997) permitiendo que cada variable tenga un número de padres, sin contar la variable clase C, acotado por k. El autor comenta una posible mejora del algoritmo flexibilizando la determinación de k por medio de la obtención de un umbral de cantidad de información mutua, el cual debería de se sobrepasarlo para que el correspondiente arco fuese incluido.

2. Obtención de redes bayesianas a través de Elvira

El programa Elvira es fruto de un proyecto de investigación financiado por la CICYT y el Ministerio de Ciencia y Tecnología Español, en el que participan investigadores de varias universidades españolas y de otros centros.

El programa Elvira está destinado a la edición y evaluación de modelos gráficos probabilistas, concretamente redes bayesianas y diagramas de influencia. Elvira cuenta con un formato propio para la codificación de los modelos, un lector-intérprete para los modelos codificados, una interfaz gráfica para la construcción de redes, con opciones específicas para modelos canónicos (puertas OR, AND, MAX, etc.), algoritmos exactos y aproximados (estocásticos) de razonamiento tanto para variables discretas como continuas, métodos de explicación del razonamiento, algoritmos de toma de decisiones, aprendizaje de modelos a partir de bases de datos, fusión de redes, etc. Elvira está escrito y compilado en el lenguaje Java, lo cual permite que funcione en diferentes plataformas y sistemas operativos (MS-DOS/Windows, linux, Solaris, etc.).

2.1. Uso de Elvira

Este programa permite el ingreso de las redes Bayesianas de dos formas: (a) por un lado el ingreso manual, donde el usuario dibuja la red bayeasiana en la pantalla y carga los valores de probabilidad asociados a cada nodo, (b) mediante la importación de archivos de casos.

El presente trabajo se centrará en la generación de redes Bayesianas a partir de un archivo de datos, ya que esta operativa es la que se vincula con la minería de datos.

2.1.1. Primeros Pasos

El programa Elvira puede obtenerse gratis desde Internet a través de: http://leo.ugr.es/ elvira.

Para poder utilizarlos se debe instalar previamente “La Máquina Virtual” de Java, que será el encargado de interpretar el programa. Este Software puede obtenerse desde la página “www.java.com”**.,** una vez instalada la “Máquina Virtual”, lo que resta es descomprimir el archivo Elvira.zip que automáticamente generará una carpeta Elvira donde se le indique (por ejemplo C:); dentro de la carpeta Elvira aparecerá un archivo *Elvira.jar* el cual se ejecutará con solo hacerle doble clic .

Una vez iniciado Elvira aparecerá una pantalla como la que se muestra en la figura 7.

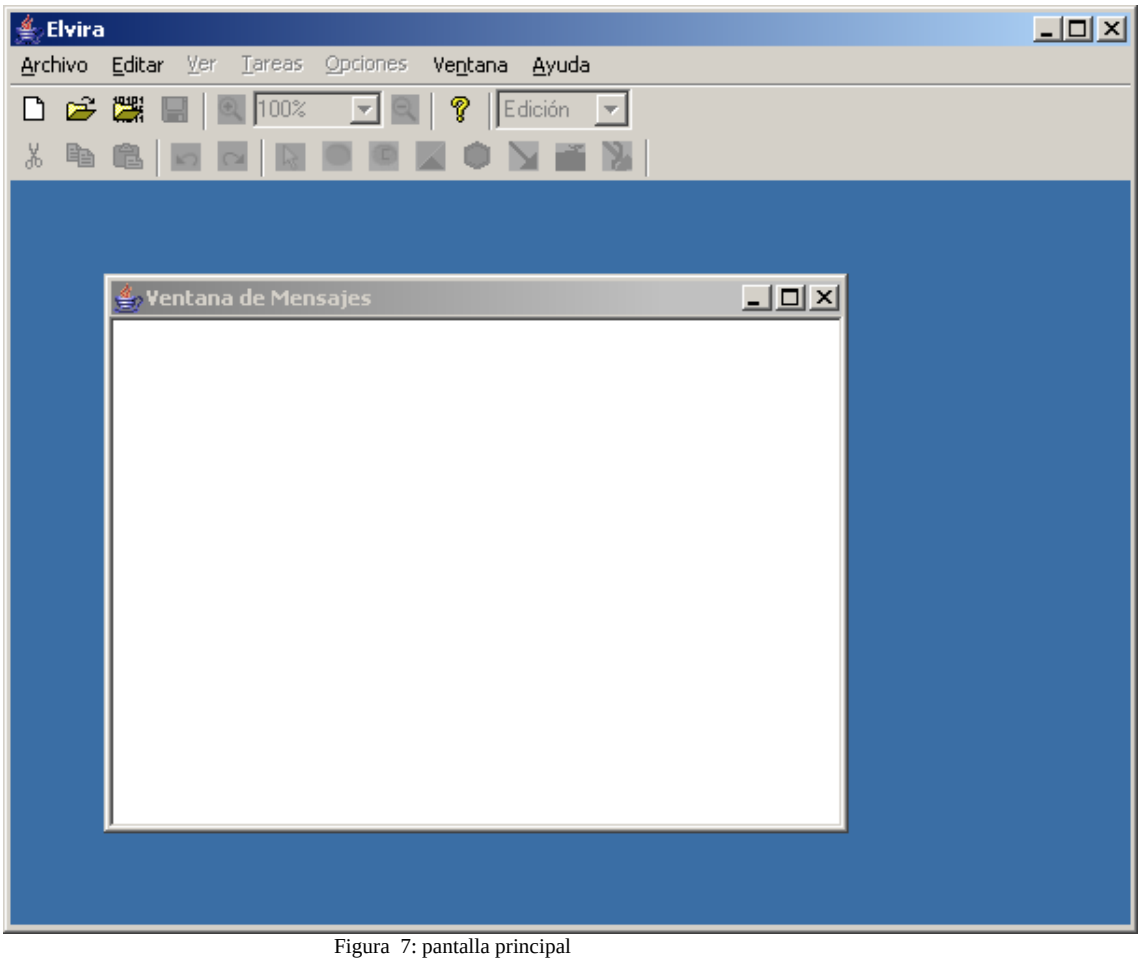

Figura 7: pantalla principal

Aquí la ventana de mensajes que aparece activa será utilizada para que el programa detalle mensaje a indicar al operador por inconvenientes con el uso del mismo.

A continuación vamos a importar un archivo de datos. Los archivos a importar deberán estar en formato .csv (texto plano con los campos dentro del mismo separados por ";" o ","). El sistema asume que el primer registro del archivo contiene el nombre de cada nodo, este punto es muy importante ya que si se omite el sistema asumirá como nombre de los nodos el valor de las variables indicadas en la primer fila..
A continuación, en la figura 8, se muestra un fragmento de un archivo a importar con Elvira, el mismo lleva en el primer registro el nombre de los campos que se detallan en los subsiguientes registros:

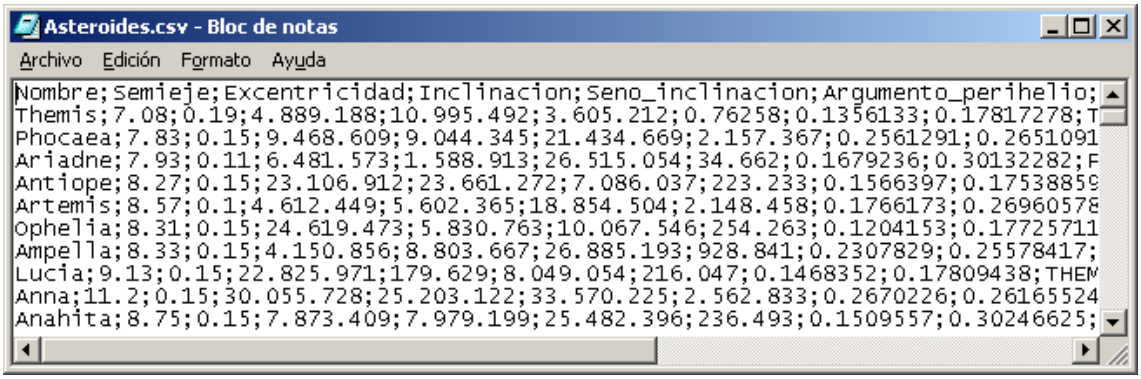

Figura 8: archivo de datos a importar

Para ingresar esta información al sistema se debe importar el archivo al formato .csv. para ello desde el menú “Archivo” se debe seleccionar la opción “Importar Fichero de Casos”, como se muestra a continuación en la figura 9:

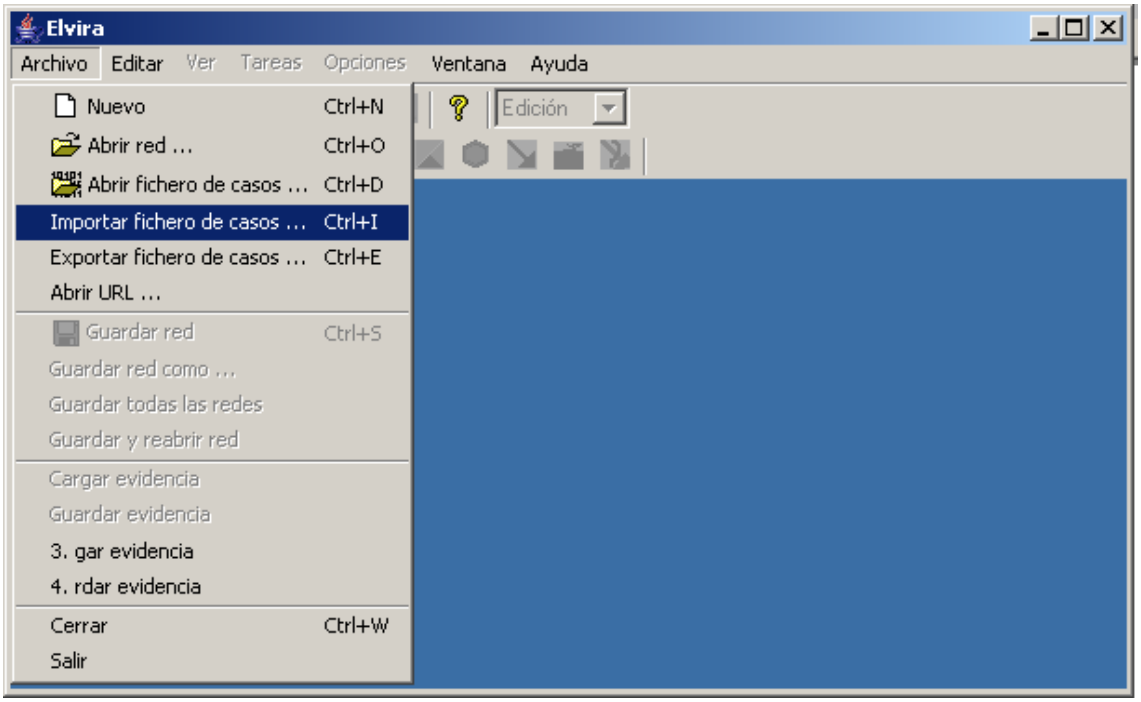

Figura 9: Ingreso a la opción de importación de fichero

Esta opción desplegará una ventana con tres solapas, a continuación, en la figura 10, se describe la primera de ellas:

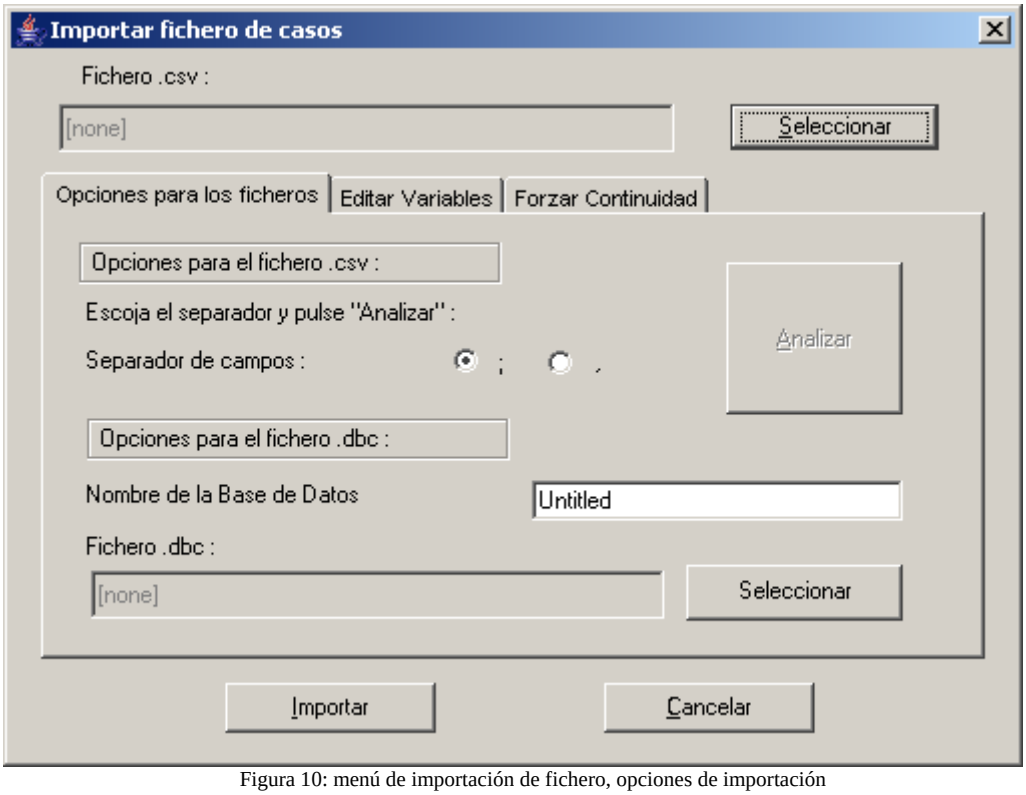

Figura 10: menú de importación de fichero, opciones de importación

El botón seleccionar permite indicar la ubicación del archivo a procesar. Opcionalmente desde esta solapa se podrá indicar el carácter de separación entre campos (“,” o “;”) y analizar el estado del archivo (mediante el botón analizar) para ver si el mismo podrá ser importado o no. Por último se debe seleccionar la ubicación y el nombre del nuevo archivo a crear como resultado de la importación de los datos.

Si el análisis previo del archivo da un resultado favorable, se está en condiciones de seleccionar el botón “Importar” para obtener el archivo .dbc, en caso contrario se podrá analizar la información a importar con mas precisión desde las otras dos solapas que se describen a continuación:

Solapa “editar variables”: Aquí en función del nombre que cada variable tenga en el primer registro del archivo se podrá modificar su nombre, editar sus posibles estados o ingresar algún comentario. A continuación, en la figura 11, se detalla el contenido de esta solapa:

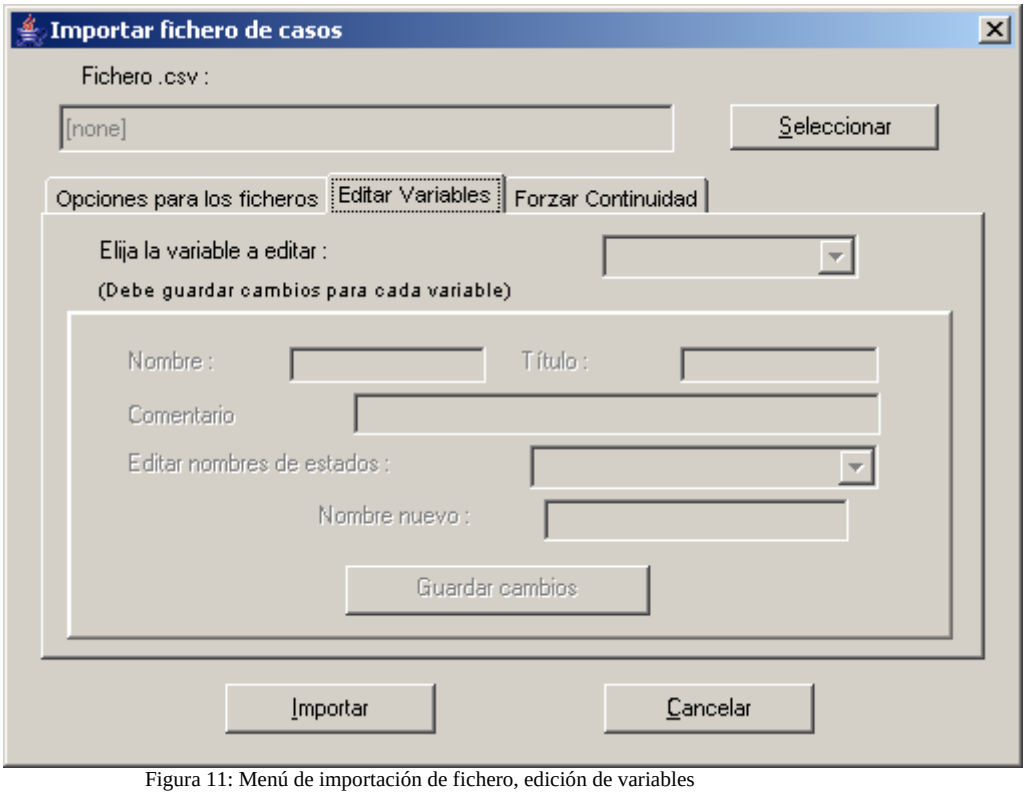

Figura 11: Menú de importación de fichero, edición de variables

Solapa “Forzar Continuidad”: Aquí se analizan las variables de forma similar a la solapa anterior y en caso de que se necesite forzar la continuidad de alguna o todas la variables a importar. A continuación, en la figura 12, se detalla el contenido de esta solapa:

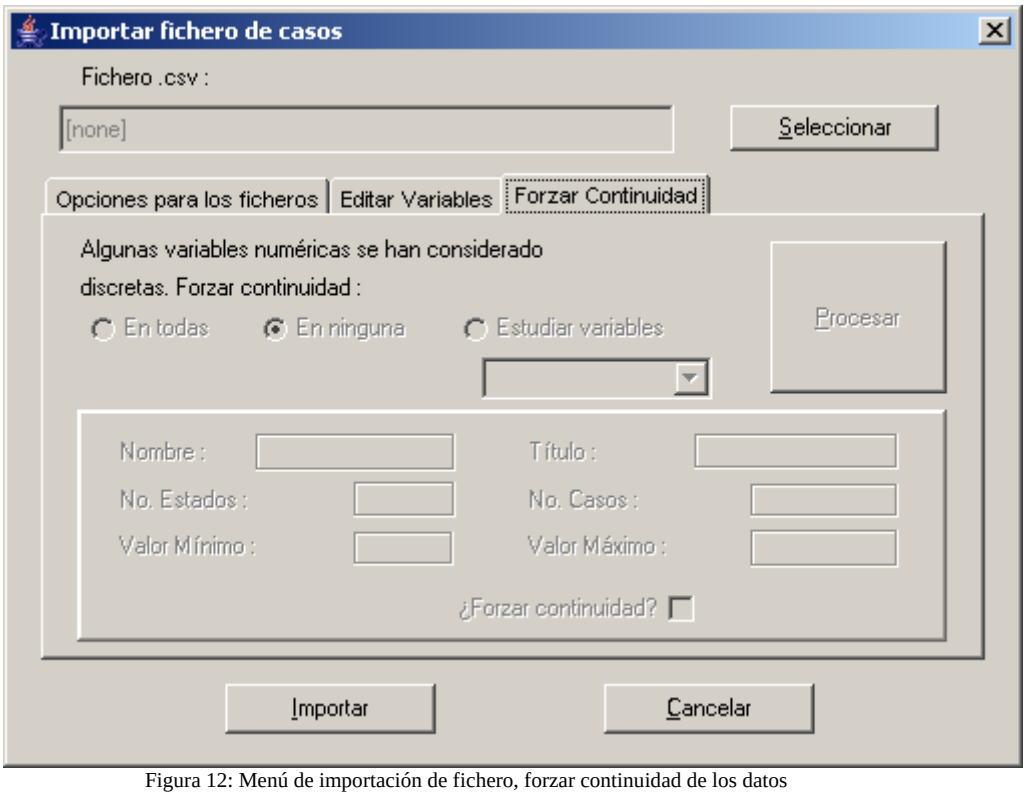

Figura 12: Menú de importación de fichero, forzar continuidad de los datos

A continuación, en la figura 13, se muestra el resultado de haber importado el archivo creditos.csv

