Trabajo Final de Especialidad en
Tecnologías de Explotación de Información

NIVEL DE SIGNIFICACIÓN ESTADÍSTICA PARA EL APRENDIZAJE DE UNA RED BAYESIANA

Autor: Matilde Inés Césari (Lic. en Sistemas de Información)

Directores:
Dr. Ramón García Martínez, M. Ing. Paola Britos

Asesor: Esp. Ing. Ricardo M. Césari

Mendoza
2006

ÍNDICE

	Pág.
1. INTRODUCCIÓN…….……..…….....	1
2. ESTADO DE LA CUESTIÓN…….……..…….....	3
2.1. Introducción…….……..…….....	
2.2. Redes bayesianas…….……..……..	4
2.2.1. Marco teórico…….……..…….....	
2.2.1.1. Definiciones…….……..…….....	
2.2.1.2. Representación de una red bayesiana……..	8
2.2.1.3. Independencia condicional…….….....	10
2.2.1.4. Propagación de probabilidades (inferencia) ……..……	11
2.2.1.5. Aprendizaje para una red bayesiana……..……..	13
2.2.2. Aprendizaje de clasificaciones. Clasificador Naïve Bayes…….	15
2.2.2.1. Hipótesis a posteriori máxima …….	
2.2.2.2. Clasificador Naïve Bayes (Friedman et al., 1997) ……..…….	16
2.2.3. Aprendizaje de asociaciones - Algoritmo K2……...	20
2.2.3.1. Algoritmo K2 ……..…….	
2.3. El procedimiento de descripción estadística……..……...	22
2.3.1. Validación estadística……..	
2.3.2. Valor de prueba estadística…….....	
2.3.2.1. Valor Test (V-test) …….....	24
2.3.2.1.1. Valor Test – Características Cuantitativas Continuas……..	
2.3.2.1.2. Valor Test – Características Cualitativas Nominales…….…....	27
2.3.2.2. Criterios de significación estadística……..…….	32
2.3.3. Técnica para la descripción de variables cualitativas categóricas – DEMOOD……....	
3. HIPÓTESIS DE TRABAJO…….	35
3.1. Demod, clasificador bayesiano…….……..	
3.2. Puntos débiles del enfoque bayesiano.	42

……..
3.3. Naïve Bayes, validación de nuevas clasificaciones……..... 45
3.4. Red Bayesiana K2 – Asociación entre variables, obtención y validación de hipótesis..... 47
4. PROPUESTA……. 51
4.1. Clasificador bayesiano. Probabilidades, gráfico…….…….... 53
4.2. Descripción óptima del modelo……... 61
4.3. Predicción de la clase para nuevos casos y nivel de significación de la inferencia……
5. CONCLUSIONES……. 62
6. FUTURAS LÍNEAS DE INVESTIGACIÓN……..... 64
6.1. Nivel de significación estadística para las Reglas…….....
6.2. Nivel de significación estadística para Árboles de Regresión. ……..... 66
7. BIBLIOGRAFÍA……... 67
8. ANEXOS…… 70

1. INTRODUCCIÓN

Las redes Bayesianas juegan diversos papeles importantes dentro de la Inteligencia Artificial. Uno de ellos es su actuación dentro del manejo de incertidumbre en los sistemas expertos. Otro papel importante lo tienen en lo que se conoce como *descubrimiento de conocimiento* en bases de datos; las redes Bayesianas permiten encontrar, de una manera consistente, *relaciones probabilistas entre variables*.

Las Redes Bayesianas (RBs) son un *formalismo* que en los últimos años ha demostrado su potencialidad como *modelo de representación de conocimiento* con incertidumbre. [Hernández O.J. 2004]. El hecho de utilizar una representación gráfica para la explicación del modelo hace de las RBs sean una herramienta realmente muy atractiva en su uso, como representación del conocimiento. No sólo modelan de forma cualitativa el conocimiento sino que además expresan de forma numérica la *fuerza de las relaciones entre las variables*. Esta parte cuantitativa del modelo suele especificarse mediante distribuciones de probabilidad como una medida de la creencia que tenemos sobre las relaciones entre variables de modelo.

Siendo entonces las redes Bayesianas modelos que describen las relaciones (*relaciones de independencia/dependencia*) entre variables, estas pueden ser aplicadas a casi cualquier tipo de problema. Se trata de utilizar las RBs para realizar *procesos eficientes de razonamiento* una vez especificado el modelo completo. Es decir, podemos realizar procesos de inferencia a partir de este modelo, conociendo alguna evidencia de las variables pronosticar cómo se comportarán el resto.

Por lo tanto, una red bayesiana una vez construida constituye un dispositivo potente para el *razonamiento probabilístico*. Sin embargo nos queda la tarea de *construcción de tal modelo*. Una posibilidad es que un experto, en el dominio que se quiere modelar, construya la red bayesiana a partir de su conocimiento en el problema. Debido al gran volumen de datos de los que habitualmente se dispone en dominios concretos, es de enorme interés proporcionarles a estos expertos herramientas que adquieran este tipo de conocimiento de forma automática a partir de datos de ejemplos del problema en cuestión, para que de esta manera tengan una herramienta de soporte para la decisión. [Hernández O.J. 2004].

Un aspecto importante en la construcción de la red, es la de calcular una *medida de adecuación de cada red a los datos de partida* y por consiguiente poder comparar para quedarnos con la mejor estructura, entre distintas redes bayesianas. Existen muchos tipos de medidas de calidad para calcular la adecuación de una red bayesiana a un conjunto de datos.

Ahora bien, otro aspecto importante en el aprendizaje es *medir la calidad* no de cómo fue estructurada la red a partir de los datos, sino de *con qué se determina esta estructura*; es decir *evaluar la significación (validar) de los datos de entrenamiento independientemente del tipo o estructura de la red bayesiana*. Esto permitirá mejorar el proceso de inferencia, simplificándolo, ya que sólo se tomarán aquellas relaciones que son realmente significativas estadísticamente.

Hay dos aplicaciones importantes de las RBs: para *aprender a clasificar* y para *aprender asociaciones*. Como clasificador podemos representar este modelo con una estructura fija como la del Naïve Bayes (un nodo padre clase y tanto nodos hijos como características) o usar una estructura más compleja como la de K2 (independientemente de si hay un clasificador se visualiza las relaciones entre variables). Este último, tiene más uso no tanto como clasificador sino para analizar las asociaciones entre variables, para de esta manera obtener hipótesis del modelo representado, que una vez validadas serán la base para la toma decisiones.

Entre los procedimientos de la estadística mutlivarada, se encuentra una técnica denominada DEMOD[1], que permite caracterizar variables cualitativas en función de otras (cualitativas y cuantitativas). En este trabajo veremos que esta técnica además de mostrarme las relaciones probabilistas entre una variable de clase y otras descriptivas, (expresan de forma numérica la fuerza de las relaciones entre las variables, igual que las RBs); también *ordenan* este conocimiento en función de un indicador estadístico llamado "Valor de Test", que determina el *nivel de significación o certeza* de la relación entre las variables, simplificando el modelo en sólo lo que es relevante.