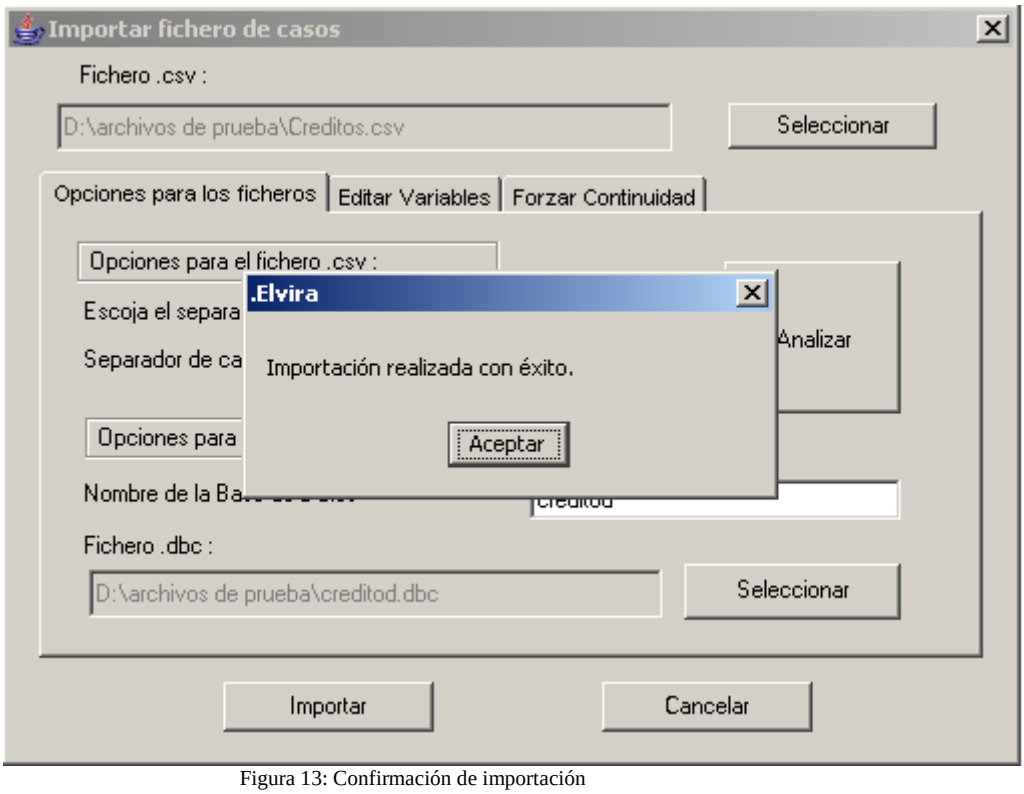

Figura 13: Confirmación de importación

Una vez importado el archivo a Formato .dbc podremos comenzar a trabajar con el. Para ello debemos ir nuevamente a el menú “Archivo” y aquí seleccionar la opción “Abrir fichero de Casos”, como se muestra a continuación en la figura 14

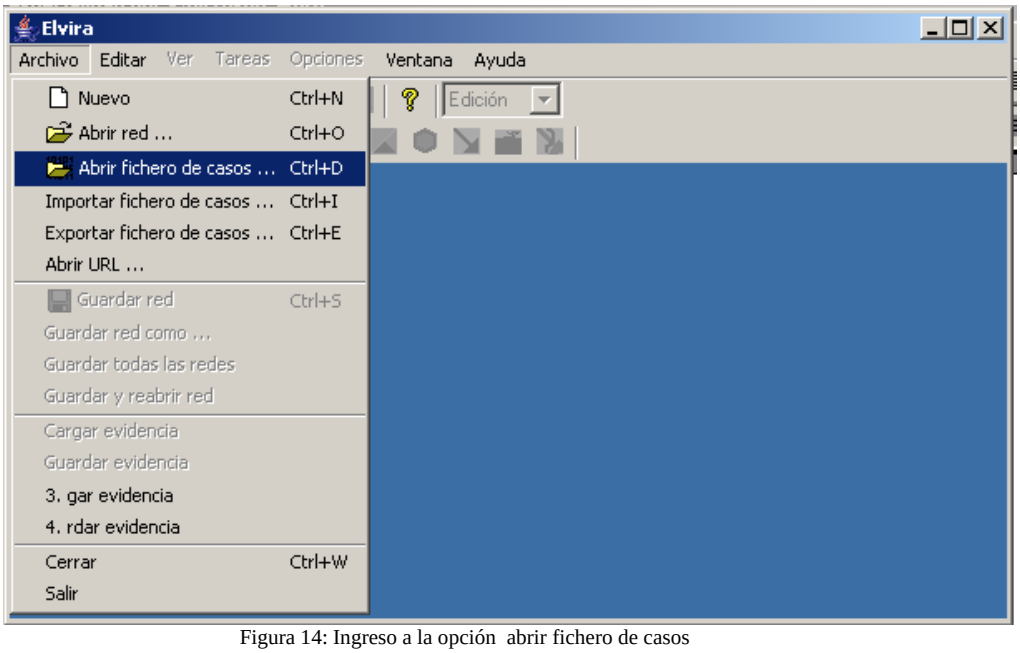

Figura 14: Ingreso a la opción abrir fichero de casos

En esta opción se deberá seleccionar en primer lugar el archivo .dbc a procesar, el cual debe ser ingresado, mediante su selección, en el campo “Fichero de Casos”. Esta pantalla además posee tres solapas, las cuales se describen a continuación:

- Opciones de preproceso: Aquí se encuentran los filtros que se pueden aplicar al archivo de casos para que luego pueda ser interpretado de forma exitosa por el clasificador.
- Aprendizaje automático: Aquí se encuentran los clasificadores a aplicar para realizar el aprendizaje (por ejemplo: Naïve Bayes, TAN o KDB)
- Opciones de Post aprendizaje: Aquí se pueden aplicar opciones de testeo de clasificación, como puede ser por ejemplo aplicar validaciones cruzadas.

A continuación, en la figura 15, se muestra la pantalla del programa:

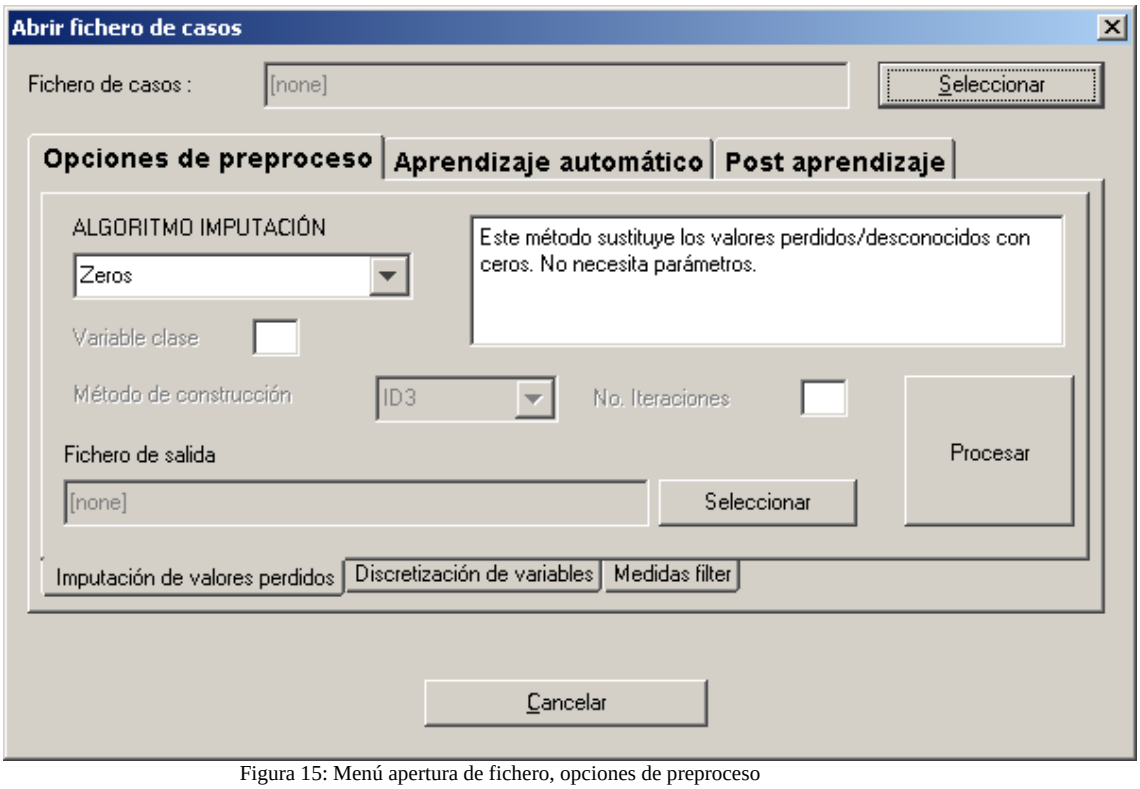

Figura 15: Menú apertura de fichero, opciones de preproceso

A partir de ahora centraremos la explicación en función del clasificador y el algoritmo de imputación a aplicar al archivo importado para generar la Red Bayesiana.

2.1.2. Elección del clasificador y algoritmos complementarios.

Los clasificadores son algoritmos dedicados a interpretar los datos provistos en el archivo de entrenamiento para generar la red Bayesiana. Actualmente existen varios algoritmos de este tipo, el presente trabajo solo analizará los soportados por el programa Elvira, estos fueron descriptos en el punto 1.2 y son Naïve Bayes, TAN y KDB.
Opcionalmente estos clasificadores pueden combinarse son con otros algoritmos, como por ejemplo el ID3, para obtener una red de mayor precisión en la red.

2.1.2.1.Caso 1 obtención de una red mediante el clasificador Naïve Bayes únicamente

Para este caso en la primer solapa no se hará selección alguna, solo se reemplazara por ceros la falta de valor en alguna de la variables a importar. Esto se hace con el fin de mostrar como opera el clasificador sin influencia de otros algoritmos que lo pueden complementar en la generación de la red Bayesiana. Por lo tanto pasaremos directamente a la solapa "Aprendizaje automático" (ver figura 16).

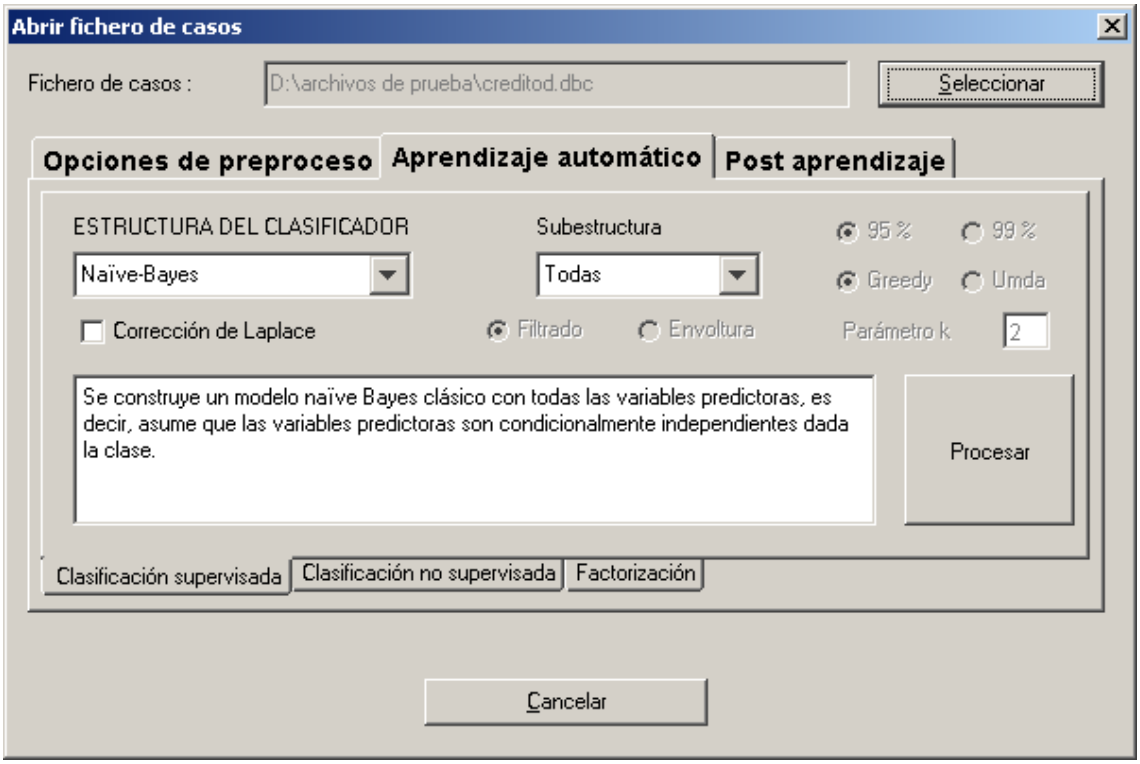

Figura 16: Menú apertura de fichero, opciones de aprendizaje automático

Aquí, continuando con el ejemplo anterior, se ha seleccionado "creditos.dbc" y como por defecto el sistema sugiere utilizar el clasificador Naïve Bayes no hay mas que ejecutar el botón "Procesar", una vez ejecutado se debe cerrar la ventana de incorporación de archivo y en la ventana principal aparecerá la gráfica de la red generada. A continuación, en la figura 17, se muestra la red obtenida:

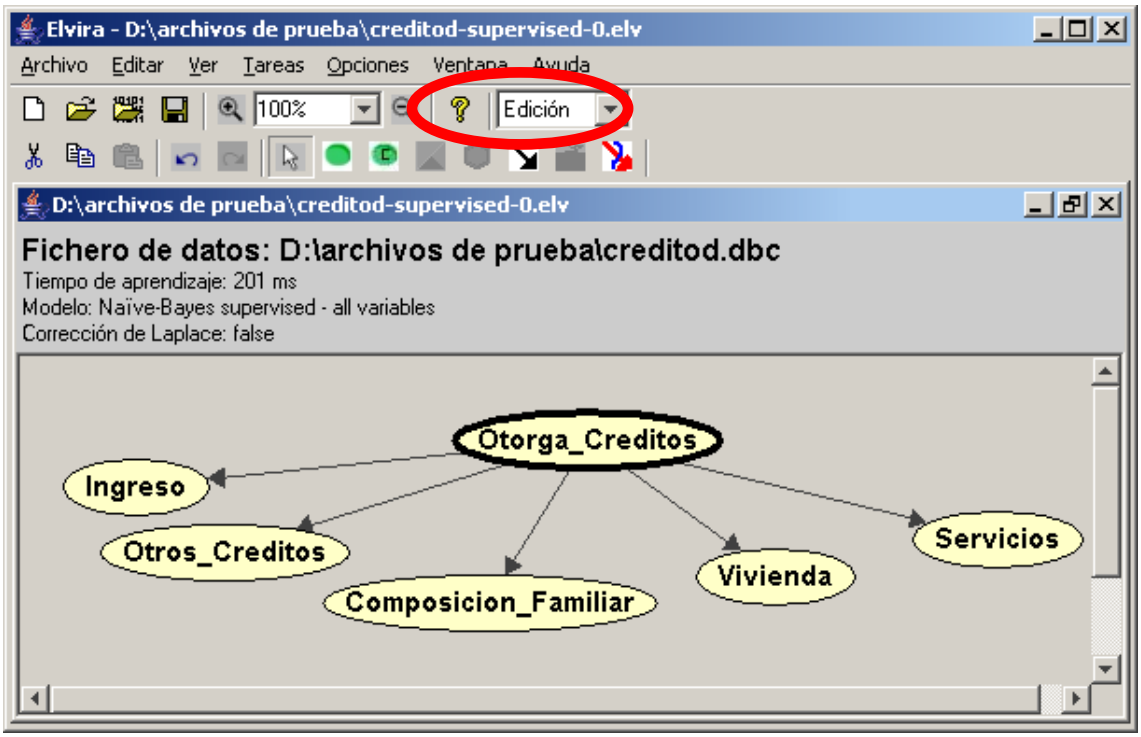

Figura 17: Red Bayesiana en modo edición

Para obtener mayores detalle sobre la red obtenida se debe cambiar, en la barra de tareas, la opción “Edición” por la opción “Inferencia”, a continuación, en la figura 18, se detalla esta última pantalla.

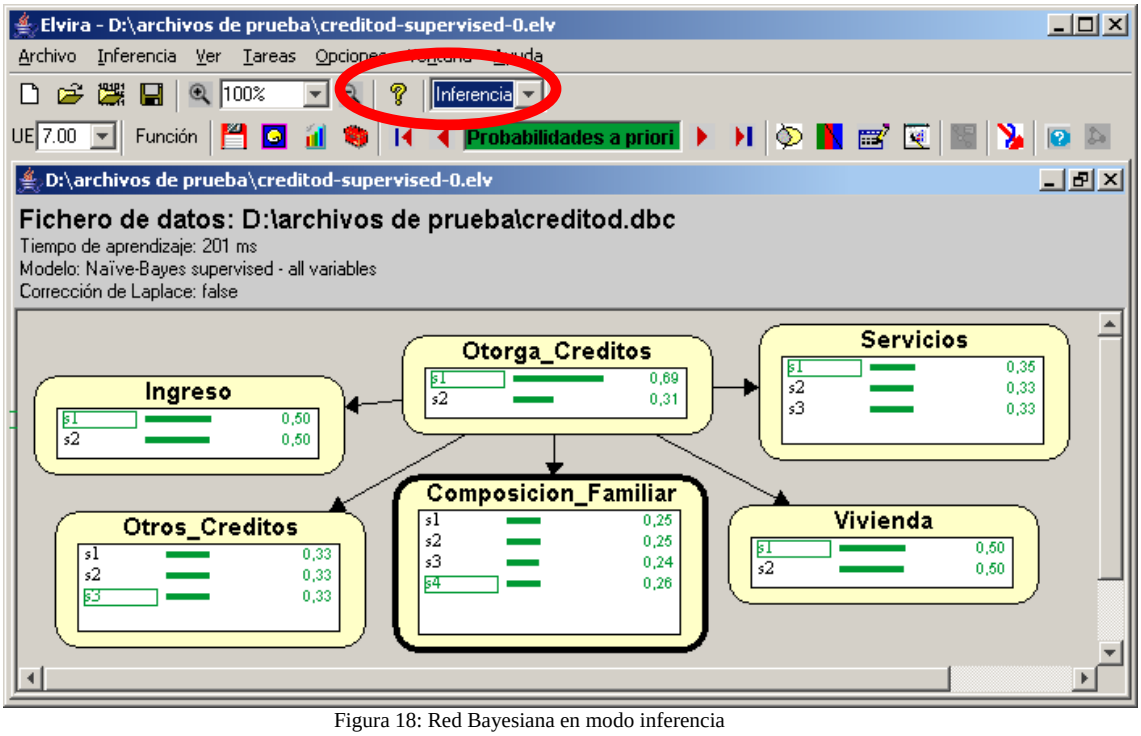

Figura 18: Red Bayesiana en modo inferencia

Aquí pueden verse las probabilidades de cada una de las distintas variable.
Una vez analizados todos los tipos de Clasificadores analizaremos con mayor detalle las opciones del menú de Inferencia.

A continuación se detallan los valores de probabilidad asociados a cada nodo en particular

Valor	Probabilidad
S1	0,6944
S2	0,3056

Tabla 6: Nodo: Otorga_Créditos

Nodo Padre	Valores	
Otorga crédito	S1	S2
Valor	Probabilidad	
S1	0,47	0.5682
S2	0,53	0.4318

Tabla 7: Nodo: Ingreso

Nodo Padre	Valores	
Otorga crédito	S1	S2
Valor	Probabilidad	
S1	0.45	0.0682
S2	0.41	0.1591
S3	0.14	0.7727

Tabla8: Nodo: Otros_Créditos

Nodo Padre	Valores	
Otorga crédito	S1	S2
Valor	Probabilidad	
S1	0.25	0.25
S2	0.3	0.1364
S3	0.27	0.1818
S4	0.18	0.4318

Tabla 9: Nodo: Composición_familiar

Nodo Padre	Valores	
Otorga crédito	S1	S2
Valor	Probabilidad	
S1	0.39	0.75
S2	0,61	0.25

Tabla 10: Nodo: Vivienda

Nodo Padre	Valores	
Otorga crédito	S1	S2
Valor	Probabilidad	
S1	0.36	0.3182
S2	0.33	0.3182
S3	0.31	0.3636

Tabla 11: Nodo: Servicios

2.1.2.2. Caso 2 obtención de una red mediante el clasificador TAN

Para generar una red mediante el Clasificador TAN se deberá cambiar la opción "Estructura del Clasificador", que por defecto sugiere Naïve Bayes, a TAN. Como se indica a continuación, en la figura 19:

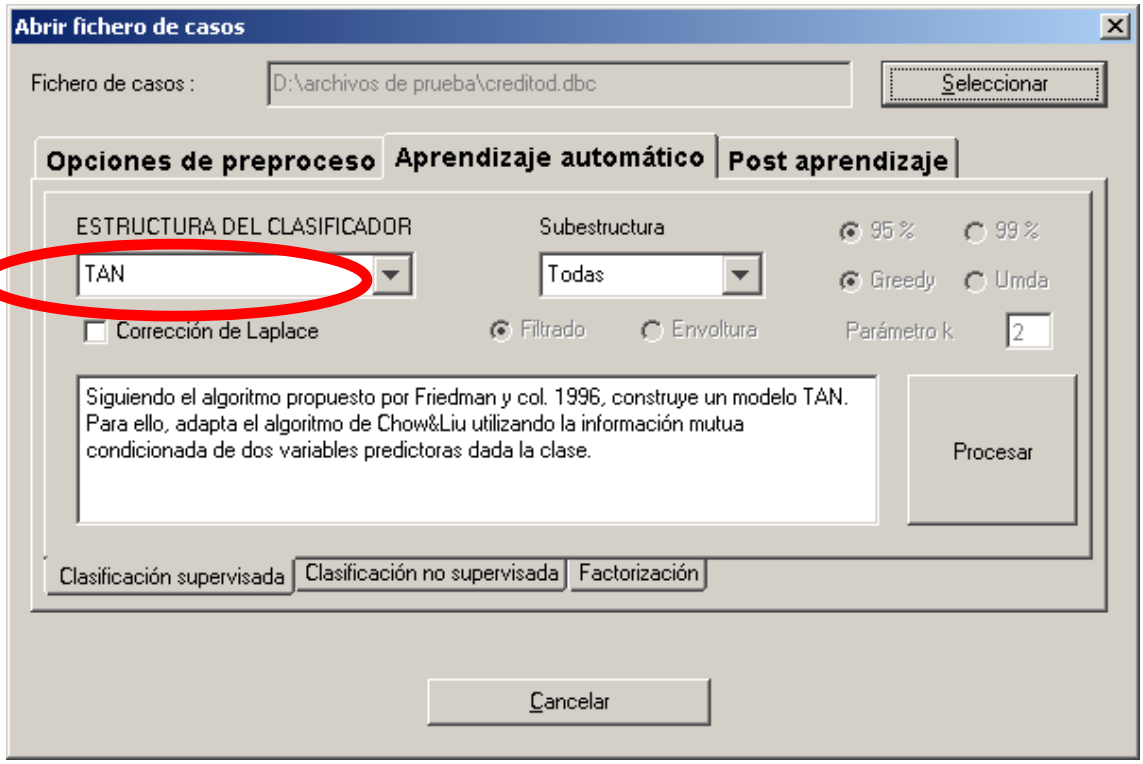

Figura 19: Menú apertura de fichero, opciones de aprendizaje automático

A continuación, en la figura 20, se detalla la red obtenida en modo Edición:

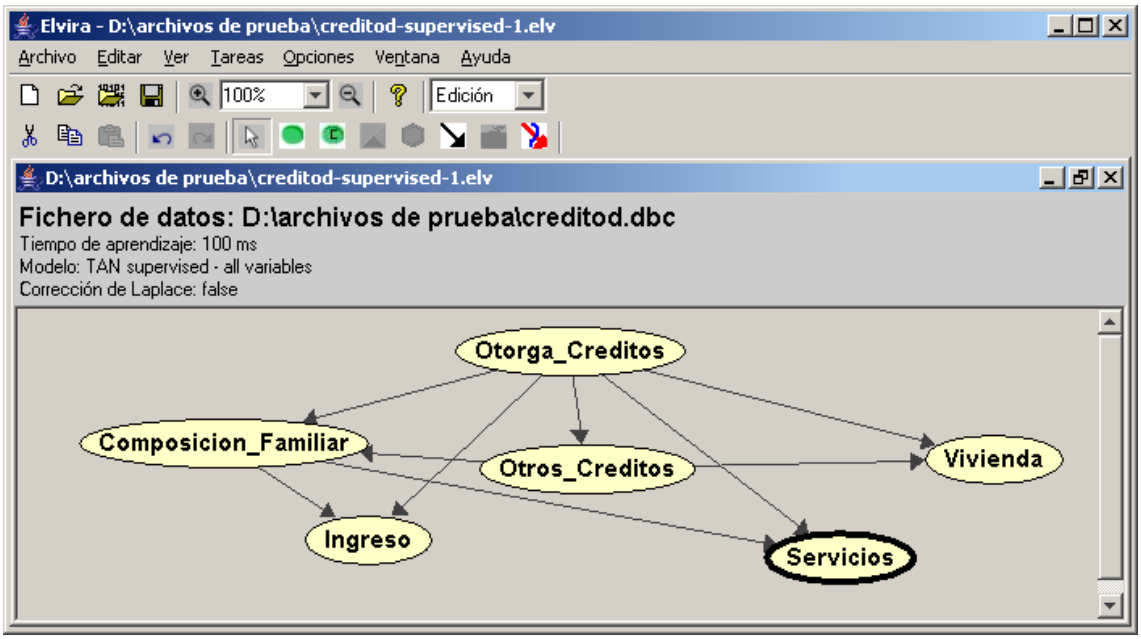

Figura 20: Red Bayesiana en modo edición

Por último, en la figura 21, se indican los valores de probabilidad obtenidos desde el modo Inferencia:

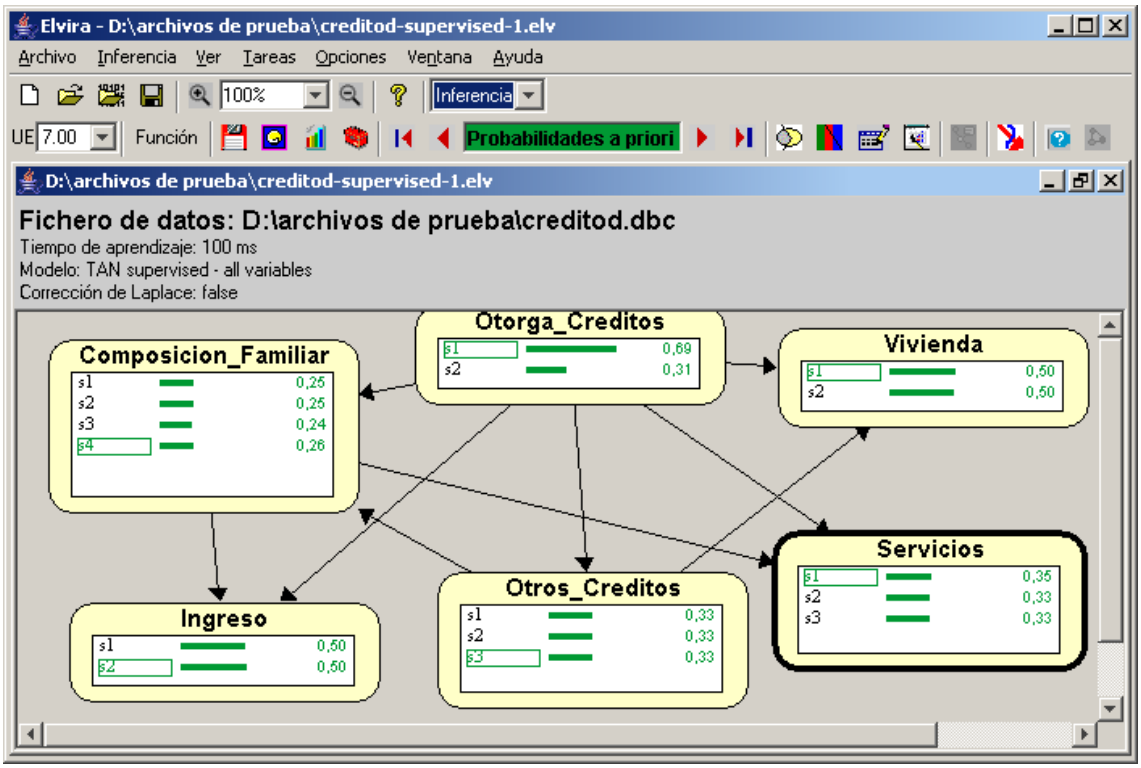

Figura 21: Red Bayesiana en modo inferencia

A continuación se detallan los valores de probabilidad asociados a cada nodo en particular

Valor	Probabilidad
S1	0,6944
S2	0,3056

Tabla 12: Nodo: Otorga_Créditos

Nodo Padre	Valores							
Compocisión_Familiar	S1	S1	S2	S2	S3	S3	S4	S4
Otorga_Crédito	S1	S2	S1	S2	S1	S2	S1	S2
Valor	Probabilidad							
S1	0.6	0.2727	0.5	0.5	0.4444	0.625	0.2778	0.7368
S2	0.4	0.7273	0.5	0.5	0.5556	0.375	0.7222	0.2632

Tabla 13: Nodo: Ingreso

Nodo Padre	Valores	
Otorga crédito	S1	S2
Valor	Probabilidad	
S1	0.45	0.0682
S2	0.41	0.1591
S3	0.14	0.7727

Tabla 14: Nodo: Otros_Créditos

Nodo Padre	Valores					
Otros_Creditos	S1	S1	S2	S2	S3	S3
Otorga_Crédito	S1	S2	S1	S2	S1	S2
Valor	Probabilidad					
S1	0.2667	0	0.2195	0.4286	0.2857	0.2353
S2	0.2667	0	0.2927	0	0.4286	0.1765
S3	0.2667	0	0.2927	0	0.2143	0.2353
S4	0.1999	1	0.1951	0.5714	0.0714	0.3529

Tabla 15: Nodo: Composición_familiar

Nodo Padre	Valores					
Otorga_Crédito	S1	S1	S2	S2	S3	S3
Otros_Creditos	S1	S2	S1	S2	S1	S2
Valor	Probabilidad					
S1	0.4667	1	0.439	0.8571	0	0.7059
S2	0.5333	0	0.561	0.1429	1	0.2941

Tabla 16: Nodo: Vivienda

Nodo Padre	Valores							
Compocisión_Familiar	S1	S1	S2	S2	S3	S3	S4	S4
Otorga_Crédito	S1	S2	S1	S2	S1	S2	S1	S2
Valor	Probabilidad							
S1	0.4	0.2727	0.3333	0.3333	0.3333	0.5	0.3889	0.2632
S2	0.28	0.3636	0.3333	0.3333	0.3704	0.25	0.3333	0.3158
S3	0.32	0.3637	0.3334	0.3334	0.2963	0.25	0.2778	0.421

Tabla 17: Nodo: Servicios:

2.1.2.3. Caso 3 obtención de una red mediante el clasificador KDB

Para generar una red mediante el Clasificador KDB se deberá cambiar la opción “Estructura del Clasificador”, que por defecto sugiera Naïve Bayes, a KDB. Como se indica a continuación, en la figura 22:

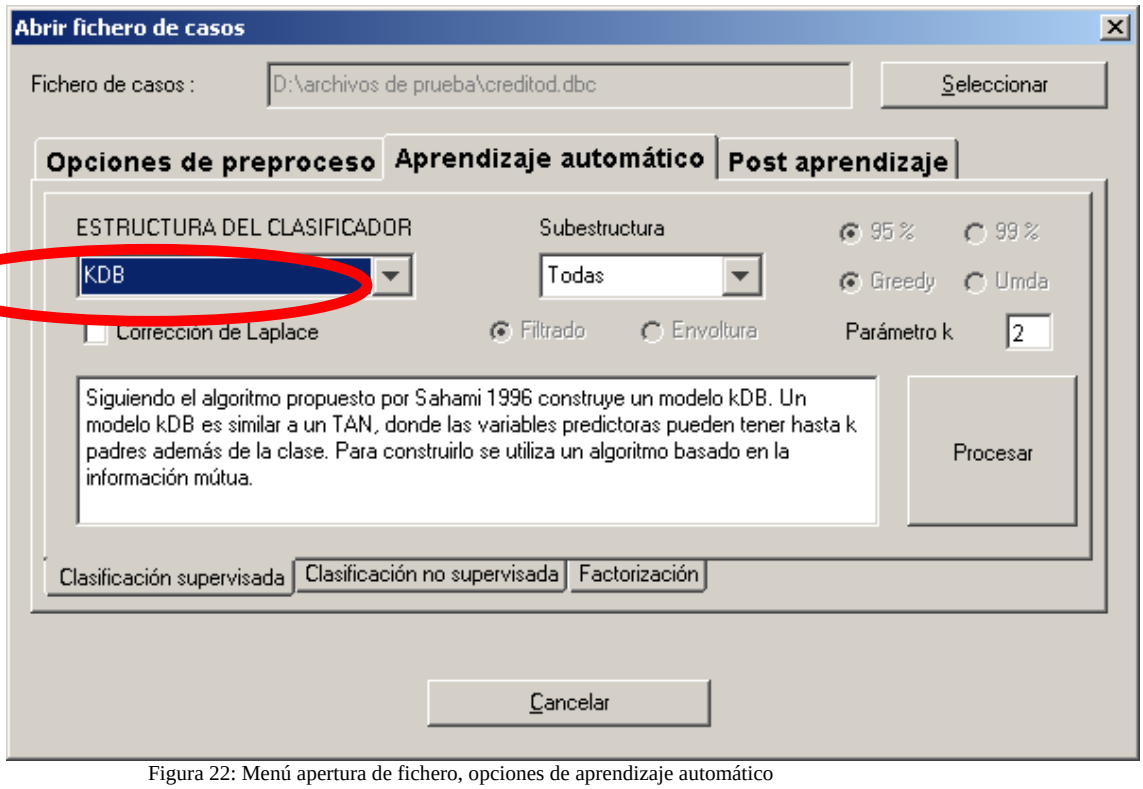

Figura 22: Menú apertura de fichero, opciones de aprendizaje automático

A continuación, en la figura 23, se detalla la red obtenida en modo Edición:

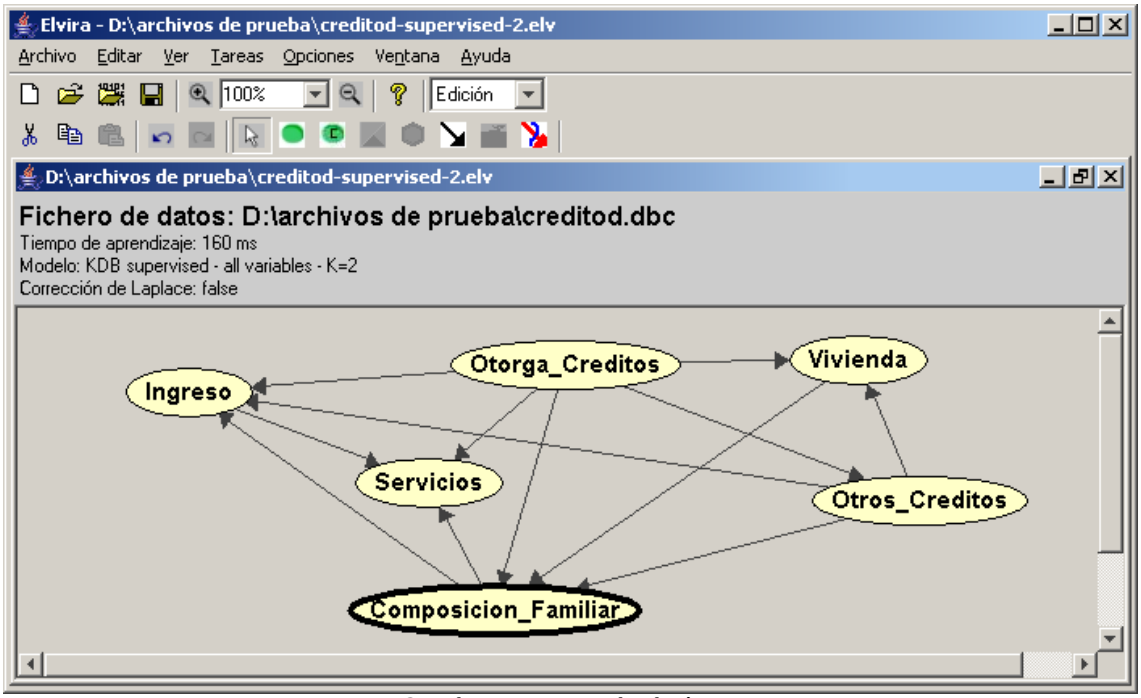

Figura 23: Red Bayesiana en modo edición

Por último, en la figura 24, se indican los valores de probabilidad obtenidos desde el modo Inferencia:

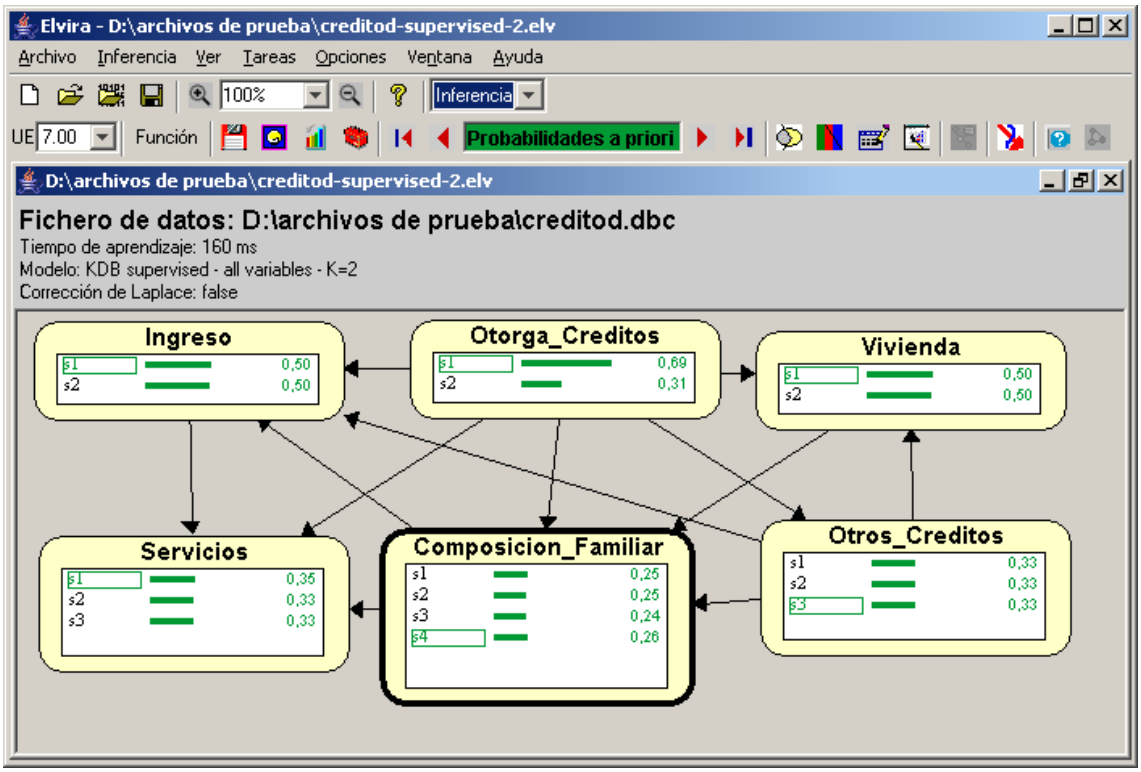

Figura 24: Red Bayesiana en modo inferencia

A continuación se detallan los valores de probabilidad asociados a cada nodo en particular

Valor	Probabilidad
S1	0,6944
S2	0,3056

Tabla 18: Nodo: Otorga_Créditos

Nodo Padre	Valores											
Otorga_Crédito	S1	S1	S1	S1	S1	S1	S2	S2	S2	S2	S2	S2
Otros_Créditos	S1	S1	S2	S2	S3	S3	S1	S1	S2	S2	S3	S3
Vivienda	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
Valor	Probabilidad											
S1	0.2857	0.25	0.1667	0.2609	0	0.2857	0	0	0.5	0	0.25	0.2
S2	0.2857	0.25	0.3333	0.2609	0	0.4286	0	0	0	0	0.25	0
S3	0.2857	0.25	0.3333	0.2609	0	0.2143	0	0	0	0	0.25	0.2
S4	0.1429	0.25	0.1667	0.2173	1	0.0714	1	1	0.5	0	0.25	0.6

Tabla 19: Nodo: Composición_familiar

Nodo Padre	Valores	
Otorga crédito	S1	S2
Valor	Probabilidad	
S1	0.45	0.0682
S2	0.41	0.1591
S3	0.14	0.7727

Tabla 20: Nodo: Otros_Créditos

Nodo Padre	Valores																							
Otorga_ Crédito	S1	S1	S1	S1	S1	S1	S1	S1	S1	S1	S1	S1	S2	S2	S2	S2	S2	S2	S2	S2	S2	S2	S2	S2
Composi-ción_ Familiar	S1	S1	S1	S2	S2	S2	S3	S3	S3	S4	S4	S4	S1	S1	S1	S2	S2	S2	S3	S3	S3	S4	S4	S4
Otros_ Creditos	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3
Valor	Probabilidad																							
S1	0.5	0.6667	0,5	0.5	0.5	0,5	0.5	0.5	0	0.3333	0.25	0	0	0	0.375	0	0	0.5	0	0	0.625	1	1	0.5833
S2	0.5	0.3333	0,5	0.5	0.5	0,5	0.5	0.5	1	0.6667	0.75	1	1	1	0.625	1	1	0.5	1	1	0.375	0	0	0.4167

Tabla 21: Nodo: Ingreso

Nodo Padre	Valores					
Otros_Creditos	S1	S1	S1	S2	S2	S2
Otorga_Crédito	S1	S2	S3	S1	S2	S3
Valor	Probabilidad					
S1	0.4667	0.439	0	1	0.8571	0.7059
S2	0.5333	0.561	1	0	0.1429	0.2941

Tabla 22: Nodo: Vivienda

Nodo Padre	Valores															
Otorga_ Crédito	S1	S1	S1	S1	S1	S1	S1	S1	S2	S2	S2	S2	S2	S2	S2	S2
Composición_ Familiar	S1	S1	S2	S2	S3	S3	S4	S4	S1	S1	S2	S2	S3	S3	S4	S4
Ingreso	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
Valor	Probabilidad															
S1	0.4	0.4	0.3333	0.3333	0.3333	0.3333	0.4	0.3846	0.3333	0.25	0.3333	0.3333	0.6	0.3333	0.2857	0.2
Sz2	0.2667	0.3	0.3333	0.3333	0.4167	0.3333	0.4	0.3077	0.3333	0.375	0.3333	0.3333	0.2	0.3333	0.2857	0.4
S3	0.3333	0.3	0.3334	0.3334	0.25	0.3334	0.2	0.3077	0.3334	0.375	0.3334	0.3334	0.2	0.3334	0.4286	0.4

Tabla 23: Nodo: Servicios:

2.1.3. Combinación de las redes bayesianas y los Algoritmos de inducción

En este punto se convinarán los clasificadores Bayesianos vistos en la sección anterior (naïve Bayes, TAN y KDB) con los algoritmos de inducción de árboles de decisión (ID3 y C4.5). Esto puede realizarse gracias a una opción de preprocesamiento de archivo que brinda el software Elvira.