Por lo tanto en este trabajo, se propone una *métrica estadística* que permita por un lado validar los datos *con que* se arma la estructura del grafo independientemente del tipo de RB. Por otro lado, validar las hipótesis que se infieren sobre una red bayesiana (relaciones de dependencia e independencia).

Concretamente buscar un valor de prueba que determina la significación estadística de las relaciones representadas en la red. De este modo si utilizamos la red para clasificación, podremos validar estadísticamente el modelo de predicción, por otro lado, en caso de que utilicemos la red para abducir nuevo conocimiento (que servirá de base para la toma de decisiones), podremos validar estadísticamente este conocimiento (hipótesis). La propuesta es complementar los métodos bayesianos con las técnicas de descripción estadística mencionadas.

Con este fin se ha estructurado este trabajo en 6 secciones, siendo l, esta introducción:

En la sección 2 se establecerán los conceptos teóricos fundamentales tanto de las RBs cómo de los procedimientos de descripción estadística y los métodos de validación y prueba estadística. En la sección 3, se mostraran 4 hipótesis de trabajo, que relacionan los conceptos teóricos de la herramienta inteligente y de técnica de inferencia de la estadística clásica. En la sección 4, mediante un caso práctico se efectuará la transferencia conceptual. Las secciones 5 y 6, resumen las conclusiones y líneas de investigación futuras.

[1] Técnica que en conjunto con el "cartografiado de datos" se utiliza para el "diagnóstico por imagen de datos ", en la investigación profesional (Ricardo y Matilde Césari – 1998-2006)

2. ESTADO DE LA CUESTIÓN

2.1. Introducción

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, 1995; Díaz & Corchado, 1999] y pueden manejar bases de datos incompletas [Heckerman, 1995; Heckerman & Chickering, 1996; Ramoni & Sebastiani, 1996].

Las RBs representan el conocimiento cualitativo del modelo mediante un grafo dirigido acíclico. Este conocimiento se articula en la definición de relaciones de independencia/dependencia entre las variables que componen el modelo. Estas relaciones abarcan desde una independencia completa hasta una dependencia funcional entre variables del modelo. El hecho de utilizar una representación gráfica para la explicación del modelo hace de las RBs una herramienta realmente muy atractiva en su uso como representación del conocimiento. No sólo modelan de forma cualitativa el conocimiento sino que además expresan de forma numérica la fuerza de las relaciones entre las variables. Esta parte cuantitativa del modelo suele especificarse mediante distribuciones de probabilidad como una medida de la creencia que tenemos sobre las relaciones entre variables de modelo.

Un aspecto importante en el aprendizaje 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. [Cowell, 1990; Ramoni & Sebastiani, 1999].

Por lo tanto es substancial en la construcción de la red, *calcular una medida de adecuación de cada red a los datos de partida* y por consiguiente poder comparar para quedarnos con la mejor estructura, entre distintas redes bayesianas. Existen muchos tipos de medidas de calidad para calcular la adecuación de una red bayesiana a un conjunto de datos.

Pero otro aspecto importante en el aprendizaje es medir la calidad **no de cómo** fue construida la red a partir de los datos, *evaluando la representación del conocimiento en la red*, sino medir la calidad **de con qué** se determina esta estructura de una red; es decir validad los datos de entrenamiento independientemente del tipo o estructura de la red bayesiana en que se representan. Los métodos estadísticos multivaridos me proveen de técnicas que son la clave para resolver este último aspecto.

A continuación se sintetizan los principales conceptos de redes bayesianas, funcionamiento de los algoritmos Naïve Bayes y K2; y conceptos fundamentales sobre validación estadística, valor de prueba y niveles de significación; además de una explicación del algoritmo Demod utilizado para la caracterización de variables cualitativas.

2.2. Redes bayesianas

2.2.1. Marco teórico

Las redes bayesianas o probabilísticas se fundamentan en la teoría de la probabilidad y combinan la potencia del teorema de Bayes con la expresividad semántica de los grafos dirigidos; las mismas permiten representar un modelo causal por medio de una representación gráfica de las independencias / dependencias entre las variables que forman parte del dominio de aplicación [Pearl, 1988].

Se puede interpretar a una red bayesiana de dos formas:

1. Distribución de probabilidad: Representa la distribución de la probabilidad conjunta de las variables representadas en la red.
2. Base de reglas: Cada arco representa un conjunto de reglas que asocian a las variables involucradas. Dichas reglas están cuantificadas por las probabilidades respectivas.

A continuación se describirán los fundamentos teóricos de las redes bayesianas y distintos algoritmos de propagación.

Una **red bayesiana** es un **grafo conexo acíclico dirigido**, donde las uniones entre los nodos tienen definidas una *dirección*, en el que los **nodos** representan *variables aleatorias* que pueden ser continuas o discretas; y las **flechas (arcos)** representan *influencias causales*, el que un nodo sea padre de otro implica que es causa directa del mismo.

Los estados que puede tener una variable deben cumplir con dos propiedades:

- Ser *mutuamente excluyentes*, es decir, un nodo sólo puede encontrarse en uno de sus estados en un momento dado.
- Ser un conjunto *exhaustivo*, es decir, un nodo no puede tener ningún valor fuera de ese conjunto

2.2.1.1. Definiciones:

- Un **nodo X** es una *variable aleatoria* que puede tener *varios estados* $\mathbf{x_i}$.

  La *probabilidad* de que el nodo **X** este en el estado **x** se denotará como **P(x) = P(X = x)**

- Un **arco** es la *unión entre dos nodos* y representa la dependencia entre dos variables del modelo.

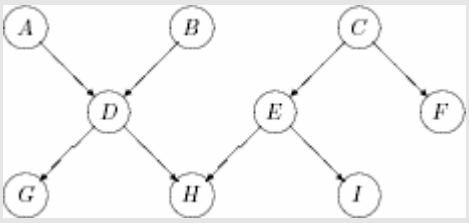
  Un arco queda definido por un *par ordenado de nodos* **(X, Y).** Esta definición de arco corresponde a lo que en otros lugares se denomina arco *dirigido*.

  En la representación gráfica, un arco **(X, Y)** viene dado por una flecha desde **X** hasta **Y**.

- **Grafo dirigido**. Es un par **G = (N, A)** donde **N** es un *conjunto de nodos* y **A** un *conjunto de arcos* definidos sobre los nodos.

- El nodo **X** es un **padre** del nodo **Y**, si existe un arco **(X, Y)** entre los dos nodos.

- El nodo **Y** es un **hijo** del nodo **X**, si existe un arco **(X ,Y)** entre los dos nodos
- **Antepasado**. **X** es un antepasado de **Z** si y sólo si existe (al menos) un nodo **Y** tal que **X** es padre de **Y** e **Y** es antepasado de **Z**.
- **Descendiente**. **Z** es un descendiente de **X** si y sólo si **X** es un antepasado de **Z**.
- **Familia X**. Es el conjunto formado por **X** y los padres de **X**, **pa(X)**.
- **Nodo terminal**. Es el nodo que no tiene hijos.



Los padres de D son A y B: pa (D) = {A, B}.
Los hijos de D son G y H.
Los antepasados de G son A, B y D.
Los descendientes de A son D, G y H.
Las nueve familias (tantas como nodos) son {A}, {B}, {C}, {D, A, B}, {E, C }, {F, C }, {G, D}, {H, D, E} e {I , E}.

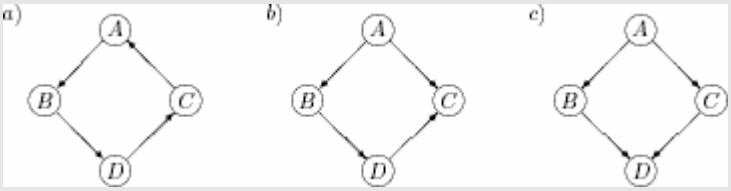
Ejemplo 2.2.1.1. Un pequeño poliárbol

- **Camino**. Un camino entre $\mathbf{X_1}$ y $\mathbf{X_N}$ en una *sucesión de nodos* $\mathbf{\{X_1, \ldots, X_N\}}$ pertenecientes a un grafo $\mathbf{G = (N, A)}$, tal que $\mathbf{X_i = X_j}$ para $\mathbf{1 \le i < j \le N}$ y $\mathbf{(X_i; X_{i+1}) \in A}$ ó $\mathbf{(X_{i+1}; X_i) \in A}$; $\mathbf{\forall i;\ 1 \le i < N}$. Es decir, dos nodos consecutivos de un camino ($\mathbf{X_i}$ y $\mathbf{X_{i+1}}$) están unidos por un arco del primero al segundo o viceversa.
- **Grafo acíclico**. Es el grafo en que *no hay ciclos*.

Tanto el *ciclo* como el *bucle* corresponden a lo que a veces se denominan *caminos cerrados simples*. La diferencia es que en un ciclo los arcos van de cada nodo al siguiente (nunca a la inversa), mientras que la definición de bucle permite que los arcos tengan cualquiera de los dos sentidos, con la única condición de que no formen un ciclo. La distinción entre ambos es muy importante, pues las *redes bayesianas se definen a partir de los grafos dirigidos acíclico*, lo cual permite que contengan *bucles* pero *no* que contengan *ciclos*.

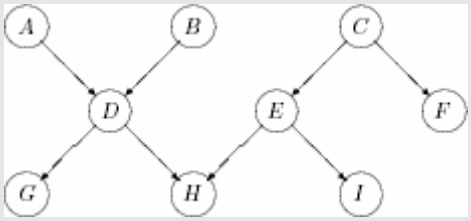
Ciclo. Es una sucesión de nodos $\mathbf{\{X_1, \ldots, X_N\}}$ pertenecientes a un grafo $\mathbf{G = (N , A)}$, tal que $\mathbf{X_i = X_j}$ para $\mathbf{1 \le i < j \le N}$, para todo $\mathbf{i < N}$ existe en **A** un arco $\mathbf{(X_i , X_{i+1})}$, y existe además un arco $\mathbf{(X_N , X_1)}$.

Bucle. Sucesión de nodos $\mathbf{\{X_1, \ldots, X_N\}}$ pertenecientes a un grafo $\mathbf{G = (N , A)}$, tal que $\mathbf{X_i = X_j}$ para $\mathbf{1 \le i < j \le N}$, para todo $\mathbf{i < N}$ existe en **A** un arco $\mathbf{(X_i , X_{i+1})}$ ó $\mathbf{(X_{i+1} , X_i)}$, existe además un arco $\mathbf{(X_N , X_1)}$ ó $\mathbf{(X_1, X_N)}$ y los arcos no forman un ciclo.



Ejemplo 2.2.1.2. En la figura **a**, vemos que entre **B** y **C** hay dos caminos: **{B, A, C}** y **{B, D, C }**, y lo mismo ocurre en **b** y **c**. El primero de estos tres grafos es un ciclo, mientras que los dos últimos son bucles. Por eso estos dos últimos podrían servir para definir redes bayesianas, pero el primero no.

- **Grafo conexo**. Un grafo es *conexo* si *entre dos cualquiera de sus nodos hay al menos un camino.* Por tanto, un grafo no conexo es aquél que *está formado por dos o más partes inconexas* entre sí. Todo grafo conexo ha de pertenecer a una de las dos categorías siguientes
  - *Grafo simplemente conexo o poliárbol*. Un grafo es *simplemente conexo* si entre dos cualquiera de sus nodos hay exactamente un camino
  - *Grafo múltiplemente conexo*. Es el que contiene ciclos o bucles
  - *Árbol*. Es un caso particular de poliárbol, en que *cada nodo tiene un sólo padre*, excepto el *nodo raíz*, que no tiene padres.



Ejemplo 2.2.1.3. Es un *poliárbol*, porque no contiene bucles; no es un árbol porque algunos de sus nodos (D y H) tienen más de un padre.

- **Probabilidad conjunta**:

Dado un *conjunto de variables* **{X ,Y, Z}** , la probabilidad conjunta *especifica la probabilidad de cada combinación posible de estados de cada variable* $\mathbf{P(x_i,y_j,\ldots,z_k)}$ $\mathbf{\forall i,j,\ldots k}$ de manera que se cumple que:

(Ecuación 2.2.1.1)

$$\sum_{i,j,\ldots k} P(x_i, y_j, \ldots, z_k) = 1$$

Probabilidad marginal es la *probabilidad particular de una de las variables*

$$P(x_i) = \sum_j P(x_i, y_j)$$

(Ecuación 2.2.1.2)

P(A,B)	b_1	b_2	P(A)
a_1	0.16	0.24	**0.4**
a_2	0.12	0.28	**0.4**
a_3	0.12	0.08	**0.2**
P(B)	**0.4**	**0.6**	**1**

Probabilidad marginal P(A), P(B)

$P(a_i) = \Sigma j\ P(a_i, b_j) P(a_1) = P(a_1,b_1) + P(a_1,b_2) = 0.16 + 0.24 = 0.4$

$P(b_j) = \Sigma i\ P(a_i, b_j)$

Probabilidad conjunta $1 = \Sigma i,j\ P(a_i, b_j)$

Ejemplo 2.2.1.4. Probabilidad conjunta y marginal

- **Probabilidad condicional**: Dadas dos variables **X** e **Y**, la probabilidad de que ocurra $\mathbf{y_j}$ dado que ocurrió el evento $\mathbf{x_i}$, es la probabilidad condicional de **Y** dado **X**, y se denota como: $\mathbf{P(y_j \mid x_i)}$.