2.1.3.1. Caso 4 obtención de una red Naïve Bayes con ID3

A continuación, en la figura 25, se muestra como cargar la parametría de preproceso. En este preprocesamiento se puede aplicar al archivo de casos algún algoritmo del tipo “Inteligente” para refinar la información contenida en el mismo y dar de esta forma una mayor precisión a la red Bayesiana que se obtenga con el uso de los clasificadores.

La figura 19 indica como seleccionar en la solapa de preproceso el algoritmo ID3.

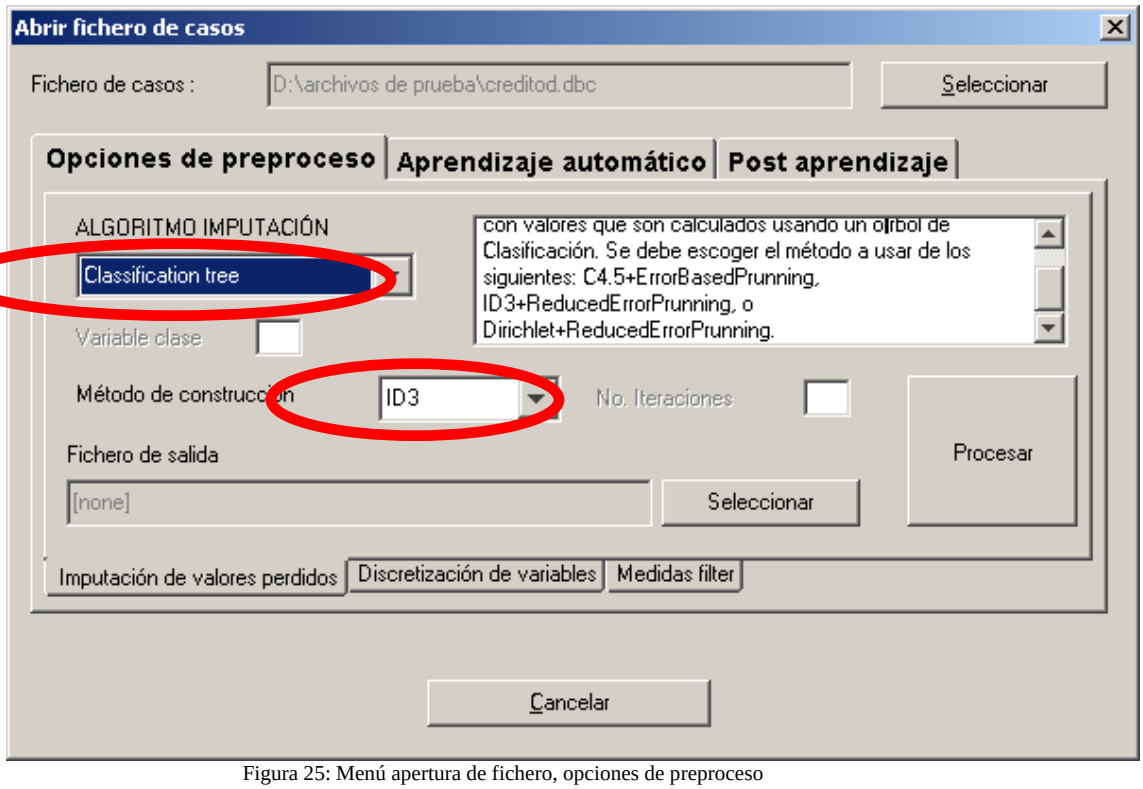

Figura 25: Menú apertura de fichero, opciones de preproceso

La figura 26 muestra la red obtenida mediante el clasificador Naïve Bayes luego de haber realizado el preprocesamiento. En esta ocasión, no se indica el camino para seleccionar este clasificador mencionado, por que es mismo que se indico en el caso 1.

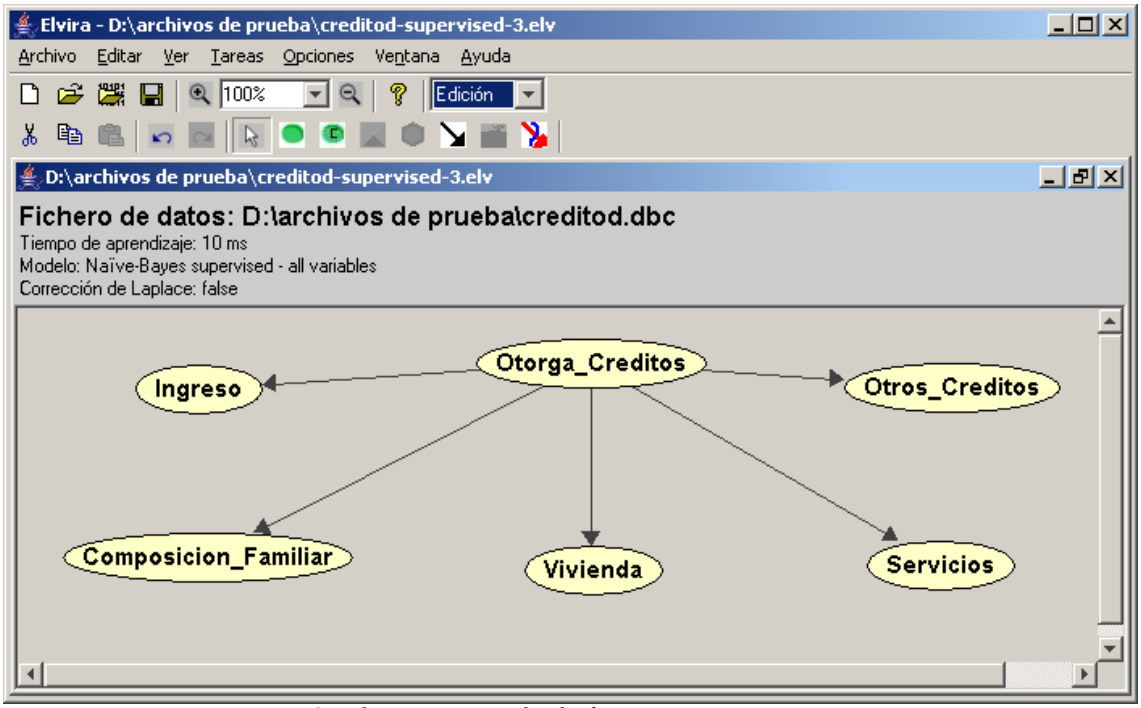

Figura 26: Red Bayesiana en modo edición

Por último, en la figura 27, se indican los valores de probabilidad obtenidos desde el modo Inferencia:

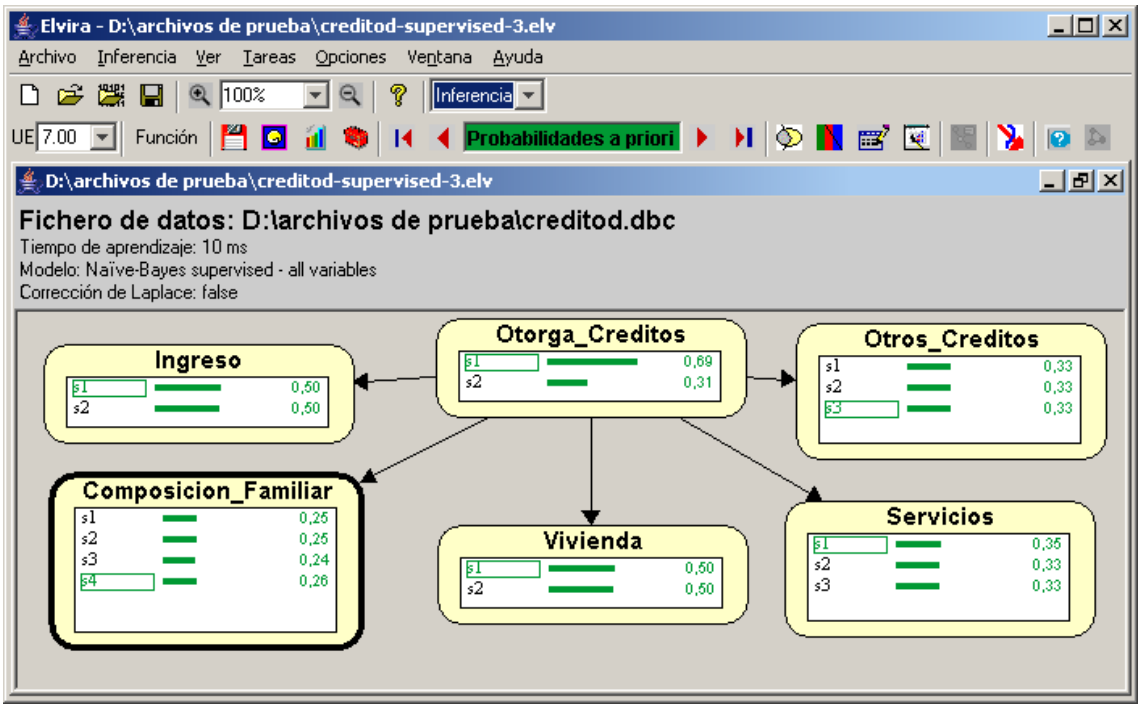

Figura 27: Red Bayesiana en modo inferencia

A continuación se detallan los valores de probabilidad asociados a cada nodo en particular

Valor	Probabilidad
S1	0,6944
S2	0,3056

Tabla 24: Nodo: Otorga_Créditos

Nodo Padre	Valores	
Otorga crédito	S1	S2
Valor	Probabilidad	
S1	0,47	0.5682
S2	0,53	0.4318

Tabla 25: Nodo: Ingreso

Nodo Padre	Valores	
Otorga crédito	S1	S2
Valor	Probabilidad	
S1	0.45	0.0682
S2	0.41	0.1591
S3	0.14	0.7727

Tabla 26: Nodo: Otros_Créditos

Nodo Padre	Valores	
Otorga crédito	S1	S2
Valor	Probabilidad	
S1	0.25	0.25
S2	0.3	0.1364
S3	0.27	0.1818
S4	0.18	0.4318

Tabla 27: Nodo: Composición_familiar

Nodo Padre	Valores	
Otorga crédito	S1	S2
Valor	Probabilidad	
S1	0.39	0.75
S2	0,61	0.25

Tabla 28 Nodo: Vivienda

Nodo Padre	Valores	
Otorga crédito	S1	S2
Valor	Probabilidad	
S1	0.36	0.3182
S2	0.33	0.3182
S3	0.31	0.3636

Tabla 29: Nodo: Servicios

2.1.3.2. Caso 5 obtención de una red TAN con ID3

En este punto solo se indica en la figura 28 la red obtenida aplicando el algoritmo ID3 en la etapa de preprocesamiento y el clasificador TAN, sin dar mayores detalles de cómo se llega a ellos, por que la forma de hacerlo se explico en el caso 4 (como seleccionar el algoritmo ID3) y el caso 2 (como seleccionar el clasificador TAN).

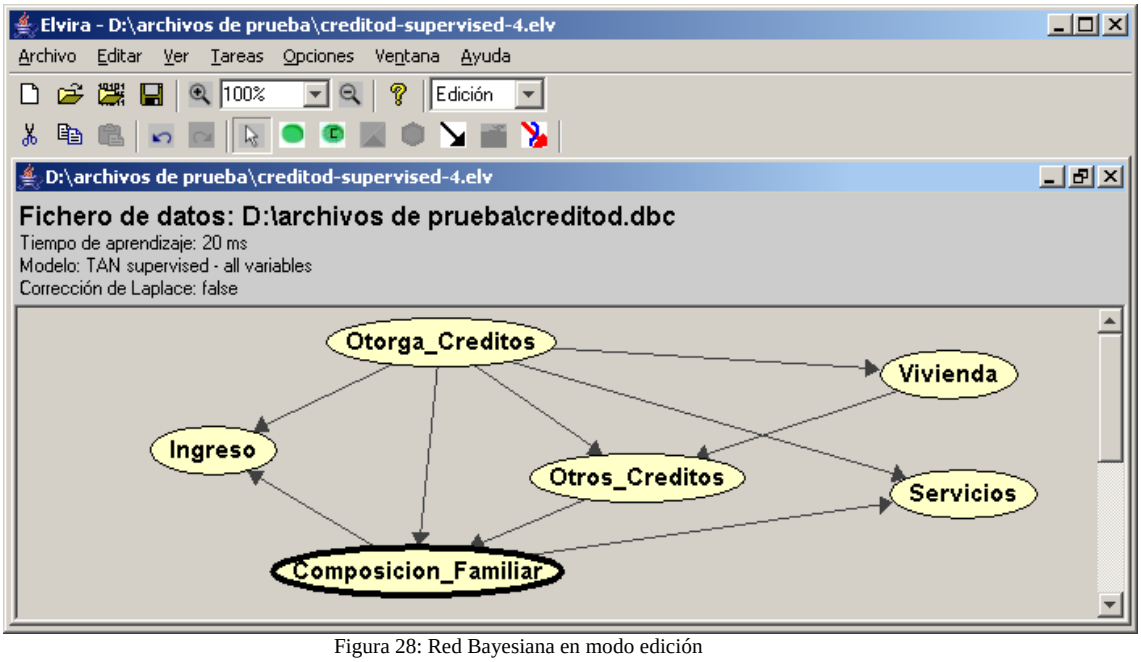

Figura 28: Red Bayesiana en modo edición

Por último, en la figura 29, se indican los valores de probabilidad obtenidos desde el modo Inferencia:

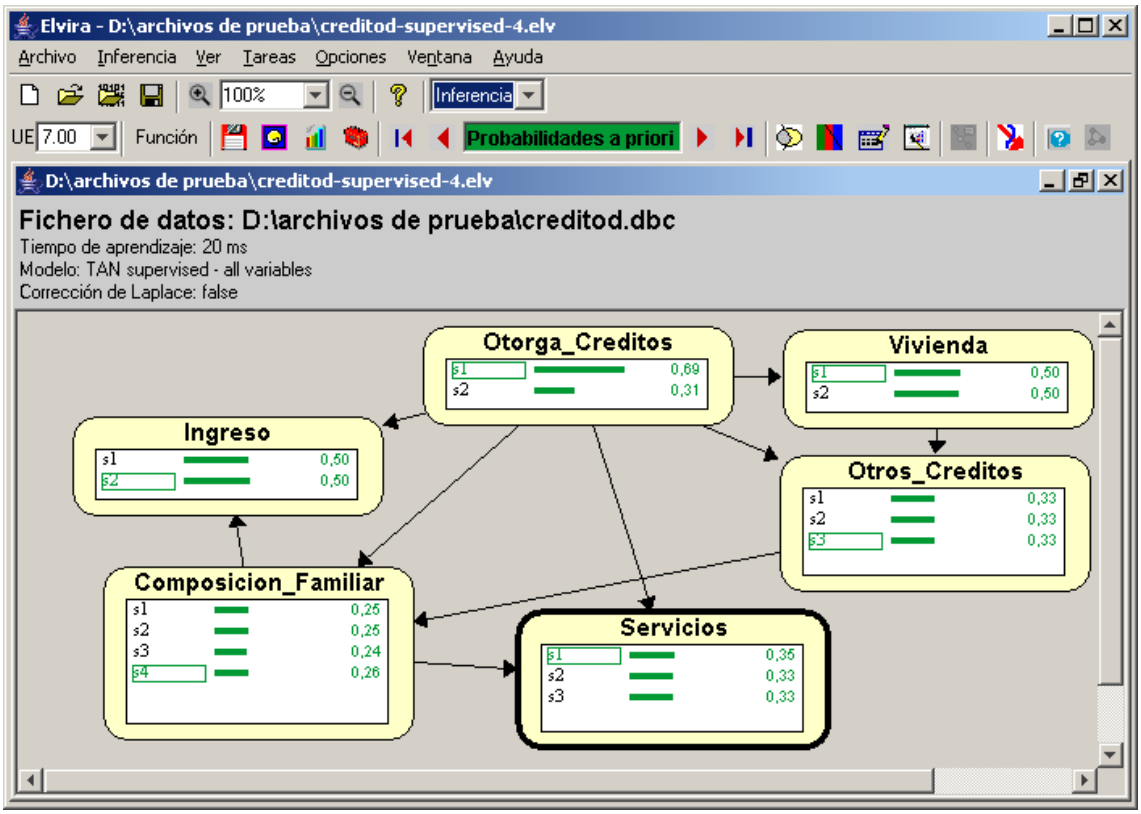

Figura 29: Red Bayesiana en modo inferencia

A continuación se detallan los valores de probabilidad asociados a cada nodo en particular

Valor	Probabilidad
S1	0,6944
S2	0,3056

Tabla 30: Nodo: Otorga_Créditos

Nodo Padre	Valores							
Compocisión_Familiar	S1	S1	S2	S2	S3	S3	S4	S4
Otorga_Crédito	S1	S2	S1	S2	S1	S2	S1	S2
Valor	Probabilidad							
S1	0.6	0.2727	0.5	0.5	0.4444	0.625	0.2778	0.7368
S2	0.4	0.7273	0.5	0.5	0.5556	0.375	0.7222	0.2632

Tabla 31: Nodo: Ingreso

Nodo Padre	Valores			
Vivienda	S1	S1	S2	S2
Otorga crédito	S1	S2	S1	S2
Valor	Probabilidad			
S1	0.5385	0.0909	0.3934	0
S2	0.4615	0.1818	0.377	0.0909
S3	0	0.7273	0.2296	0.9091

Tabla 32: Nodo: Otros_Créditos

Nodo Padre	Valores					
Otros_Creditos	S1	S1	S2	S2	S3	S3
Otorga_Crédito	S1	S2	S1	S2	S1	S2
Valor	Probabilidad					
S1	0.2667	0	0.2195	0.4286	0.2857	0.2353
S2	0.2667	0	0.2927	0	0.4286	0.1765
S3	0.2667	0	0.2927	0	0.2143	0.2353
S4	0.1999	1	0.1951	0.5714	0.0714	0.3529

Tabla 33: Nodo: Composición_familiar

Nodo Padre	Valores	
Otorga_Crédito	S1	S2
Valor	Probabilidad	
S1	0.39	0.75
S2	0.61	0.25

Tabla 34: Nodo: Vivienda

Nodo Padre	Valores							
Compocisión_Familiar	S1	S1	S2	S2	S3	S3	S4	S4
Otorga_Crédito	S1	S2	S1	S2	S1	S2	S1	S2
Valor	Probabilidad							
S1	0.4	0.2727	0.3333	0.3333	0.3333	0.5	0.3889	0.2632
S2	0.28	0.3636	0.3333	0.3333	0.3704	0.25	0.3333	0.3158
S3	0.32	0.3637	0.3334	0.3334	0.2963	0.25	0.2778	0.421

Tabla 35: Nodo: Servicios:

2.1.3.3. Caso 6 obtención de una red KDB con ID3

En este punto solo se indica en la figura 30 la red obtenida aplicando el algoritmo ID3 en la etapa de preprocesamiento y el clasificador KDB, sin dar mayores detalles de cómo se llega a ellos, por que la forma de hacerlo se explico en el caso 4 (como seleccionar el algoritmo ID3) y el caso 3 (como seleccionar el clasificador KDB).