La probabilidad condicional por definición es:

$$P(y_j \mid x_i) = \frac{P(y_j, x_i)}{P(x_i)}, \text{ dado } P(x_i) > 0$$

Análogamente, si se intercambia el orden de las variables

$$P(x_i \mid y_j) = \frac{P(y_j, x_i)}{P(y_j)}$$

A partir de las dos fórmulas anteriores se obtiene

$$\mathbf{P(y_j \mid x_i) = \frac{P(y_j) \cdot P(x_i \mid y_j)}{P(x_i)}}$$

(Ecuación 2.2.1.3)

Esta expresión se conoce como el *Teorema de Bayes*,

que en su forma más general es:

$$P(y_j \mid x_i) = \frac{P(y_j)P(x_i \mid y_j)}{\sum_j P(x_i \mid y_j)P(y_j)}$$

Al denominador se le conoce como el *Teorema de la Probabilidad Total*

En las redes bayesianas el *conjunto de valores que componen la probabilidad condicional* de un hijo dados sus padres, se representa en las llamadas *tablas de probabilidad condicional*

P(A | B)	b_1	b_2
a_1	0.4	0.6
a_2	0.3	0.7
a_3	0.6	0.4

$P(a_1|b_1) = P(a_1) \cdot P(b_1,a_1) / P(b_1) =$
$= 0.16 / 0.4 = 0.4$

Regla del producto: $P(A, B) = P(A|B) \cdot P(B)$

P(A,B)	b_1	b_2	P(A)
a_1	0.16	0.24	0.4
a_2	0.12	0.28	0.4
a_3	0.12	0.08	0.2
P(B)	0.4	0.6	1

Regla del producto condicionada: $P(A, B \mid C) = P(A \mid B, C) \cdot P(B \mid C)$

Ejemplo 2.2.1.5. Probabilidad condicional

Expresamos el teorema de **Bayes en forma Normalizada**:

$$\mathbf{P(y_j \mid x_i) = \alpha \cdot P(y_j) \cdot \lambda x_i(y_j)}$$

(Ecuación 2.2.1.4)

donde

$$\mathbf{\lambda x_i(y_j) = P(x_i \mid y_j)}$$ (Ecuación 2.2.1.5)

$$\mathbf{\alpha = [P(x_i)]^{-1}}$$ (Ecuación 2.2.1.6)

Dos variables **X** e **Y** son *independientes*, si la ocurrencia de una no tiene que ver con la ocurrencia de la otra.

Dos variables aleatorias **{X, Y}** son independientes si su probabilidad conjunta es igual al producto de las marginales, esto es:

$$P(x_i, y_j) = P(x_i)P(y_j), \ \forall(i.j)$$

esto implica que:

$$P(y_j \mid x_i) = P(y_j) \forall i,$$
$$P(x_i \mid y_j) = P(x_i) \forall i,$$

- La observación es la determinación del estado de un nodo (X □ x), a partir de un dato obtenido en el exterior del modelo.
- La **evidencia** es el *conjunto de observaciones*, $e = \{X = x, Y = y, \ldots, Z = z\}$, en un momento dado.
- **Probabilidad a priori**. Es la *probabilidad de una variable en ausencia de evidencia*.

Conociendo la *probabilidad a priori* de **X** y la *probabilidad condicional* $\mathbf{P(y_1|X)}$, podemos calcular la probabilidad a priori de $\mathbf{y_1}$ por el teorema de probabilidad total.

$$\mathbf{P(y_1) = \sum_x P(y_1 \mid x) \cdot P(x)}$$ (Ecuación 2.2.1.7)

- **Probabilidad a posteriori**. Es la *probabilidad de una variable condicionada a la existencia de una determinada evidencia*.

La probabilidad a posteriori de **X** cuando se dispone de la evidencia **e** se calcula como:

$$\mathbf{P^*(x) = P(x \mid e)}$$ (Ecuación 2.2.1.8)

Dada la evidencia $\mathbf{e=\{y_j\}}$

$$\mathbf{P^*(x) = (x \mid y) = \frac{P(x) \cdot P(y \mid x)}{P(y)}}$$ (Ecuación 2.2.1.9)

En forma Normalizada:

$$\mathbf{P^*(x) = \alpha \cdot P(x) \cdot \lambda y(x)}$$ (Ecuación 2.2.1.10)

donde

$$\mathbf{\lambda y(x) \equiv P(e \mid x) = P(y \mid x)}$$ (Ecuación 2.2.1.11)

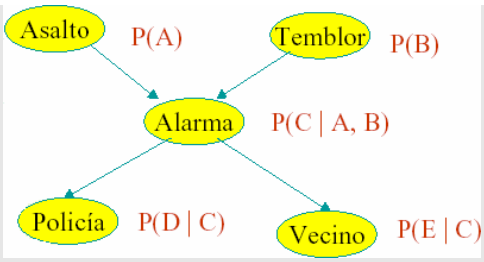
$$\mathbf{\alpha \equiv [P(e)]^{-1} = [P(y)]^{-1}}$$ (Ecuación 2.2.1.12)

2.2.1.2. Representación de una red bayesiana

Una red bayesiana representa relaciones causales en el dominio del conocimiento a través de una *estructura gráfica* y las *tablas de probabilidad condicional* entre los nodos.

El conocimiento que representa la red está compuesto por los siguientes elementos:

- Un *conjunto de nodos* $\mathbf{\{X_i\}}$ que *representan cada una de las variables del modelo*. Cada una de ellas tiene un conjunto exhaustivo de estados $\mathbf{\{x_i\}}$ mutuamente excluyentes.
- Un *conjunto de enlaces o arcos* $\mathbf{(X_i, X_j)}$ entre aquellos nodos que tienen una relación causal. De esta manera *todas las relaciones están explícitamente representadas* en el grafo.
- Una *tabla de probabilidad condicional* asociada a cada nodo $\mathbf{X_i}$ *indicando la probabilidad de sus estados para cada combinación de los estados de sus padres*. Si un nodo no tiene padres, se indican sus probabilidades a priori.



P(A=false) = 1- 0.001	P(A=true) = 0.001
P(B=false) = 1- 0.002	P(B=true) = 0.002
P(C=false | A=false, B=false) = 0.99	P(C=true | A=false, B=false) = 1- 0.99
P(C=false | A=false, B=true) = 0.71	P(C=true | A=false, B=true) = 1- 0.71
P(C=false | A=true, B=false) = 0.06	P(C=true | A=true, B=false) = 1- 0.06
P(C=false | A=true, B=true) = 0.05	P(C=true | A=true, B=true) = 1- 0.05
P(D=false | C=false) = 0.95	P(D=true | C=false) = 1- 0.95
P(D=false | C=true) = 0.1	P(D=true | C=true) = 1- 0.1
P(E=false | C=false) = 0.99	P(E=true | C=false) = 1-0.99
P(E=false | C=true) = 0.3	P(E=true | C=true) = 1-0.3

Ejemplo 2.2.1.6. Representación de una red bayesiana

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).

La *estructura* de una red bayesiana se puede determinar de la siguiente manera:

Se asigna un **vértice o nodo** a cada variable **(X_i)** y se indica de qué otros vértices es una causa directa; a ese conjunto de vértices "causa del nodo X_i" se lo denota como el conjunto $\boldsymbol{\pi}_X$ y se lo llamará "padres de X_i". Se une cada padre con sus hijos con **flechas** que parten de los padres y llegan a los hijos.

A cada variable **X** se le asigna una **matriz P ($\mathbf{x_i} \mid \boldsymbol{\pi}_{\mathbf{xi}}$)** que estima la *probabilidad condicional* de un evento $\mathbf{X_i = x_i}$ dada una combinación de valores de los $\boldsymbol{\pi}_{\mathbf{X}}$.