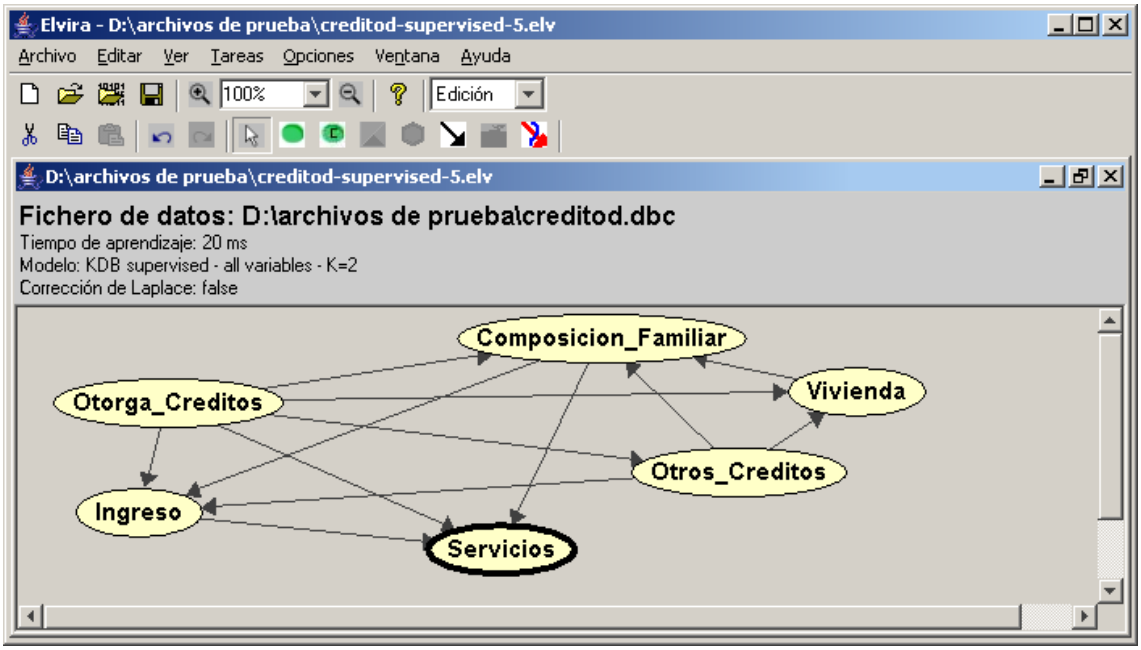

Figura 30: Red Bayesiana en modo edición

Por último, en la figura 31, se indican los valores de probabilidad obtenidos desde el modo Inferencia:

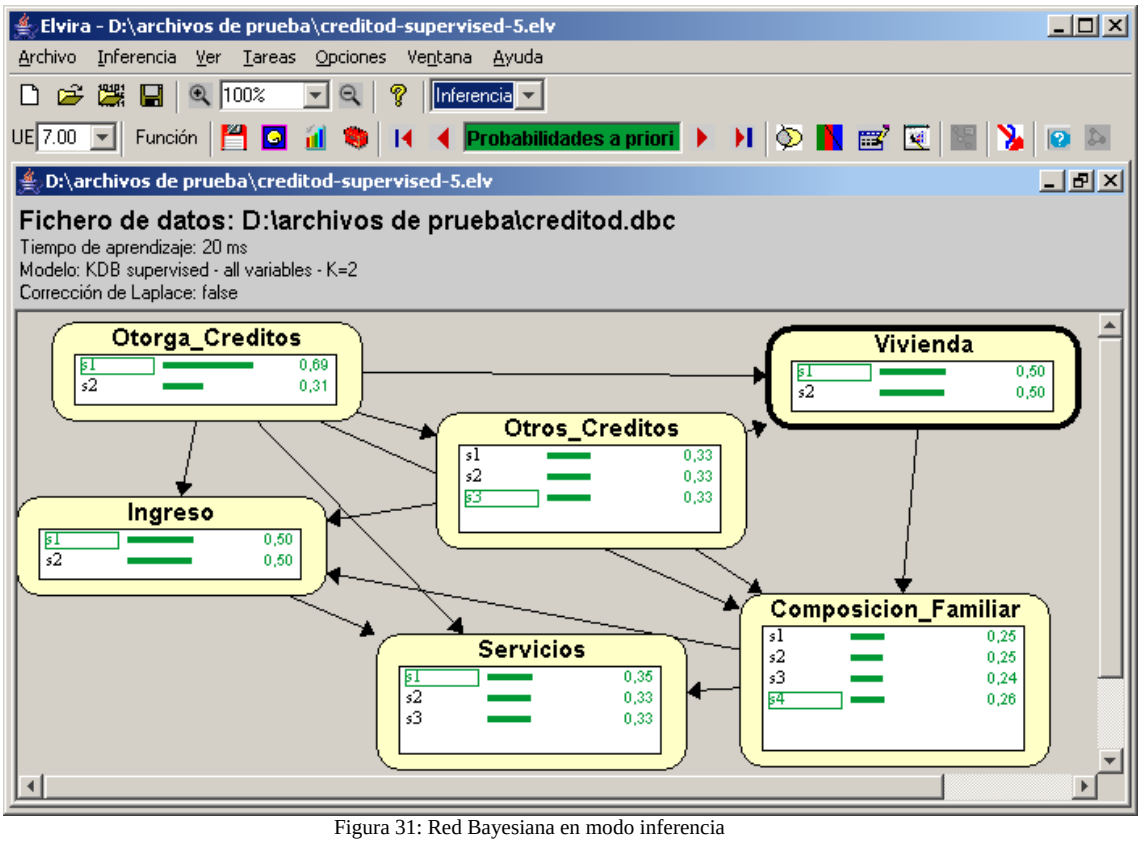

Figura 31: Red Bayesiana en modo inferencia

A continuación se detallan los valores de probabilidad asociados a cada nodo en particular

Valor	Probabilidad
S1	0,6944
S2	0,3056

Tabla 36: Nodo: Otorga_Créditos

Nodo Padre	Valores											
Otorga_Crédito	S1	S1	S1	S1	S1	S1	S2	S2	S2	S2	S2	S2
Otros_Créditos	S1	S1	S2	S2	S3	S3	S1	S1	S2	S2	S3	S3
Vivienda	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
Valor	Probabilidad											
S1	0.2857	0.25	0.1667	0.2609	0	0.2857	0	0	0.5	0	0.25	0.2
S2	0.2857	0.25	0.3333	0.2609	0	0.4286	0	0	0	0	0.25	0
S3	0.2857	0.25	0.3333	0.2609	0	0.2143	0	0	0	0	0.25	0.2
S4	0.1429	0.25	0.1667	0.2173	1	0.0714	1	1	0.5	0	0.25	0.6

Tabla 37: Nodo: Composición_familiar

Nodo Padre	Valores	
Otorga crédito	S1	S2
Valor	Probabilidad	
S1	0.45	0.0682
S2	0.41	0.1591
S3	0.14	0.7727

Tabla 38: Nodo: Otros_Créditos

Nodo Padre	Valores																							
Otorga_ Crédito	S1	S1	S1	S1	S1	S1	S1	S1	S1	S1	S1	S1	S2	S2	S2	S2	S2	S2	S2	S2	S2	S2	S2	S2
Composi-ción_ Familiar	S1	S1	S1	S2	S2	S2	S3	S3	S3	S4	S4	S4	S1	S1	S1	S2	S2	S2	S3	S3	S3	S4	S4	S4
Otros_ Creditos	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3
Valor	Probabilidad																							
S1	0.5	0.6667	0,5	0.5	0.5	0,5	0.5	0.5	0	0.3333	0.25	0	0	0	0.375	0	0	0.5	0	0	0.625	1	1	0.5833
S2	0.5	0.3333	0,5	0.5	0.5	0,5	0.5	0.5	1	0.6667	0.75	1	1	1	0.625	1	1	0.5	1	1	0.375	0	0	0.4167

Tabla 39: Nodo: Ingreso

Nodo Padre	Valores					
Otros_Creditos	S1	S1	S1	S2	S2	S2
Otorga_Crédito	S1	S2	S3	S1	S2	S3
Valor	Probabilidad					
S1	0.4667	0.439	0	1	0.8571	0.7059
S2	0.5333	0.561	1	0	0.1429	0.2941

Tabla 40: Nodo: Vivienda

Nodo Padre	Valores															
Otorga_ Crédito	S1	S1	S1	S1	S1	S1	S1	S1	S2	S2	S2	S2	S2	S2	S2	S2
Composición_ Familiar	S1	S1	S2	S2	S3	S3	S4	S4	S1	S1	S2	S2	S3	S3	S4	S4
Ingreso	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
Valor	Probabilidad															
S1	0.4	0.4	0.3333	0.3333	0.3333	0.3333	0.4	0.3846	0.3333	0.25	0.3333	0.3333	0.6	0.3333	0.2857	0.2
Sz2	0.2667	0.3	0.3333	0.3333	0.4167	0.3333	0.4	0.3077	0.3333	0.375	0.3333	0.3333	0.2	0.3333	0.2857	0.4
S3	0.3333	0.3	0.3334	0.3334	0.25	0.3334	0.2	0.3077	0.3334	0.375	0.3334	0.3334	0.2	0.3334	0.4286	0.4

Tabla 41: Nodo: Servicios:

2.1.3.4. Caso 7 obtención de una red Naïve Bayes con C4.5

A continuación, en la figura 32, se muestra como cargar la parametría de preproceso. En este preprocesamiento se puede aplicar al archivo de casos algún algoritmo del tipo “Inteligente” para refinar la información contenida en el mismo y dar de esta forma una mayor precisión a la red Bayesiana que se obtenga con el uso de los clasificadores.

La figura 32 indica como seleccionar para la etapa de preproceso el algoritmo C4.5.

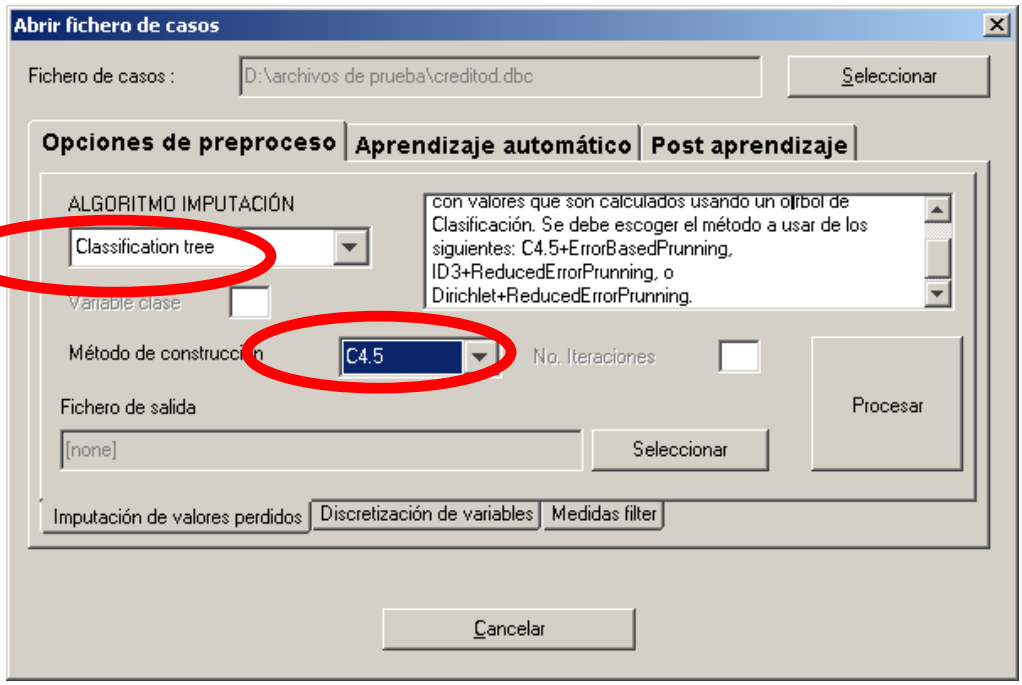

Figura 32: Menú apertura de fichero, opciones de preproceso

La figura 33 muestra la red obtenida mediante el clasificador Naïve Bayes luego a haber realizado el preprocesamiento. En esta ocasión, no se indica el camino para seleccionar este clasificador mencionado, por que es mismo que se indico en el caso 1.

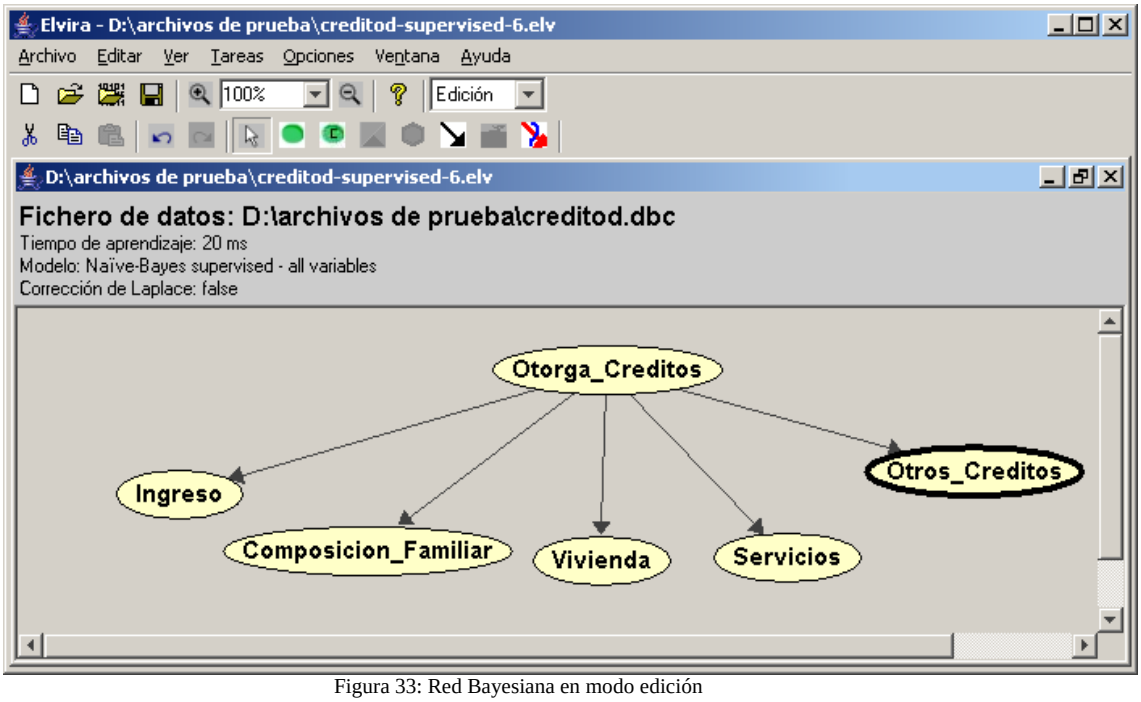

Figura 33: Red Bayesiana en modo edición

Por último, en la figura 34, se indican los valores de probabilidad obtenidos desde el modo Inferencia:

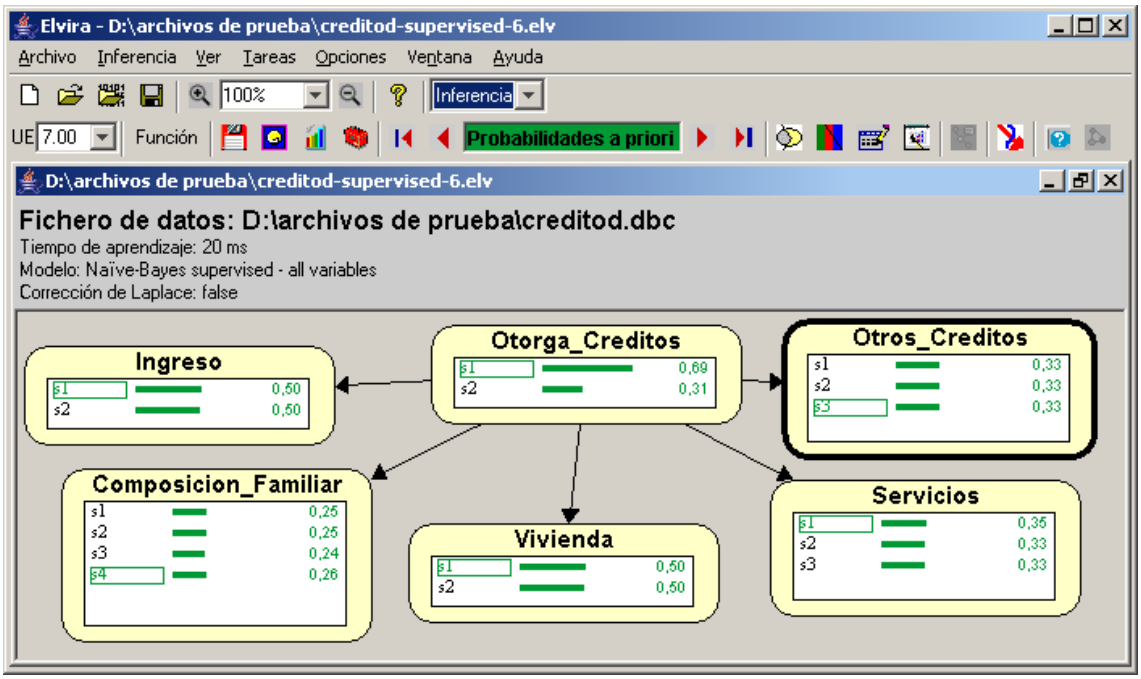

Figura 34: Red Bayesiana en modo inferencia

A continuación se detallan los valores de probabilidad asociados a cada nodo en particular

Valor	Probabilidad
S1	0,6944
S2	0,3056

Tabla 42: Nodo: Otorga_Créditos

Nodo Padre	Valores	
Otorga crédito	S1	S2
Valor	Probabilidad	
S1	0,47	0.5682
S2	0,53	0.4318

Tabla 43: Nodo: Ingreso

Nodo Padre	Valores	
Otorga crédito	S1	S2
Valor	Probabilidad	
S1	0.45	0.0682
S2	0.41	0.1591
S3	0.14	0.7727

Tabla 44: Nodo: Otros_Créditos

Nodo Padre	Valores	
Otorga crédito	S1	S2
Valor	Probabilidad	
S1	0.25	0.25
S2	0.3	0.1364
S3	0.27	0.1818
S4	0.18	0.4318

Tabla 45: Nodo: Composición_familiar

Nodo Padre	Valores	
Otorga crédito	S1	S2
Valor	Probabilidad	
S1	0.39	0.75
S2	0,61	0.25

Tabla 46: Nodo: Vivienda

Nodo Padre	Valores	
Otorga crédito	S1	S2
Valor	Probabilidad	
S1	0.36	0.3182
S2	0.33	0.3182
S3	0.31	0.3636

Tabla 47: Nodo: Servicios

2.1.3.5. Caso 8 obtención de una red TAN con C4.5

En este punto solo se indica en la figura 35 la red obtenida aplicando el algoritmo C4.5 en la etapa de preprocesamiento y el clasificador TAN, sin dar mayores detalles de cómo se llega a ellos, por que la forma de hacerlo se explico en el caso 7 (como seleccionar el algoritmo C4.5) y el caso 2 (como seleccionar el clasificador TAN).

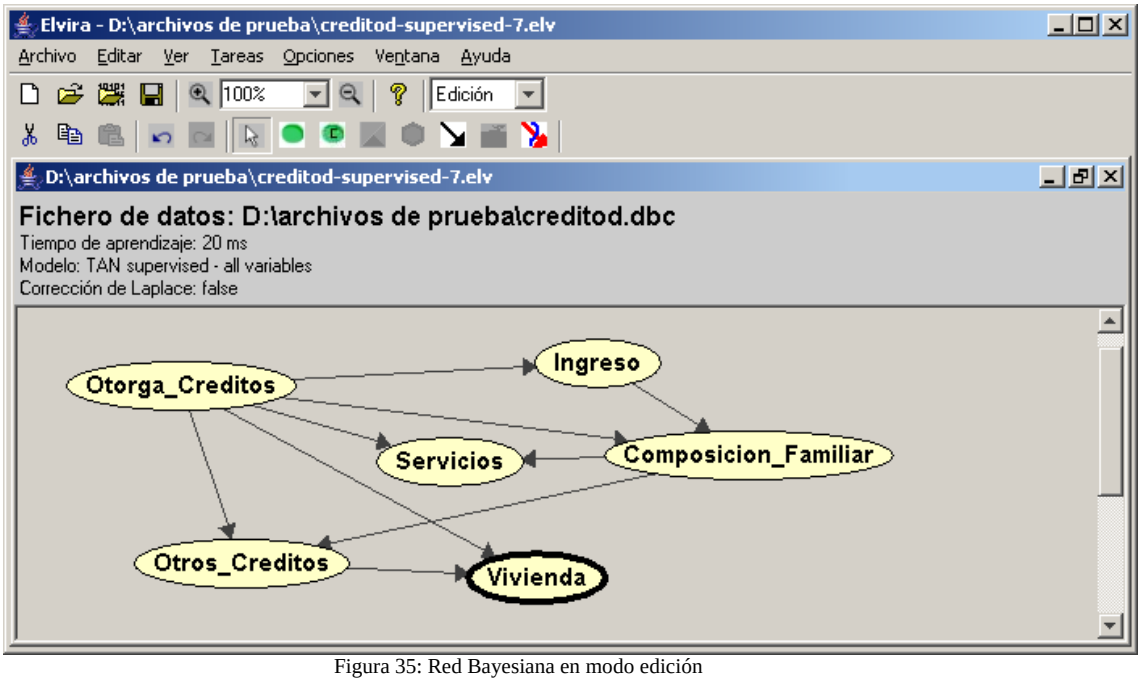

Figura 35: Red Bayesiana en modo edición

Por último, en la figura 36, se indican los valores de probabilidad obtenidos desde el modo Inferencia:

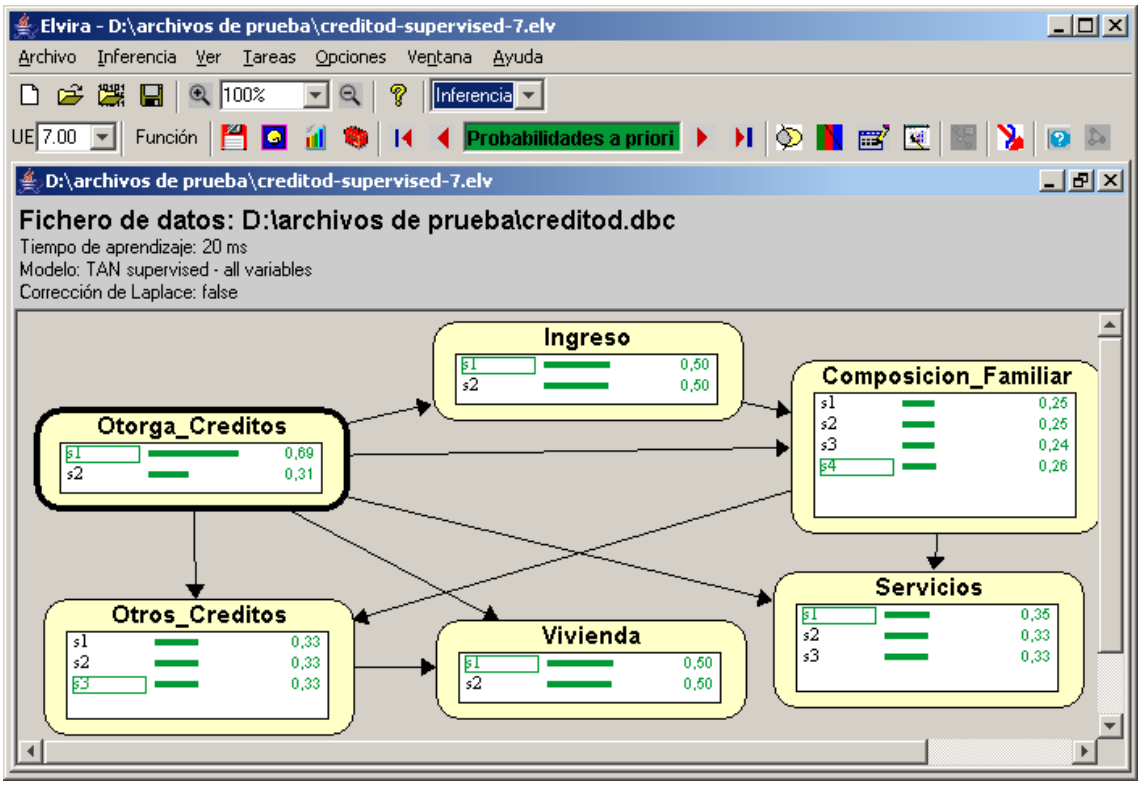

Figura 36: Red Bayesiana en modo inferencia

A continuación se detallan los valores de probabilidad asociados a cada nodo en particular

Valor	Probabilidad
S1	0,6944
S2	0,3056

Tabla 48: Nodo: Otorga_Créditos

Nodo Padre	Valores	
Otorga_Crédito	S1	S2
Valor	Probabilidad	
S1	0.47	0.5682
S2	0.53	0.4318

Tabla 49:Nodo: Ingreso

Nodo Padre	Valores							
Compocisión_Familiar	S1	S1	S2	S2	S3	S3	S4	S4
Otorga crédito	S1	S2	S1	S2	S1	S2	S1	S2
Valor	Probabilidad							
S1	0.48	0	0.4	0	0.4444	0	0.5	0.1579
S2	0.36	0.2727	0.4	0	0.4444	0	0.4444	0.2105
S3	0.16	0.7273	0.2	1	0.1112	1	0.0556	0.6316

Tabla 50: Nodo: Otros_Créditos

Nodo Padre	Valores			
Ingreso	S1	S1	S2	S2
Otorga_Crédito	S1	S2	S1	S2
Valor	Probabilidad			
S1	0.3191	0.12	0.1887	0.4211
S2	0.3191	0.12	0.283	0.1579
S3	0.2553	0.2	0.283	0.1579
S4	0.1065	0.56	0.2453	0.2631

Tabla 51: Nodo: Composición_familiar

Nodo Padre	Valores					
Otros_Creditos	S1	S1	S2	S2	S3	S3
Otorga_Crédito	S1	S2	S1	S2	S1	S2
Valor	Probabilidad					
S1	0.4667	1	0.439	0.8571	0	0.7059
S2	0.5333	0	0.561	0.1429	1	0.2941

Tabla 52: Nodo: Vivienda

Nodo Padre	Valores							
Compocisión_Familiar	S1	S1	S2	S2	S3	S3	S4	S4
Otorga_Crédito	S1	S2	S1	S2	S1	S2	S1	S2
Valor	Probabilidad							
S1	0.4	0.2727	0.3333	0.3333	0.3333	0.5	0.3889	0.2632
S2	0.28	0.3636	0.3333	0.3333	0.3704	0.25	0.3333	0.3158
S3	0.32	0.3637	0.3334	0.3334	0.2963	0.25	0.2778	0.421

Tabla 53: Nodo: Servicios:

2.1.3.6. Caso 9 obtención de una red KDB con C4.5

En este punto solo se indica en la figura 37 la red obtenida aplicando el algoritmo C4.5 en la etapa de preprocesamiento y el clasificador KDB, sin dar mayores detalles de cómo se llega a ellos, por que la forma de hacerlo se explico en el caso 7 (como seleccionar el algoritmo C4.5) y el caso 3 (como seleccionar el clasificador KDB).

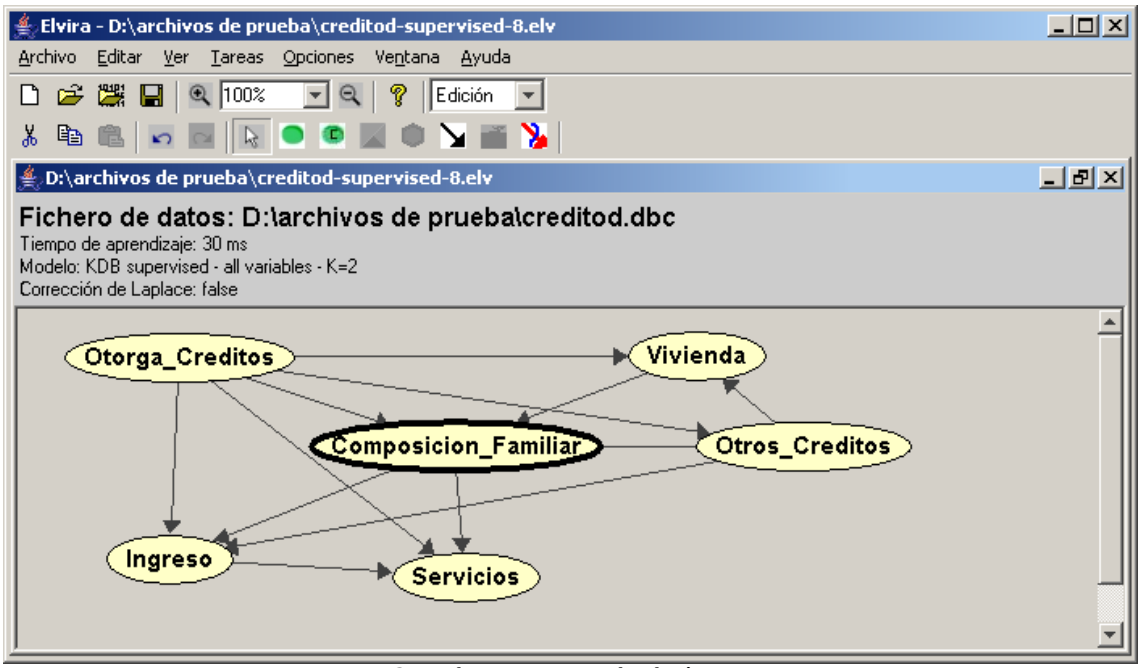

Figura 37: Red Bayesiana en modo edición

Por último, en la figura 38 se indican los valores de probabilidad obtenidos desde el modo Inferencia:

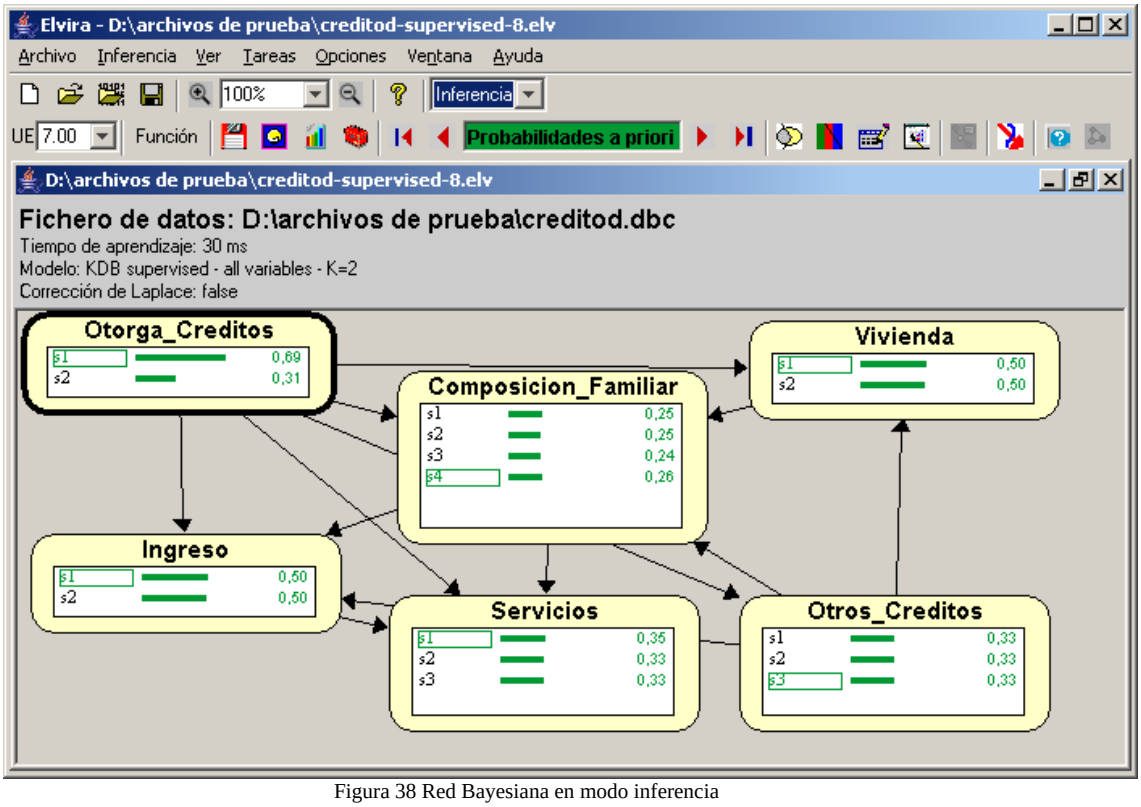

Figura 38 Red Bayesiana en modo inferencia

A continuación se detallan los valores de probabilidad asociados a cada nodo en particular

Valor	Probabilidad
S1	0,6944
S2	0,3056

Tabla 54: Nodo: Otorga_Créditos

Nodo Padre	Valores											
Otorga_Crédito	S1	S1	S1	S1	S1	S1	S2	S2	S2	S2	S2	S2
Otros_Créditos	S1	S1	S2	S2	S3	S3	S1	S1	S2	S2	S3	S3
Vivienda	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
Valor	Probabilidad											
S1	0.2857	0.25	0.1667	0.2609	0	0.2857	0	0	0.5	0	0.25	0.2
S2	0.2857	0.25	0.3333	0.2609	0	0.4286	0	0	0	0	0.25	0
S3	0.2857	0.25	0.3333	0.2609	0	0.2143	0	0	0	0	0.25	0.2
S4	0.1429	0.25	0.1667	0.2173	1	0.0714	1	1	0.5	0	0.25	0.6

Tabla 55: Nodo: Composición_familiar

Nodo Padre	Valores	
Otorga crédito	S1	S2
Valor	Probabilidad	
S1	0.45	0.0682
S2	0.41	0.1591
S3	0.14	0.7727

Tabla 56: Nodo: Otros_Créditos

Nodo Padre	Valores																							
Otorga_ Crédito	S1	S1	S1	S1	S1	S1	S1	S1	S1	S1	S1	S1	S2	S2	S2	S2	S2	S2	S2	S2	S2	S2	S2	S2
Composi-ción_ Familiar	S1	S1	S1	S2	S2	S2	S3	S3	S3	S4	S4	S4	S1	S1	S1	S2	S2	S2	S3	S3	S3	S4	S4	S4
Otros_ Creditos	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3
Valor	Probabilidad																							
S1	0.5	0.6667	0,5	0.5	0.5	0,5	0.5	0.5	0	0.3333	0.25	0	0	0	0.375	0	0	0.5	0	0	0.625	1	1	0.5833
S2	0.5	0.3333	0,5	0.5	0.5	0,5	0.5	0.5	1	0.6667	0.75	1	1	1	0.625	1	1	0.5	1	1	0.375	0	0	0.4167

Tabla 57: Nodo: Ingreso

Nodo Padre	Valores					
Otros_Creditos	S1	S1	S1	S2	S2	S2
Otorga_Crédito	S1	S2	S3	S1	S2	S3
Valor	Probabilidad					
S1	0.4667	0.439	0	1	0.8571	0.7059
S2	0.5333	0.561	1	0	0.1429	0.2941

Tabla 58: Nodo: Vivienda

Nodo Padre	Valores															
Otorga_ Crédito	S1	S1	S1	S1	S1	S1	S1	S1	S2	S2	S2	S2	S2	S2	S2	S2
Composición_ Familiar	S1	S1	S2	S2	S3	S3	S4	S4	S1	S1	S2	S2	S3	S3	S4	S4
Ingreso	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
Valor	Probabilidad															
S1	0.4	0.4	0.3333	0.3333	0.3333	0.3333	0.4	0.3846	0.3333	0.25	0.3333	0.3333	0.6	0.3333	0.2857	0.2
Sz2	0.2667	0.3	0.3333	0.3333	0.4167	0.3333	0.4	0.3077	0.3333	0.375	0.3333	0.3333	0.2	0.3333	0.2857	0.4
S3	0.3333	0.3	0.3334	0.3334	0.25	0.3334	0.2	0.3077	0.3334	0.375	0.3334	0.3334	0.2	0.3334	0.4286	0.4

Tabla 59: Nodo: Servicios:

3. Comparación de los resultados obtenidos con cada Clasificador

3.1. Comparación de las redes obtenidas:

Como se vio las variables definidas para cada nodo y su probabilidad independiente son las mismas independiente del método utilizado para la obtención de la red. Esto se debe a que el análisis que cada método hace sobre los datos de entrada es similar en todos los casos, donde se observan diferencia significativas es en la relación de los nodos de la red y por consiguiente las probabilidades condicionadas. A continuación se muestran un conjunto de tablas que comparan los resultados obtenidos con cada método de generación:

3.1. Comparación de los resultados obtenidos con el clasificador naïve Bayes

Pura	Con Preproceso ID3	Con preproceso C45
Otorga_credito (P) ➢ Ingreso (I) ➢ Otros_Créditos (I) ➢ Composición_Familiar (I) ➢ Vivienda (I) ➢ Servicios (I)	Otorga_credito (P) ➢ Ingreso (I) ➢ Otros_Créditos (I) ➢ Composición_Familiar (I) ➢ Vivienda (I) ➢ Servicios (I)	Otorga_credito (P) ➢ Ingreso (I) ➢ Otros_Créditos (I) ➢ Composición_Familiar (I) ➢ Vivienda (I) ➢ Servicios (I)

Tabla 60: Comparación de las redes obtenidas mediante Naïve Bayes

En la tabla 60 se puede observar que el preprocesamiento no influye en el resultado final obtenido mediante el clasificador Naïve Bayes.

3.1.2. Comparación de los resultados obtenidos con el clasificador TAN

Pura	Con Preproceso ID3	Con preproceso C45
Otorga_credito (P) ➢ Ingreso (I) ➢ Otros_Créditos (I) ➢ Composición_Familiar (I) ➢ Vivienda (I) ➢ Servicios (I) Otros_Créditos (P) ➢ Vivienda (I) ➢ Composición_Familiar(I) Composición_Familar (P) ➢ Servicios (I) ➢ Ingreso (I)	Otorga_credito (P) ➢ Ingreso (I) ➢ Otros_Créditos (I) ➢ Composición_Familiar (I) ➢ Vivienda (I) ➢ Servicios (I) Vivienda (P) ➢ Otros_Créditos (I) Otros_Créditos (P) ➢ Composición_Familiar(I) Composición_Familar (P) ➢ Servicios (I) ➢ Ingreso (I)	Otorga_credito (P) ➢ Ingreso (I) ➢ Otros_Créditos (I) ➢ Composición_Familiar (I) ➢ Vivienda (I) ➢ Servicios (I) Ingreso (P) ➢ Composición_Familiar (I) Composición_Familar (P) ➢ Servicios (I) ➢ Otros_Créditos (I) Otros_Créditos (P) ➢ Vivienda (I)

Tabla 61: Comparación de las redes obtenidas mediante TAN

En la tabla 61 se puede observar que el preprocesamiento provoca cambios en las relaciones de la red obtenida el clasificador TAN.

3.1.3. Comparación de los resultados obtenidos con el clasificador KDB

Pura	Con Preproceso ID3	Con preproceso C45
Otorga_credito (P) ➢ Ingreso (I) ➢ Otros_Créditos (I) ➢ Composición_Familiar (I) ➢ Vivienda (I) ➢ Servicios (I) Ingreso (P) ➢ Servicios (I) Composición_Familar (P) ➢ Servicios (I) ➢ Ingreso (I) Otros_Créditos (P) ➢ Vivienda (I) ➢ Composición_Familiar(I) ➢ Ingreso (I) Vivienda (P) ➢ Composición_Familiar(I)	Otorga_credito (P) ➢ Ingreso (I) ➢ Otros_Créditos (I) ➢ Composición_Familiar (I) ➢ Vivienda (I) ➢ Servicios (I) Ingreso (P) ➢ Servicios (I) Composición_Familar (P) ➢ Servicios (I) ➢ Ingreso (I) Otros_Créditos (P) ➢ Vivienda (I) ➢ Composición_Familiar(I) ➢ Ingreso (I) Vivienda (P) ➢ Composición_Familiar(I)	Otorga_credito (P) ➢ Ingreso (I) ➢ Otros_Créditos (I) ➢ Composición_Familiar (I) ➢ Vivienda (I) ➢ Servicios (I) Ingreso (P) ➢ Servicios (I) Composición_Familar (P) ➢ Servicios (I) ➢ Ingreso (I) Otros_Créditos (P) ➢ Vivienda (I) ➢ Composición_Familiar(I) ➢ Ingreso (I) Vivienda (P) ➢ Composición_Familiar(I)

Tabla 62: Comparación de las redes obtenidas mediante KDB

En la tabla 62 se puede observar que el preprocesamiento no influye en el resultado final obtenido mediante el clasificador KDB

3.1.4. Comparación de los resultados obtenidos para entre los distintos clasificadores sin utilización de preprocesamiento:

Naïve Bayes	TAN	KDB
Otorga_credito (P) ➢ Ingreso (I) ➢ Otros_Créditos (I) ➢ Composición_Familiar (I) ➢ Vivienda (I) ➢ Servicios (I)	Otorga_credito (P) ➢ Ingreso (I) ➢ Otros_Créditos (I) ➢ Composición_Familiar (I) ➢ Vivienda (I) ➢ Servicios (I) Otros_Créditos (P) ➢ Vivienda (I) ➢ Composición_Familiar(I) Composición_Familar (P) ➢ Servicios (I) ➢ Ingreso (I)	Otorga_credito (P) ➢ Ingreso (I) ➢ Otros_Créditos (I) ➢ Composición_Familiar (I) ➢ Vivienda (I) ➢ Servicios (I) Ingreso (P) ➢ Servicios (I) Composición_Familar (P) ➢ Servicios (I) ➢ Ingreso (I) Otros_Créditos (P) ➢ Vivienda (I) ➢ Composición_Familiar(I) ➢ Ingreso (I) Vivienda (P) ➢ Composición_Familiar(I)

Tabla 63: Comparación de las redes obtenidas con los distintos clasificadores sin preproceso

Como puede observarse en la figura 63 las redes obtenidas por los distintos clasificadores son muy distintas en cuanto a su estructura. Esto se debe a que el clasificador naïve Bayes solo genera un nodo padre por cada nodo hijo, mientras que

el clasificador TAN puede tener por cada nodo, además del padre principal (que es el mismo que obtiene el clasificador naïve Bayes), un padre mas y el clasificador KDB aporta k variables padre por cada nodo hijo.

3.1.5. Comparación de los resultados obtenidos entre los distintos clasificadores utilizando el algoritmo ID3 en la etapa de preprocesamiento:

Naïve Bayes	TAN	KDB
Otorga_credito (P) ➢ Ingreso (I) ➢ Otros_Créditos (I) ➢ Composición_Familiar (I) ➢ Vivienda (I) ➢ Servicios (I)	Otorga_credito (P) ➢ Ingreso (I) ➢ Otros_Créditos (I) ➢ Composición_Familiar (I) ➢ Vivienda (I) ➢ Servicios (I) Vivienda (P) ➢ Otros_Créditos (I) Otros_Créditos (P) ➢ Composición_Familiar(I) Composición_Familar (P) ➢ Servicios (I) ➢ Ingreso (I)	Otorga_credito (P) ➢ Ingreso (I) ➢ Otros_Créditos (I) ➢ Composición_Familiar (I) ➢ Vivienda (I) ➢ Servicios (I) Ingreso (P) ➢ Servicios (I) Composición_Familar (P) ➢ Servicios (I) ➢ Ingreso (I) Otros_Créditos (P) ➢ Vivienda (I) ➢ Composición_Familiar(I) ➢ Ingreso (I) Vivienda (P) ➢ Composición_Familiar(I)

Tabla 64: Comparación de las redes obtenidas con los distintos clasificadores prepreprocesando con ID3

3.1.6. Comparación de los resultados obtenidos entre los distintos clasificadores utilizando el algoritmo C4.5 en la etapa de preprocesamiento:

Naïve Bayes	TAN	KDB
Otorga_credito (P) ➢ Ingreso (I) ➢ Otros_Créditos (I) ➢ Composición_Familiar (I) ➢ Vivienda (I) ➢ Servicios (I)	Otorga_credito (P) ➢ Ingreso (I) ➢ Otros_Créditos (I) ➢ Composición_Familiar (I) ➢ Vivienda (I) ➢ Servicios (I) Ingreso (P) ➢ Composición_Familiar (I) Composición_Familar (P) ➢ Servicios (I) ➢ Otros_Créditos (I) Otros_Créditos (P) ➢ Vivienda (I)	Otorga_credito (P) ➢ Ingreso (I) ➢ Otros_Créditos (I) ➢ Composición_Familiar (I) ➢ Vivienda (I) ➢ Servicios (I) Ingreso (P) ➢ Servicios (I) Composición_Familar (P) ➢ Servicios (I) ➢ Ingreso (I) Otros_Créditos (P) ➢ Vivienda (I) ➢ Composición_Familiar(I) ➢ Ingreso (I) Vivienda (P) ➢ Composición_Familiar(I)

Tabla 65: Comparación de las redes obtenidas con los distintos clasificadores prepreprocesando con ID3

3.2. Comparación de los tiempos de proceso

En este punto se analizan los tiempos de proceso de cada clasificador tratando un archivo con 150 registros :

Clasificador	Tiempo de proceso
Naïve Bayes	Menor al segundo
TAN	Aproximadamente 2 segundos
KDB	Aproximadamente 4 minutos

Tabla 66: Comparación de tiempos de procesamiento

Como puede verse la complejidad de cálculo del Clasificador KDB hace que los tiempos de proceso sean considerablemente superior a los mostrados por los otros algoritmos.
Desde el punto de vista de la performance podría decirse que el Clasificador TAN es el mas conveniente, ya que no solo aporta un tiempo de proceso pequeño sino que además genera una red mas completa que la generada con Naïve Bayes.

3.3. Respuesta de las distintas redes al ingreso de evidencia

Antes de comenzar con el análisis de los resultados obtenidos mediante la instaciación de los nodos de cada red, vamos a mostrar como se instancian los nodos de una red en el software Elvira

3.3.1. Instanciación de nodos

En primer lugar el software Elvira debe estar en modo de trabajo “Inferencia” (como se muestra en la figura 39).

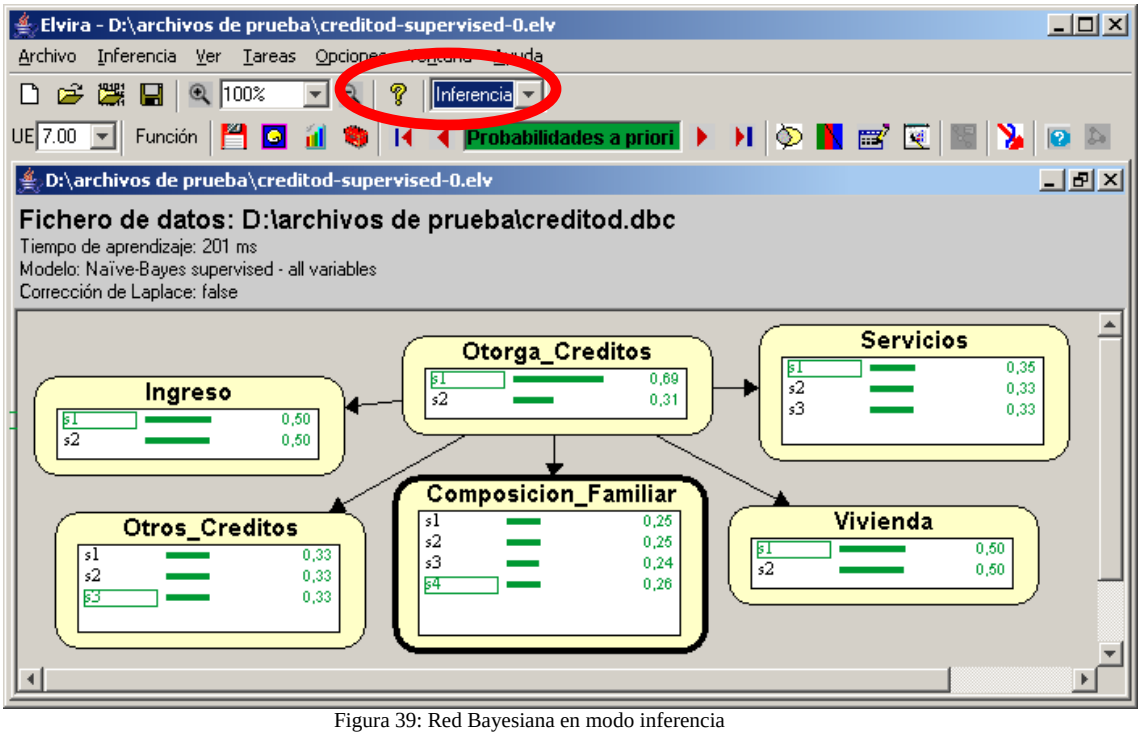

Figura 39: Red Bayesiana en modo inferencia

Desde esta pantalla y con solo hacer doble click sobre el valor del nodo que deseamos instanciar, el programa inmediatamente recalculara los valores de probabilidad de toda la red. A continuación en la figura 40 se muestra la respuesta del software ante la instaciación del valor “S1” del nodo “Otorga_Creditos”:

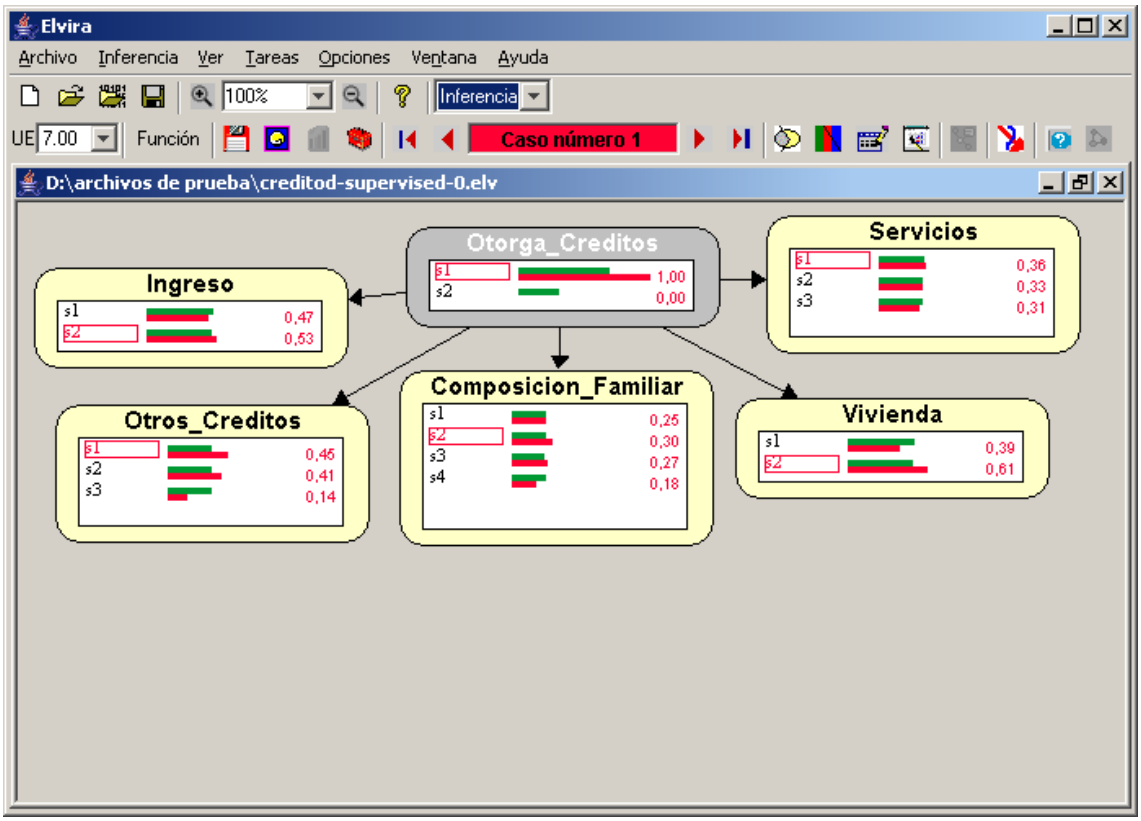

Figura 40: Red Bayesiana en modo inferencia

La figura 40 muestra la respuesta del sistema ante una instanciación, mediante el color verde se indican los valores de probabilidad a priori (sin evidencia) y con color rojo se detallan los valores de probabilidad fruto del ingreso de la evidencia.

A continuación vamos a analizar como se comportaron las distintas redes ante el ingreso de la evidencia.

3.3.2. Análisis de las probabilidades ante la ausencia de evidencia

Los valores de probabilidades asignadas a cada valor sin ingreso de evidencia son iguales para todos los clasificadores. A continuación, en la tabla 67, se detalla una tabla que muestra estos valores.

Nodo	Valor	Probabilidad
Otorga_Créditos	S1	0.69
	S2	0.39
Composición_Familiar	S1	0.25
	S2	0.25
	S3	0.24
	S4	0.26
Vivienda	S1	0.50
	S2	0.50
Ingreso	S1	0.50
	S2	0.50
Servicios	S1	0.35
	S2	0.33
	S3	0.33
Otros_créditos	S1	0.33
	S2	0.33
	S3	0.33

Tabla 67: Valores de probabilidad a priori

3.3.3 Análisis de las probabilidades ante el ingreso de evidencia

3.3.3.1. Comparación de redes que no fueron preprocesadas con otros algoritmos

Para comprender mas fácilmente las tablas se utilizará el color rojo para mostrar los valores de probabilidad del nodo que recibió la evidencia y el color azul para indicar los valores de probabilidad que cambiaron con el ingreso de la evidencia.

Comparación de la respuesta de las distintas redes ante el ingreso de evidencia en el nodo Otorga_Créditos, la cual indica que se confirma que el valor S1 es verdadero

Nodo	Valor	Probabilidad Sin evidencia	Probabilidad Naïve Bayes	Probabilidad TAN	Probabilidad KDB
Otorga_Créditos	S1	0.69	1	1	1
	S2	0.39	0	0	0
Composición_Familiar	S1	0.25	0.25	0.25	0.25
	S2	0.25	0.30	0.30	0.30
	S3	0.24	0.27	0.27	0.27
	S4	0.26	0.18	0.18	0.18
Vivienda	S1	0.50	0.39	0.39	0.39
	S2	0.50	0.61	0.61	0.61
Ingreso	S1	0.50	0.47	0.47	0.47
	S2	0.50	0.53	0.53	0.53
Servicios	S1	0.35	0.36	0.36	0.36
	S2	0.33	0.33	0.33	0.33
	S3	0.33	0.31	0.31	0.31
Otros_créditos	S1	0.33	0.45	0.45	0.45
	S2	0.33	0.41	0.41	0.41
	S3	0.33	0.14	0.14	0.14

Tabla 68: Instaciación del nodo Otorga_Credito

Comparación de la respuesta de las distintas redes ante el ingreso de evidencia en el nodo Composición_Familiar, la cual indica que se confirma que el valor S1 es verdadero

Nodo	Valor	Probabilidad Sin evidencia	Probabilidad Naïve Bayes	Probabilidad TAN	Probabilidad KDB
Otorga_Créditos	S1	0.69	0.69	0.69	0.69
	S2	0.39	0.31	0.31	0.31
Composición_Familiar	S1	0.25	1	1	1
	S2	0.25	0	0	0
	S3	0.24	0	0	0
	S4	0.26	0	0	0
Vivienda	S1	0.50	0.50	0.49	0.50
	S2	0.50	0.50	0.51	0.50
Ingreso	S1	0.50	0.50	0.50	0.50
	S2	0.50	0.50	0.50	0.50
Servicios	S1	0.35	0.35	0.36	0.36
	S2	0.33	0.33	0.31	0.31
	S3	0.33	0.33	0.33	0.33
Otros_créditos	S1	0.33	0.33	0.33	0.33
	S2	0.33	0.33	0.33	0.33
	S3	0.33	0.33	0.33	0.33

Tabla 69: Instaciación del nodo Composición Familiar

Comparación de la respuesta de las distintas redes ante el ingreso de evidencia en el nodo Vivienda, la cual indica que se confirma que el valor S1 es verdadero

Nodo	Valor	Probabilidad Sin evidencia	Probabilidad Naïve Bayes	Probabilidad TAN	Probabilidad KDB
Otorga_Créditos	S1	0.69	0.54	0.54	0.54
	S2	0.39	0.46	0.46	0.46
Composición_Familiar	S1	0.25	0.25	0.25	0.25
	S2	0.25	0.23	0.21	0.25
	S3	0.24	0.23	0.23	0.25
	S4	0.26	0.30	0.31	0.25
Vivienda	S1	0.50	1	1	1
	S2	0.50	0	0	0
Ingreso	S1	0.50	0.52	0.51	0.52
	S2	0.50	0.48	0.49	0.48
Servicios	S1	0.35	0.34	0.34	0.34
	S2	0.33	0.32	0.33	0.33
	S3	0.33	0.33	0.34	0.33
Otros_créditos	S1	0.33	0.27	0.33	0.33
	S2	0.33	0.29	0.33	0.33
	S3	0.33	0.43	0.33	0.33

Tabla 70: Instaciación del nodo vivienda

Comparación de la respuesta de las distintas redes ante el ingreso de evidencia en el nodo Servicios, la cual indica que se confirma que el valor S1 es verdadero

Nodo	Valor	Probabilidad Sin evidencia	Probabilidad Naïve Bayes	Probabilidad TAN	Probabilidad KDB
Otorga_Créditos	S1	0.69	0.65	0.65	0.65
	S2	0.39	0.36	0.35	0.35
Composición_Familiar	S1	0.25	0.25	0.25	0.25
	S2	0.25	0.24	0.25	0.25
	S3	0.24	0.24	0.24	0.24
	S4	0.26	0.27	0.26	0.26
Vivienda	S1	0.50	0.52	0.51	0.52
	S2	0.50	0.48	0.49	0.48
Ingreso	S1	0.50	1	1	1
	S2	0.50	0	0	0
Servicios	S1	0.35	0.35	0.35	0.36
	S2	0.33	0.33	0.33	0.32
	S3	0.33	0.33	0.33	0.32
Otros_créditos	S1	0.33	0.32	0.32	0.33
	S2	0.33	0.32	0.32	0.33
	S3	0.33	0.36	0.36	0.33

Tabla 71: Instaciación del nodo Ingreso

Comparación de la respuesta de las distintas redes ante el ingreso de evidencia en el nodo Servicios, la cual indica que se confirma que el valor S1 es verdadero

Nodo	Valor	Probabilidad Sin evidencia	Probabilidad Naïve Bayes	Probabilidad TAN	Probabilidad KDB
Otorga_Créditos	S1	0.69	0.72	0.72	0.72
	S2	0.39	0.28	0.28	0.28
Composición_Familiar	S1	0.25	0.25	0.26	0.26
	S2	0.25	0.25	0.24	0.24
	S3	0.24	0.25	0.26	0.26
	S4	0.26	0.25	0.24	0.24
Vivienda	S1	0.50	0.49	0.49	0.49
	S2	0.50	0.51	0.51	0.51
Ingreso	S1	0.50	0.50	0.50	0.52
	S2	0.50	0.50	0.50	0.48
Servicios	S1	0.35	1	1	1
	S2	0.33	0	0	0
	S3	0.33	0	0	0
Otros_créditos	S1	0.33	0.34	0.34	0.34
	S2	0.33	0.34	0.33	0.33
	S3	0.33	0.32	0.33	0.32

Tabla 72: Instaciación del nodo Servicios

Comparación de la respuesta de las distintas redes ante el ingreso de evidencia en el nodo Otros_créditos, la cual indica que se confirma que el valor S1 es verdadero

Nodo	Valor	Probabilidad Sin evidencia	Probabilidad Naïve Bayes	Probabilidad TAN	Probabilidad KDB
Otorga_Créditos	S1	0.69	0.94	0.94	0.94
	S2	0.39	0.06	0.06	0.06
Composición_Familiar	S1	0.25	0.25	0.25	0.25
	S2	0.25	0.29	0.25	0.25
	S3	0.24	0.26	0.25	0.25
	S4	0.26	0.20	0.25	0.25
Vivienda	S1	0.50	0.41	0.50	0.50
	S2	0.50	0.59	0.50	0.50
Ingreso	S1	0.50	0.48	0.48	0.50
	S2	0.50	0.52	0.52	0.50
Servicios	S1	0.35	0.36	0.36	0.36
	S2	0.33	0.33	0.33	0.33
	S3	0.33	0.31	0.32	0.31
Otros_créditos	S1	0.33	1	1	1
	S2	0.33	0	0	0
	S3	0.33	0	0	0

Tabla 73: Instaciación del nodo Otros_creditos

3.3.3.2. Comparación de redes preprocesadas con el algoritmo ID3

Comparación de la respuesta de las distintas redes ante el ingreso de evidencia en el nodo Otorga_Créditos, la cual indica que se confirma que el valor S1 es verdadero

Nodo	Valor	Probabilidad Sin evidencia	Probabilidad Naïve Bayes	Probabilidad TAN	Probabilidad KDB
Otorga_Créditos	S1	0.69	1	1	1
	S2	0.39	0	0	0
Composición_Familiar	S1	0.25	0.25	0.25	0.25
	S2	0.25	0.30	0.30	0.30
	S3	0.24	0.27	0.27	0.27
	S4	0.26	0.18	0.18	0.18
Vivienda	S1	0.50	0.39	0.39	0.39
	S2	0.50	0.61	0.61	0.61
Ingreso	S1	0.50	0.47	0.47	0.47
	S2	0.50	0.53	0.53	0.53
Servicios	S1	0.35	0.36	0.36	0.36
	S2	0.33	0.33	0.33	0.33
	S3	0.33	0.31	0.31	0.31
Otros_créditos	S1	0.33	0.45	0.45	0.45
	S2	0.33	0.41	0.41	0.41
	S3	0.33	0.14	0.14	0.14

Tabla 74: Instaciación del nodo Otorga_Creditos

Comparación de la respuesta de las distintas redes ante el ingreso de evidencia en el nodo Composición_Familiar, la cual indica que se confirma que el valor S1 es verdadero

Nodo	Valor	Probabilidad Sin evidencia	Probabilidad Naïve Bayes	Probabilidad TAN	Probabilidad KDB
Otorga_Créditos	S1	0.69	0.69	0.69	0.69
	S2	0.39	0.31	0.31	0.31
Composición_Familiar	S1	0.25	1	1	1
	S2	0.25	0	0	0
	S3	0.24	0	0	0
	S4	0.26	0	0	0
Vivienda	S1	0.50	0.50	0.49	0.50
	S2	0.50	0.50	0.51	0.50
Ingreso	S1	0.50	0.50	0.50	0.50
	S2	0.50	0.50	0.50	0.50
Servicios	S1	0.35	0.35	0.36	0.36
	S2	0.33	0.33	0.31	0.31
	S3	0.33	0.33	0.33	0.33
Otros_créditos	S1	0.33	0.33	0.33	0.33
	S2	0.33	0.33	0.33	0.33
	S3	0.33	0.33	0.33	0.33

Tabla 75: Instaciación del nodo Composición_familiar

Comparación de la respuesta de las distintas redes ante el ingreso de evidencia en el nodo Vivienda, la cual indica que se confirma que el valor S1 es verdadero

Nodo	Valor	Probabilidad Sin evidencia	Probabilidad Naïve Bayes	Probabilidad TAN	Probabilidad KDB
Otorga_Créditos	S1	0.69	0.54	0.54	0.54
	S2	0.39	0.46	0.46	0.46
Composición_Familiar	S1	0.25	0.25	0.25	0.25
	S2	0.25	0.23	0.21	0.25
	S3	0.24	0.23	0.23	0.25
	S4	0.26	0.30	0.31	0.25
Vivienda	S1	0.50	1	1	1
	S2	0.50	0	0	0
Ingreso	S1	0.50	0.52	0.51	0.52
	S2	0.50	0.48	0.49	0.48
Servicios	S1	0.35	0.34	0.34	0.34
	S2	0.33	0.32	0.33	0.33
	S3	0.33	0.33	0.34	0.33
Otros_créditos	S1	0.33	0.27	0.33	0.33
	S2	0.33	0.29	0.33	0.33
	S3	0.33	0.43	0.33	0.33