Una vez que se ha *diseñado la estructura de la red* y se han *especificado todas las tablas de probabilidad condicional* se está en condiciones de conocer la probabilidad de una determinada variable, dependiendo del estado de cualquier combinación del resto de variables de la red; para ello se debe *calcular la probabilidad a posteriori de cada variable condicionada a la evidencia*, estas probabilidades a posteriori se podrán obtener de forma inmediata a partir de la probabilidad conjunta de todas las variables $P(x_1, x_2, \ldots, x_i)$.

A continuación se indica cómo el **proceso** se ve simplificado al aplicar la propiedad de independencia condicional, que permite *obtener la probabilidad conjunta a partir de las probabilidades condicionales de cada nodo en función de sus padres.*

2.2.1.3. Independencia condicional

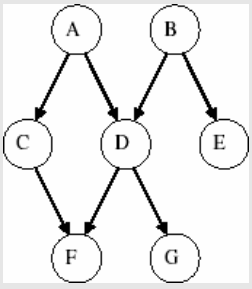
Una variable **X** es condicionalmente independiente de otra **Y** dada una tercer variable **Z**, si el conocer **Z** hace que **X** e **Y** sean independientes. Es decir, si conozco **Z**, **Y** no tiene influencia en **X**.

Esto es: **P(x | y, z) = P(x | z)** siempre que **P(y, z) > 0.** Conocer **Y** no provee información adicional acerca de **X**, una vez que conocemos a **Z**.

Alternativamente: **P(x, y | z) = P(x | z)·P(y | z)** (Ecuación 2.2.1.13)

Cada variable es independiente de todos aquellos nodos que no son sus "descendientes" una vez que se conocen sus propios nodos padres.

Gráficamente se verifica en los casos en que los nodos **X** e **Y** están separados por **Z** en el grafo. Esto implica que todos los caminos para ir de **X** a **Y** pasarán necesariamente por **Z**.



Ejemplo 2.2.1.7. {E} es condicionalmente independiente de [A, C, D, F, G}dado {B}; con lo cual P(E | A,B,C,D,F,G) = P(E | B); esto se conoce como Separación-D.

El término *direccional* hace referencia a la *asimetría de dicha propiedad*, que se manifiesta en las siguientes propiedades de las redes bayesianas

1. Si **A** no tiene padres, entonces **P (x|pa(x)) = P (x|∅) = P (x),** y la ecuación (3.59) se traduce en **P (e**|a) **= P (e)** para cada nodo **E** que no sea uno de los descendientes de **A**; en otras palabras, **E** es a priori independiente de **A**. En consecuencia, dos nodos cualesquiera **D** y **E** que no tengan ningún antepasado común son independientes a priori.
2. Si **D** es descendiente de **A** y antepasado de **H** , y no existe ningún otro camino desde **A** hasta **H** , entonces estos dos nodos quedan condicionalmente separados por **D: P (h|d, a) = P (h|d)**
3. Si tanto **G** como **H** son hijos de **D** y no tienen ningún otro antepasado común, este último separa **G** y **H** , haciendo que sean condicionalmente independientes: **P (g|d, h) = P (g|d)**

En general, *la independencia (a priori o condicional)* de dos nodos –por ejemplo, **A** y **E**— se pierde al conocer el valor de cualquiera de sus descendientes comunes —**H** es descendiente tanto de **A** como de **E**— pues en este caso la propiedad de separación direccional ya no es aplicable.

La importancia de este teorema es que nos permite describir una red bayesiana a partir de la probabilidad condicionada de cada nodo, en vez de dar la distribución de probabilidad conjunta, que requerirá un número de parámetros exponencial en el número de nodos y plantearía el grave problema de verificar la propiedad de separación direccional; sin embargo, el número de parámetros requerido para dar las probabilidades condicionadas es proporcional al número de nodos (suponiendo que el número de padres y el número de valores posibles están acotados para cada variable).

2.2.1.4. Propagación de probabilidades (inferencia).

Dado que se conoce el valor de alguna(s) variable(s) podemos *actualizar las probabilidades* del resto de las variables; esto comúnmente se llama *propagación de probabilidades*, *propagación de evidencia o inferencia*.

La *inferencia* es el proceso de *introducir nuevas observaciones y calcular las nuevas probabilidades que tendrán el resto de las variables*, por lo tanto, dicho proceso consiste en calcular las probabilidades a posteriori $\mathbf{P(X|Y)=y_i}$ de un conjunto de variables **X**, después de obtener un conjunto de observaciones $\mathbf{Y=y_i}$ donde **Y** es la lista de variables observadas e $\mathbf{y_i}$ es la lista correspondiente a los valores observados).

El fundamento matemático en el que se basan las redes probabilísticas para llevar a cabo la inferencia es el *Teorema de Bayes*, (ecuación 2.2.1.3) se expresa como:

$$P(y_j \mid x_i) = \frac{P(y_j)P(x_i \mid y_j)}{P(x_i)} = \frac{P(y_j)P(x_i \mid y_j)}{\sum_j P(x_i \mid y_j)P(y_j)}$$

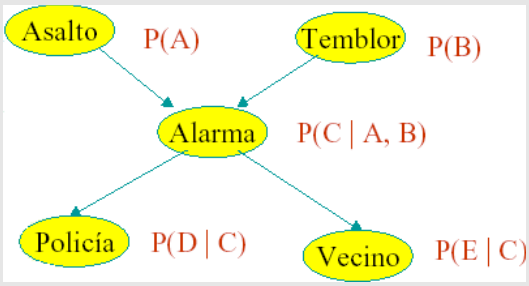
Según el teorema Bayes (Ecuación 2.2.1.3):

Dada $\mathbf{e = \{y_1, y_2\}} \rightarrow \mathbf{P^*(x) = (x \mid y_1, y_2) = \frac{P(x) \cdot P(y_1, y_2 \mid x)}{P(y_1, y_2)}} \equiv \mathbf{P^*(x) = \alpha \cdot P(x) \cdot \lambda y(x)}$

Ahora si aplicamos la independencia condicional (Ecuación 2.2.1.4) **podemos averiguar**

$$\boxed{\mathbf{P(y_1, y_2 \mid x) = P(y_1 \mid x) \cdot P(y_2 \mid x)}} \equiv \mathbf{\lambda y(x)} = \boxed{\mathbf{\lambda y_1(x) \cdot \lambda y_2(x)}}$$

Las probabilidades a posteriori $\mathbf{P(X|Y=y_i)}$, se pueden obtener a partir de la probabilidad marginal **P(X|Y)** , que a su vez puede obtenerse de la probabilidad conjunta $\mathbf{P(x_1, x_2, x_i)}$ sumando los valores para todas las variables que no pertenezcan al conjunto $\mathbf{X \cup Y}$. En la práctica, esto no es viable por el tiempo necesario para llevarlo a cabo, ya que incrementar el número de nodos de la red aumentaría exponencialmente el número de sumas necesarias; por este motivo se han desarrollado varios algoritmos de propagación que se citan a continuación.