Tabla 76: Instaciación del nodo Vivienda

Comparación de la respuesta de las distintas redes ante el ingreso de evidencia en el nodo Servicios, la cual indica que se confirma que el valor S1 es verdadero

Nodo	Valor	Probabilidad Sin evidencia	Probabilidad Naïve Bayes	Probabilidad TAN	Probabilidad KDB
Otorga_Créditos	S1	0.69	0.65	0.65	0.65
	S2	0.39	0.36	0.35	0.35
Composición_Familiar	S1	0.25	0.25	0.25	0.25
	S2	0.25	0.24	0.25	0.25
	S3	0.24	0.24	0.24	0.24
	S4	0.26	0.27	0.26	0.26
Vivienda	S1	0.50	0.52	0.51	0.52
	S2	0.50	0.48	0.49	0.48
Ingreso	S1	0.50	1	1	1
	S2	0.50	0	0	0
Servicios	S1	0.35	0.35	0.35	0.36
	S2	0.33	0.33	0.32	0.32
	S3	0.33	0.33	0.33	0.32
Otros_créditos	S1	0.33	0.32	0.32	0.33
	S2	0.33	0.32	0.32	0.33
	S3	0.33	0.36	0.36	0.33

Tabla 77: Instaciación del nodo Ingreso

Comparación de la respuesta de las distintas redes ante el ingreso de evidencia en el nodo Servicios, la cual indica que se confirma que el valor S1 es verdadero

Nodo	Valor	Probabilidad Sin evidencia	Probabilidad Naïve Bayes	Probabilidad TAN	Probabilidad KDB
Otorga_Créditos	S1	0.69	0.72	0.72	0.72
	S2	0.39	0.28	0.28	0.28
Composición_Familiar	S1	0.25	0.25	0.26	0.26
	S2	0.25	0.25	0.24	0.24
	S3	0.24	0.25	0.26	0.26
	S4	0.26	0.25	0.24	0.24
Vivienda	S1	0.50	0.49	0.49	0.49
	S2	0.50	0.51	0.51	0.51
Ingreso	S1	0.50	0.50	0.50	0.52
	S2	0.50	0.50	0.50	0.48
Servicios	S1	0.35	1	1	1
	S2	0.33	0	0	0
	S3	0.33	0	0	0
Otros_créditos	S1	0.33	0.34	0.34	0.34
	S2	0.33	0.34	0.33	0.33
	S3	0.33	0.32	0.33	0.32

Tabla 78: Instaciación del nodo Servicios

Comparación de la respuesta de las distintas redes ante el ingreso de evidencia en el nodo Otros_créditos, la cual indica que se confirma que el valor S1 es verdadero

Nodo	Valor	Probabilidad Sin evidencia	Probabilidad Naïve Bayes	Probabilidad TAN	Probabilidad KDB
Otorga_Créditos	S1	0.69	0.94	0.94	0.94
	S2	0.39	0.06	0.06	0.06
Composición_Familiar	S1	0.25	0.25	0.25	0.25
	S2	0.25	0.29	0.25	0.25
	S3	0.24	0.26	0.25	0.25
	S4	0.26	0.20	0.25	0.25
Vivienda	S1	0.50	0.41	0.50	0.50
	S2	0.50	0.59	0.50	0.50
Ingreso	S1	0.50	0.48	0.48	0.50
	S2	0.50	0.52	0.52	0.50
Servicios	S1	0.35	0.36	0.36	0.36
	S2	0.33	0.33	0.33	0.33
	S3	0.33	0.31	0.32	0.31
Otros_créditos	S1	0.33	1	1	1
	S2	0.33	0	0	0
	S3	0.33	0	0	0

Tabla 79: Instaciación del nodo Otros_creditos

3.3.3.3. Comparación de redes preprocesadas con el algoritmo C4.5

Comparación de la respuesta de las distintas redes ante el ingreso de evidencia en el nodo Otorga_Créditos, la cual indica que se confirma que el valor S1 es verdadero

Nodo	Valor	Probabilidad Sin evidencia	Probabilidad Naïve Bayes	Probabilidad TAN	Probabilidad KDB
Otorga_Créditos	S1	0.69	1	1	1
	S2	0.39	0	0	0
Composición_Familiar	S1	0.25	0.25	0.25	0.25
	S2	0.25	0.30	0.30	0.30
	S3	0.24	0.27	0.27	0.27
	S4	0.26	0.18	0.18	0.18
Vivienda	S1	0.50	0.39	0.39	0.39
	S2	0.50	0.61	0.61	0.61
Ingreso	S1	0.50	0.47	0.47	0.47
	S2	0.50	0.53	0.53	0.53
Servicios	S1	0.35	0.36	0.36	0.36
	S2	0.33	0.33	0.33	0.33
	S3	0.33	0.31	0.31	0.31
Otros_créditos	S1	0.33	0.45	0.45	0.45
	S2	0.33	0.41	0.41	0.41
	S3	0.33	0.14	0.14	0.14

Tabla 80: Instaciación del nodo Otorga_Creditos

Comparación de la respuesta de las distintas redes ante el ingreso de evidencia en el nodo Composición_Familiar, la cual indica que se confirma que el valor S1 es verdadero

Nodo	Valor	Probabilidad Sin evidencia	Probabilidad Naïve Bayes	Probabilidad TAN	Probabilidad KDB
Otorga_Créditos	S1	0.69	0.69	0.69	0.69
	S2	0.39	0.31	0.31	0.31
Composición_Familiar	S1	0.25	1	1	1
	S2	0.25	0	0	0
	S3	0.24	0	0	0
	S4	0.26	0	0	0
Vivienda	S1	0.50	0.50	0.49	0.50
	S2	0.50	0.50	0.51	0.50
Ingreso	S1	0.50	0.50	0.50	0.50
	S2	0.50	0.50	0.50	0.50
Servicios	S1	0.35	0.35	0.36	0.36
	S2	0.33	0.33	0.31	0.31
	S3	0.33	0.33	0.33	0.33
Otros_créditos	S1	0.33	0.33	0.33	0.33
	S2	0.33	0.33	0.33	0.33
	S3	0.33	0.33	0.33	0.33

Tabla 81: Instaciación del nodo Composición_Familiar

Comparación de la respuesta de las distintas redes ante el ingreso de evidencia en el nodo Vivienda, la cual indica que se confirma que el valor S1 es verdadero

Nodo	Valor	Probabilidad Sin evidencia	Probabilidad Naïve Bayes	Probabilidad TAN	Probabilidad KDB
Otorga_Créditos	S1	0.69	0.54	0.54	0.54
	S2	0.39	0.46	0.46	0.46
Composición_Familiar	S1	0.25	0.25	0.25	0.25
	S2	0.25	0.23	0.21	0.25
	S3	0.24	0.23	0.23	0.25
	S4	0.26	0.30	0.31	0.25
Vivienda	S1	0.50	1	1	1
	S2	0.50	0	0	0
Ingreso	S1	0.50	0.52	0.51	0.52
	S2	0.50	0.48	0.49	0.48
Servicios	S1	0.35	0.34	0.34	0.34
	S2	0.33	0.32	0.33	0.33
	S3	0.33	0.33	0.34	0.33
Otros_créditos	S1	0.33	0.27	0.33	0.33
	S2	0.33	0.29	0.33	0.33
	S3	0.33	0.43	0.33	0.33

Tabla 82: Instaciación del nodo Vivienda

Comparación de la respuesta de las distintas redes ante el ingreso de evidencia en el nodo Servicios, la cual indica que se confirma que el valor S1 es verdadero

Nodo	Valor	Probabilidad Sin evidencia	Probabilidad Naïve Bayes	Probabilidad TAN	Probabilidad KDB
Otorga_Créditos	S1	0.69	0.65	0.65	0.65
	S2	0.39	0.35	0.35	0.35
Composición_Familiar	S1	0.25	0.25	0.25	0.25
	S2	0.25	0.24	0.25	0.25
	S3	0.24	0.24	0.24	0.24
	S4	0.26	0.27	0.26	0.26
Vivienda	S1	0.50	0.52	0.51	0.52
	S2	0.50	0.48	0.49	0.48
Ingreso	S1	0.50	1	1	1
	S2	0.50	0	0	0
Servicios	S1	0.35	0.35	0.35	0.36
	S2	0.33	0.33	0.32	0.32
	S3	0.33	0.33	0.33	0.32
Otros_créditos	S1	0.33	0.32	0.32	0.33
	S2	0.33	0.32	0.32	0.33
	S3	0.33	0.36	0.36	0.33

Tabla 83: Instaciación del nodo Ingreso

Comparación de la respuesta de las distintas redes ante el ingreso de evidencia en el nodo Servicios, la cual indica que se confirma que el valor S1 es verdadero

Nodo	Valor	Probabilidad Sin evidencia	Probabilidad Naïve Bayes	Probabilidad TAN	Probabilidad KDB
Otorga_Créditos	S1	0.69	0.72	0.72	0.72
	S2	0.39	0.28	0.28	0.28
Composición_Familiar	S1	0.25	0.25	0.26	0.26
	S2	0.25	0.25	0.24	0.24
	S3	0.24	0.25	0.26	0.26
	S4	0.26	0.25	0.24	0.24
Vivienda	S1	0.50	0.49	0.49	0.49
	S2	0.50	0.51	0.51	0.51
Ingreso	S1	0.50	0.50	0.50	0.52
	S2	0.50	0.50	0.50	0.48
Servicios	S1	0.35	1	1	1
	S2	0.33	0	0	0
	S3	0.33	0	0	0
Otros_créditos	S1	0.33	0.34	0.34	0.34
	S2	0.33	0.34	0.33	0.33
	S3	0.33	0.32	0.33	0.32

Tabla 84: Instaciación del nodo Servicios

Comparación de la respuesta de las distintas redes ante el ingreso de evidencia en el nodo Otros_créditos, la cual indica que se confirma que el valor S1 es verdadero

Nodo	Valor	Probabilidad Sin evidencia	Probabilidad Naïve Bayes	Probabilidad TAN	Probabilidad KDB
Otorga_Créditos	S1	0.69	0.94	0.94	0.94
	S2	0.39	0.06	0.06	0.06
Composición_Familiar	S1	0.25	0.25	0.25	0.25
	S2	0.25	0.29	0.25	0.25
	S3	0.24	0.26	0.25	0.25
	S4	0.26	0.20	0.25	0.25
Vivienda	S1	0.50	0.41	0.50	0.50
	S2	0.50	0.59	0.50	0.50
Ingreso	S1	0.50	0.48	0.48	0.50
	S2	0.50	0.52	0.52	0.50
Servicios	S1	0.35	0.36	0.36	0.36
	S2	0.33	0.33	0.33	0.33
	S3	0.33	0.31	0.32	0.31
Otros_créditos	S1	0.33	1	1	1
	S2	0.33	0	0	0
	S3	0.33	0	0	0

Tabla 84: Instaciación del nodo Otros_creditos

4. Conclusión

En base a las comparaciones realizadas en el punto 3 se puede decir el que algoritmo naïve Bayes puro es un algoritmo eficiente. A pesar de que las redes que genera no tiene el nivel de presición de los clasificadores TAN y KDB, las respuesta que genera ante el ingreso de evidencia son muy similares a las brindadas por estos otros clasificadores, que en esencia son mucho mas complejos de desarrollar y lentos en cuanto a los tiempos de procesamiento requeridos para la generación de las redes.

Si bien el preprocesamiento de los archivos con un algoritmo de inducción de árboles de decisión genera pequeños cambios en la distribución de probabilidades a priori, sobre todo en las redes generadas con el clasificador TAN, de los valores de cada nodo, esta variación no genera cambios significativos cuando se ingresa evidencia en las redes.

El programa Elvira a mostrado ser muy eficiente desde el punto de vista de la generación de redes Bayesianas, desde archivos de datos, y en el tratamiento de estas redes luego de haber sido generadas.

5. Referencias

[Beinlich, I. et al, 1989] Beinlich, I.; Suermondt, H.; Chavez, R.; Cooper, G.; 1989; ***The Alarm Monitoring System: A case study with two probabilistic inference techniques for belief networks***. In proceedings of the 2º European Conference on Artificial Intelligence in Medicine.

[Blurock, E., 1996] Blurock, E.; 1996; ***The ID3 Algorithm, Research Institute Institute for Symbolic Computation;*** www.risc.unilinz.ac.at/people/blurock/analysis/manual/document ; Australia.

[Breese & Blake, 1995] Breese, J.; Blake, R.; 1995; ***Automating computer bottleneck detection with belief nets***. Proceeding of the conference on Uncertainty in artificial Intelligence. Morga Kaufmann, San Francisco, CA.

[Chen, M. et al, 1996] Chen, M.; Han, j.; Yu, P.; 1996; ***Data Mining: An overview from Database perspective.*** IEEE Transaction on Knowledge and Data Eng.

[Diaz, F. et al, 1999] Diaz, F.; Corchado, J.; 1999; ***Rough sets bases learning for Bayesian networks. International workshop on objetive bayesian methodology***; Valencia, Spain.

[Evangelos, S. et al, 1996] Evangelos, S.; Han J.; 1996; ***Proceeding of the Second International Conference on Knowledge Discovery and Data Mining.*** Portland, EE.UU.

[Felgaer, P. et al, 2003] Felgaer, P.; Britos, P.; Sicre, J.; Servetto, A.; García-Martínez, R. y Perichinsky, G.; 2003; ***Optimización de Redes Bayesianas Basada en Técnicas de Aprendizaje por Instrucción.***; Proceedings del VIII Congreso Argentino de Ciencias de la Computación. Pág. 1687.

[Fritz, W. et al, 1989] Fritz, W.; García Martínez, R.; Blanque, J.; Rama, A.; Adobbati, R.; Samo, M.; 1989; ***The Autonomous Intelligence System. Robotics and Autonomous System.*** Vol. 5 Nber. 2. Elsevier.

[Gallion, R. et al, 1993] Gallion, R.; St Clair, D.; Sabharwal, C.; Bond, W.; 1993; ***Dynamic ID3: A Symbolic Learning Algorithm for Many-Valued Attribute Domains.*** Engineering Education Center, University of Missouri-Rolla, St. Luis, EE.UU.

[García Martínez, R., 1993] García Martínez, R.; 1993; ***Heuristic Theory Formation as a machine learning method.*** Proceeding VI International Symposium on Artificial Inteligence. Páginas 294 a 298. Editorial Limusa. Mexico.

[García Martínez, R., 1995] García Martínez, R.; 1995; ***Theory Formation by Heuristics.*** Proceeding of the II International Congress on Information Engineering. PP 200-205. School of Engineering Pess. University of Buenos Aires.

[García Martínez, R., 1997] García Martínez, R.; 1997; ***Sistemas Autónomos: Aprendizaje Automático***; Nueva Librería, Buenos Aires, Argentina.

[García Martínez, R. et al, 2003] García Martínez R.; Pasqín D.;2003; ***Sistemas Inteligentes***; Capítulo 1: "Aprendizaje Automático", Capítulo 2: "Redes Neuronales Artificiales"; Nueva Librería, Buenos Aires, Argentina.

[García Martínez &Borrajo, 2000] García Martínez, R.; Borrajo D.; 2000; ***An integrated approach of learning, planning and executing;*** Journal of Intelligent and Robotic Systems. Volumen 29, Número 1, páginas 47 a 78. Kluwer Academic Press.

[Heckerman, D., 1995] Heckerman, D.; 1995; ***A Tutorial on learning bayesian network.*** Technical report MSR-TR-95-06, Microsoft research , Redmond, WA.

[Heckerman D. et al, 1996] Heckerman, D.; Chickering, M.; 1996; ***Efficient approximation for the marginal likelihood on incomplete data given a bayesian network.*** Technical report MSR-TR-96-8, Microsoft Research, , Microsoft Corporation.

[Hernández O.J., 2000] Hernández Orallo, J.; 2000;***Extracción Automática de Conocimiento de base de datos e ingeniería del software.*** Programación declarativa e ingeniería de la programación.

[Hernández O.J. et al, 2004] Hernández Orallo, J.; Ferri Ramírez, C.; Ramírez Quintana J.;2004;***Introducción a la minería de datos***; Capítulo 10: "Métodos Bayesianos"; PEARSON EDUCACION.

[Holsheimer & Siebes, 1994] Holsheimer, M. & Siebes A.; 1994 ;***Data Mining: the search for knowledge in databases.*** Computer Science/Departament of Algorithmics and Architecture, Centrum voor Wiskunde en informática , CS-R9406; Ámsterdam, Holanda.

[Larrañaga, P. et al, 2004] Larrañaga, P. e Inza, I.; ***Clasificadores Bayesianos;*** Departamento de Ciencias de la Computación e Inteligencia Artificial; Universidad del País Vasco-Euskal Herreiko Unibertsitatea; 2004; http:/www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/t6bayesianos.pdf

[Mannilla, H., 1997] Mannilla, H.; 1997; ***Methods and problems in data mining.*** Inc. Proc. of International Conference on Database Theory, Delphi, Greece.

[Michalski, R. et al, 1982] Michalski R.; Baskin, A.; Spackman, K.; 1982; ***A Logic-Based Approach to Conceptual Database Analysis,*** Sixth Annual Symposium on Computer Applications on Medical Care, George Washington University, Medical Center, Washington, DC, EE.UU.

[Michalski, R. et al, 1983] Michalski R.; 1983; ***A Theory and Methodology of Inductive Learning;*** Morgan-Kauffman; EE.UU.

[Michalski, R. et al, 1998] Michalski R.; Bratko, I.; Kubat, M.; 1998; ***Machine Learning and data mining . Methods and Applications;*** Wiley & Sons Ltd.; EE.UU:

[Ochoa, M. et al, 2004] Ochoa, M.; García Martínez, R.; Britos P.; 2004; *Trabajo final de Especialidad en Ingeniería de Sistemas Expertos* ***Herramientas Inteligentes para Explotación de Información;*** Instituto Tecnológico de Buenos Aires.

[Pearl, J., 1988] Pearl, J.; 1988; ***Probabilistic reasoning in intelligent systems***. Morgan Kaufmann, San Mateo, CA.

[Perichinsky, G. et al, 2000] Perichinsky,G.; García-Martínez, R.; Proto, A.; 2000; ***Knowledge Discovery Based on Computational Taxonomy And Intelligent Data Mining;*** CD del VI Congreso Argentino de Ciencias de la Computación. (\cacic2k\cacic\sp\is-039\IS-039.htm). Usuhaia. Octubre 2 al 6.

[Perichinsky, G. et al, 2001] Perichinsky, G.; Garcia Martinez, R.; Proto, A.; Sevetto, A.; Grossi, D.; 2001; ***Integrated Environment of Systems Automated Engineering*;** Proceedings del II Workshop de Investigadores en Ciencias de la Computación. Mayo. Editado por Universidad Nacional de San Luis en el CD Wicc2001:\Wiccflash\Areas\IngSoft\Integrated_Environment.pdf .

[Perichinsky, G. et al, 2003] Perichinsky, G.; Servente, M.; Servetto, A.; García-Martínez, R.; Orellana, R.; Plastino, A.; 2003; ***Taxonomic Evidence and Robustness of the Classification Appying Intelligent Data Mining*;** Proceedings del VIII Congreso Argentino de Ciencias de la Computación. Pág. 1797-1808.

[Perichinsky, & García M., 2000] Perichinsky, G. y Garcia Martinez, R.; 2000; ***A Data Mining Approach to Computational Taxonomy;*** Proceedings del Workshop de Investigadores en Ciencias de la Computación. Páginas 107-110. Editado por Departamento de Publicaciones de la Facultad de Informática. Universidad Nacional de La Plata. Mayo.

[Piatetski-Shapiro, G. et al, 1991] Piatetski-Shapiro, G.; Frawley, W.; Matheus C.; 1991; ***Knowledge discovery in databases: an overview***; AAAI-MIP Press; Menlo Park; California.

[Piatetski-Shapiro, G. et al, 1996] Piatetski-Shapiro, G.; Fayyad, U.; Smith P.; 1996; ***From data mining to Knowledge discovery***; AAAI Press/MIT Press; CA.

[Quinlan, J, 1993a] Quinlan J.; 1993; ***The effect of Noise on Concept Learning***, en R. Michalski, J. Carbonell & T. Mitchells (Eds.) Machine Learning, The Artificial Intelligence Approach. Morgan Kaufmann, Vol. I, Capítulo 6, páginas 149 a 167. San Mateo, CA: Morgan Kaufmann, EE.UU.

[Quinlan, J.,1993b] Quinlan J.; 1993; ***Learning Efficient Clasifications Procedures and Their Application to Chess Games***, en R. Michalski, J. Carbonell & T. Mitchells (Eds.) Machine Learning, The Artificial Intelligence Approach. Morgan Kaufmann, Vol. II, Capítulo 15, páginas 463 a 482. EE.UU.

[Servente, M. et al, 2002] Servente M. ; García Martínez R.;2002; *Tesis Doctoral* ***Algoritmos TDIDT aplicados a la minería de datos inteligente;*** Universidad de Buenos Aires.

[Ramoni & Sebastiani et al, 1996] Ramoni, M; Sebastián P.; 1996; ***Learning Bayesian networks from incomplete databases***. Technical report KMI-TR-43, knowledye Media Institute, The open University.