P(A=false) = 1- 0.001
P(B=false) = 1- 0.002

P(C=false | A=false, B=false) = 0.99
P(C=false | A=false, B=true) = 0.71
P(C=false | A=true, B=false) = 0.06
P(C=false | A=true, B=true) = 0.05

P(D=false | C=false) = 0.95
P(D=false | C=true) = 0.1

P(E=false | C=false) = 0.99
P(E=false | C=true) = 0.3
P(A=true) = 0.001
P(B=true) = 0.002

P(C=true | A=false, B=false) = 1- 0.99
P(C=true | A=false, B=true) = 1- 0.71
P(C=true | A=true, B=false) = 1- 0.06
P(C=true | A=true, B=true) = 1- 0.05

P(D=true | C=false) = 1- 0.95
P(D=true | C=true) = 1- 0.1

P(E=true | C=false) = 1-0.99
P(E=true | C=true) = 1- 0.3

P(A,B,C,D,E) =
P(A)·P(B)·P(C|A,B)·P(D|C)·P(E|C)=
= 0.001 * 0.002 * (1-0.05) * (1- 0.1) * (1- 0.3)
P(D|E) = P(D|E,C)·P(C|E)+P(D|E,¬C)·P(¬C|E)

Ejemplo 2.2.1.8. Propagación de probabilidades

Algoritmos de propagación

Existen varios *métodos computacionales que aprovechan la estructura gráfica para propagar los efectos que las observaciones del mundo real tienen sobre el resto de las variables de la red*; las diferencias entre ellos se basan principalmente en la *precisión de los resultados* y en *el consumo de recursos* durante el tiempo de ejecución.

Los algoritmos de propagación se dividen inicialmente en "**exactos**" o "**aproximados**", según cómo *calculen los valores de las probabilidades*. Los métodos exactos *calculan los valores por medio del teorema de Bayes*, mientras que los métodos aproximados *utilizan técnicas iterativas de muestreo*, en las que los valores se aproximarán más o menos a los exactos dependiendo del punto en que se detenga el proceso.

Los algoritmos de propagación dependen del tipo de estructura de la red bayesiana, existiendo las siguientes tres **topologías de red**: *Árboles*, *Poliárboles* , y *Redes multiconectadas*

2.2.1.5. Aprendizaje para una red bayesiana

El término "*aprendizaje*" es una de las características de los sistemas adaptativos que son capaces de mejorar su comportamiento en función de su experiencia pasada, por ejemplo al resolver problemas similares.

El **aprendizaje en la redes bayesianas** consiste en *definir la red probabilística a partir de datos almacenados en bases de datos en lugar de obtener el conocimiento del experto*. Este tipo de aprendizaje ofrece la posibilidad de *inducir la estructura gráfica de la red* a partir de los datos observados y de *definir las relaciones entre los nodos* basándose también en dichos casos. El aprendizaje inductivo consiste en obtener conocimiento a partir de datos. En redes bayesianas se divide en 2 aspectos:

- Obtener la *estructura de la red* – estructural
- Obtener las *probabilidades asociadas* - paramétrico

Entonces 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.

Básicamente existen tres enfoques para de determinar la topología de una red Bayesiana: de forma manual o tradicional, de forma automática y el enfoque Bayesiano que puede ser visto como una combinación de los dos anteriores

Aprendizaje a Partir de Datos

Este enfoque ha sido el más explorado durante los últimos años, y existe una gran variedad de algoritmos para la obtención de la estructura de la red bayesiana a partir de datos.

La motivación de este enfoque surge, obviamente, para evitar el enfoque tradicional en el que se extraía el conocimiento del experto. Con este enfoque el tiempo de ingeniería del conocimiento se reduce considerablemente, al determinar la estructura de manera automática. El aprendizaje de redes bayesianas a partir de datos se divide en dos: métodos basados en búsquedas y métodos basados en restricciones.

- *Algoritmos Basados en Búsquedas*

  La determinación de la estructura de la red es vista como un problema de *selección del modelo*, en este caso probabilista. Los estadísticos, que comúnmente trabajan con este tipo de problemas, usan dos enfoques para resolver este problema: *selección del modelo y selección del modelo promedio;* el primero consiste en seleccionar un solo modelo "bueno" de entre todos los posibles modelos, y usarlo como si fuera el modelo correcto, el segundo enfoque consiste en seleccionar un número manejable de modelos buenos de entre todos los modelos posibles y pretender que estos modelos son exhaustivos

Para la búsqueda del modelo, existen diferentes enfoques de optimización, como los algoritmos genéticos. Ahora bien, para saber que tan bueno es el modelo, se han desarrollado diferentes criterios para evaluar el modelo (en este caso la estructura de la red bayesiana aprendida), llamándose, en general, criterios de bondad de ajuste. Entre ellos figuran: método de calificación Bayesiano (Cooper & Herskovits, 1992; Heckerman, 1994; Ramoni y Sebastián, 1996), métodos basados en entropía (Herskovits, 1991), método de evaluación MDL (minimum description length) (Susuki, 1996; Lam and Bacchus, 1994; Bouckaert, 1994) y método de evaluación MML(minimum message length) (Wallace, 1996).

En general, el problema de la selección del modelo consiste en encontrar un modelo que, basado en datos, incluya una aproximación de la distribución de frecuencias relativas de dichos datos (i.e. distribución de probabilidad), pero que además sea un modelo de dimensiones razonables

La idea central del enfoque de búsquedas radica en encontrar el modelo mas parsimonioso que describa de mejor manera la distribución de probabilidad de los datos, basándose en algún criterio

Cuando se trata de pocas variables podemos realizar una búsqueda exhaustiva, es decir evaluar cada posible estructura y escoger la mejor, de acuerdo el criterio de evaluación. Sin embargo, cuando el número de variables no es pequeño se vuelve intratable, computacionalmente hablando, una búsqueda exhaustiva.

Uno de los principales problemas encontrados en este tipo de métodos es la cantidad tan grande de posibles estructuras; es por ello que se han desarrollado diferentes heurísticas para no explorar todo el espacio de búsqueda, entre ellos figuran los algoritmos genéticos (Larragaña) y algoritmos “greedy” como el K2 (Cooper & Herskovits, 1992).

- *Algoritmos Basados en Restricciones*

Este tipo de algoritmos asumenlo siguiente: dado un conjunto de independencias condicionales en una distribución de probabilidad, hay que encontrar el GAD que contenga todas y solamente estas independencias condicionales

Entonces, la clave de estos algoritmos es determinar las relaciones de independencia (marginal o condicional), contenidas implícitamente en los datos, con alguna medida de independencia condicional.

Una medida comúnmente usada para probar independencia marginal e independencia condicional es la información mutua y la información mutua condicional (Shannon & Weaver, 1949, Pearl, 1988), respectivamente.

Si solo contáramos con el ordenamiento, de las variables, entonces tendríamos que determinar de alguna forma el conjunto mínimo de variables que separa a cada variable de sus predecesores que no son padres.

La forma en que la mayoría algoritmos resuelven el problema de determinar el conjunto de padres es llevando a cabo muchas pruebas de Independencia condicional. El hecho de tener el ordenamiento de las variables como un conocimiento a priori, es visto como una desventaja en los algoritmos de construcción de redes bayesianas, ya que en dominios muy complejos donde hay muchas variables y sus relaciones no son muy claras ni para el experto humano, no es posible contar o determinar de manera adecuada el ordenamiento de las variables

Es por eso que algunos autores han ideado algoritmos más generales que prescindan de un ordenamiento. Solo por mencionar algunos, (Spirtes et al., 1990; Spirtes and Glymour, 1991; Martínez-Morales, 1995; Cheng, 1998). Existen algunos otros algoritmos que tampoco reciben un ordenamiento, pero casi siempre construyen un grafo acíclico parcialmente orientado. Ejemplo de este tipo es el Bayes9 (Cruz-Ramírez, 2001).

Estos dos enfoques para el aprendizaje de redes Bayesianas a partir de datos son los más usados, aunque se ha sugerido un enfoque híbrido, es decir, combinar algún algoritmo de restricciones con uno de búsquedas para tratar de aprovechar las ventajas de los dos enfoques y evadir las desventajas de los mismos que en breve se tratará.

2.2.2. Aprendizaje de clasificaciones. Clasificador Naïve Bayes

"Asignación de un ejemplo a una clase o categoría".

Un clasificador es una función que mapea un conjunto de casos (atributos) en una clase específica (Friedman 1997). La tarea de clasificar consiste en etiquetar casos a partir de un conjunto de características (Friedman 1997). Esta tarea ha sido abordada con diferentes enfoques, entre ellos podemos mencionar los árboles de decisión, redes neuronales y Naïve-Bayes (Friedman, 1997). Este último se considera bastante efectivo, en el sentido de que es competitivo con el estado del arte de los clasificadores.

La clasificación supervisada es una tarea básica dentro del análisis de datos y el reconocimiento de patrones que requiere de la construcción de un clasificador: función que asigna una clase a una instancia descrita por un conjunto de variables.

2.2.2.1. Hipótesis a posteriori máxima

Supongamos que tenemos un conjunto de hipótesis candidatas *H* y estamos interesados en encontrar la hipótesis $h \in H$ más probable dados los datos D observados. Cualquier hipótesis de máxima probabilidad se llama hipótesis a posteriori máxima (MAP, maximum a posteriori) [Mitchell, 1997].

Podemos usar el teorema de Bayes (ecuación 2.2.1.3) para calcular la probabilidad a posteriori de cada hipótesis candidata, y la que tenga mayor probabilidad será la hipótesis MAP, que denotaremos como h_{MAP}

$$h_{MAP} \equiv \arg\max_{h \in H} P(h \mid D) = \arg\max_{h \in H} \frac{P(D \mid h) P(h)}{P(D)} = \arg\max_{h \in H} P(D \mid h) P(h)$$ (Ecuación 2.2.2.1)

En algunos casos se asume que cada hipótesis en *H* tiene la misma probabilidad a priori (*P (hi) = P (h)*, para todo *hi* en *H*). En este caso podemos simplificar aun más la ecuación y solamente necesitamos considerar el término *P (D|h)* para encontrar la hipótesis más probable. *P (D|h)* es la probabilidad de los datos D dado h. Cualquier hipótesis que maximice *P(D|h)* se llama hipótesis de máxima verosimilitud(ML).

$$h_{ML} = \arg\max_{h \in H} P(D \mid h)$$

Lo anterior, Mitchell [Mitchell, 1997] lo presenta como una conexión entre el teorema de Bayes y los problemas de aprendizaje automático, refiriéndose a los datos *D* como una conjunto de ejemplos de entrenamiento y a *H* como un espacio de funciones que pueden describir los datos.

2.2.2.2. Clasificador Naïve Bayes (Friedman et al., 1997)

Modelo más simple de clasificación con redes bayesianas. La estructura de la red es fija y sólo necesitamos aprender los parámetros (probabilidades). Las hipótesis de independencia asumidas por el clasificador NB da lugar a un modelo gráfico probabilístico en el que existe un único nodo raíz (la clase) y en la que todos los atributos son nodos hojas que tienen como único padre a la variable clase.

El fundamento principal de clasificador Naïve Bayes (NB) [Langley, 1992] es la suposición de que todos *los atributos son independientes conocido el valor de la variable clase*. En este trabajo uso la implementación hecha de "NaiveBayesSimple", en el programa Weka [Weka, 2003].

Para inducir un clasificador a partir de una base de datos, se consideran dos tipos de variables: la variable clase o *clase* **C**, y el resto de variables o *predictoras*, $\mathbf{X = (X_1,\ldots, X_d, X_{d+1},\ldots, X_t)}$. Asumimos que $\mathbf{(X_1,\ldots, X_d)}$ es el conjunto de predictoras con valores numéricos continuos y $\mathbf{(X_{d+1},\ldots, X_l)}$ es el conjunto de predictoras discretas. Una red Naive Bayes se vería más o menos así:

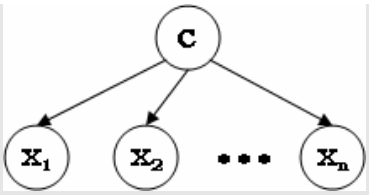


Figura 2.2.2. 1.: Estructura gráfica de un modelo Naïve Bayes.

El proceso de clasificación de una instancia **x** consiste en seleccionar la clase **c** con la máxima probabilidad a posteriori, **P(c|x)** [Pérez, Larrañaga, 2005].

Estimación de parámetros:

Sea **C** la variable aleatoria dependiente y sea $\mathbf{X = \{X_1, X_2, ..., X_n\}}$ el conjunto no ordenado de variables aleatorias independientes(atributos). Un clasificador Naïve asume que **C** es padre de todas las variables del conjunto **X** y a su vez estas variables son independientes entre si dada la variable dependiente **C**.

Para llevar a cabo la tarea de clasificación hacemos lo siguiente. Tomamos el caso que queremos clasificar, caso seguido instanciamos la red con dicho caso, es decir hacemos que cada variable independiente de la red tome el valor correspondiente del registro. Para saber a que valor mapea la red Naïve debemos *calcular la probabilidad de cada valor de la clase dado el conjunto de atributos*, de la forma siguiente usando la definición de la probabilidad condicional (Ecuación 2.2.1.3):

$$\mathbf{P(C \mid X) = \frac{P(C, X)}{P(X)}}$$ donde: **P(C | X)** es la *probabilidad conjunta* de **C** y **X**

Recordemos que esta probabilidad se encuentra codificada en la red, y es el producto de todas las probabilidades condicionales (Ecuación 2.2.1.13): $\mathbf{P(C, X)=P(C)\cdot P(X_1|C)\cdot P(X_2|C)\ldots P(X_n|C).}$

P (Y) es la *probabilidad del conjunto condicionante*. Esta probabilidad no la calcularemos, ya que se eliminará al normalizar las probabilidades calculadas. Esto es fácil de ver; la normalización que haremos consiste en dividir cada valor de probabilidad entre la suma de todas (todas las probabilidades condicionales de la clase).

El enfoque Bayesiano para clasificar este nuevo caso, consiste en asignar el valor con mayor probabilidad, lo cual corresponde a la hipótesis MAP, (Ecuación 2.2.2.1) $v_{MAP} = \arg\max_x P(x \mid y)$

MAPv es pues el valor clasificado por Naïve-Bayes (la etiqueta asignada) para el nuevo caso presentado y corresponde al valor con probabilidad máxima a posteriori (MAP). Debido a la *hipótesis de independencia* usada en el Naïve Bayes, la expresión para obtener la hipótesis MAP queda como sigue:

$$C_{MAP} = \underset{c \in \Omega_C}{\arg.\max} p(X_i,...,X_n \mid c)p(c) \qquad C_{MAP} = \underset{c \in \Omega_C}{\arg.\max} p(c)\prod_{i=1} p(X_i \mid c)$$

Es decir, la tabla de probabilidad **$P(X_1,\ldots, X_n \mid c)$** ha sido factorizada como el producto de **n** tablas que sólo involucra dos variables. Por tanto, los parámetros que tenemos que estimar son **$P(X_i \mid c)$** para cada atributo y la probabilidad a priori de la variable clase **P(c)**. [Hernández O.J., 2004]

- **Si el atributo X_i, es discreto**: la estimación de probabilidad condicional se basa en las *frecuencias de aparición que obtendremos en la tabla de datos*. Así si llamamos **$m(x_i, Pa(x_i))$** al número de registros de la tabla de datos en que la variable **X_i** toma el valor **x_i** y los padres de **x_i** ($Pa(x_i)$) toman las configuraciones denotada por **$Pa(x_i)$**, entonces la forma más simple de estimar **$P(x_i \mid Pa(x_i))$**, es:

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

El número de casos favorables divididos por el número de casos totales.

(Ecuación 2.2.2.2)

Esta técnica se conoce como *estimación por máxima verosimilitud* y tiene como desventajas que necesita una muestra de gran tamaño y que sobreajusta los datos. Existen otros estimadores más complejos que palian estos problemas, entre ellos el estimador basado en la ley de la *sucesión Laplace*:

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

El número de casos favorables más uno dividido por el número de casos totales más el número de valores posibles

(Ecuación 2.2.2.3)

Con pocos ejemplos la probabilidad se corrige por la probabilidad uniforme a priori, es decir, uno dividido por el número de valores posibles. Con esta estimación lo que se pretende es que todas las configuraciones posibles tengan una mínima probabilidad, ya que con el estimador de máxima verosimilitud cualquier configuración que no esté presente en la base de datos tendrá probabilidad cero.

- **Si el atributo X_i, es continuo**: el clasificador NB, supone que el atributo en cuestión sigue una distribución normal; por tanto, lo único que tenemos que calcular (a partir de la base de datos) es la media μ y la desviación típica σ condicionadas a cada valor de la variable clase. El inconveniente es que los datos no siempre siguen una distribución normal.

$$P(X_i \mid c) \propto N(\mu,\sigma) = \frac{1}{\sqrt{2\pi\cdot\sigma}} \exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right)$$ (Ecuación 2.2.2.4)

Veamos un ejemplo sencillo:

Supongamos un problema de clasificación con la variable clase **C** (tomando valores *+* y -) y dos atributos: **D** (discreto tomando valores *a* y *b*) y **N** (*numérico*). Sea el siguiente conjunto de tripletas [D, N, C] nuestra tabla de datos:

D	N	C
a	5	+
a	2.2	-
a	1.8	-
b	4	+
b	2	+
a	3	-

Para construir el clasificador NB tenemos que estimar **P(C)**, **P (D | C)** y **P(N | C)**:
Aplicando las ecuaciones 2.2.2.3/4 y 2.2.2.5, obtenemos las siguientes estimaciones:

P(C)	
+	0.5
-	0.5

P(D | C)	+	-
a	0.33	1
b	0.67	0
Máx. Veros.		

P(D | C)	+	-
a	0.4	0.8
b	0.6	0.2
Laplace		

P(N | C)	
+	N(μ=3.67, σ=1.53)
-	N(μ=2.33, σ=0.61)

Podemos ver como, en este caso la compensación de Laplace compensa el cero asignado a P (b | -), que puede deberse a la pequeñez de la muestra. Usando esta estimación para P (D | C), y usando la distribución normal para el atributo numérico, podríamos ahora clasificar un nuevo caso: p.e.(D=b, N=3.5):

P (+ | (b, 3.5))=P (+)·P(b | +)·N(3.64, 1.53:3.5)=0.5·0.6·0.26=0.078

P (- | (b, 3.5))=P (-)·P(b | -)·N(2.33, 0.61:3.5)=0.5·0.2·0.2=0.01

Y tras normalizar, obtnemos P (+ | (b, 3.5))=0.89 y P (- | (b, 3.5))=0.11, por lo que "+" es el clasificador para el caso tratado.

Ejemplo 2.2.2.1. Estimación de parámetros

El buen rendimiento del clasificador NB ha motivado la investigación de paradigmas basados en grafos que relajen la fuerte suposición de independencia. Uno de los primeros y que obtienen mejor rendimiento es la red Bayesiana aumentada a árbol (TAN, tree augmented Bayesian network). TAN construye un árbol de dependencias entre las predictoras que son a su vez hijas de la variable clase.

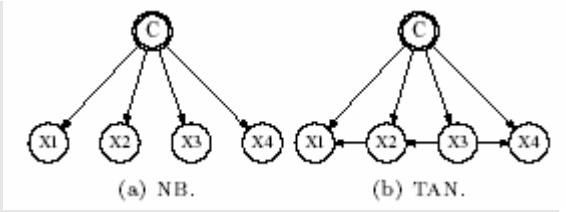


Figura 2.2.2. 2. Naïve Bayes y TAN

2.2.3. Aprendizaje de asociaciones - Algoritmo K2

Detectar asociaciones entre distintas características de los datos de entrenamiento. Generalización del problema de la clasificación: Se predice sobre cualquier valor de atributo, Puede predecirse más de un valor

El clasificador NB ha demostrado comportarse sorprendentemente bien en la clasificación supervisada a pesar de que asume que las variables predictoras son condicionalmente independientes dada la clase, lo que generalmente no se cumple. El clasificador red Bayesiana aumentada a árbol rompe con esta suposición tan fuerte ya que permite dependencias entre las variables predictoras, por lo que se comporta mejor que el NB en ciertos dominios. [Freidman, 1997]

También se usan las RBs como clasificadores. No se le da un tratamiento especial a la variable clase, aprendiendo así un modelo orientado a clasificación, sino que se aprende una red incluyendo a todas las variables (clases y atributos) del problema, que posteriormente se utilizará para clasificar. Permite identificar las variables que son útiles para la predicción, que están directamente la relacionada con la variable clase.

En el caso de una Red Bayesiana la clasificación se lleva a cabo de la misma manera, la única diferencia es que las redes Bayesianas *no consideran a las variables explicativas como condicionalmente independientes dada la clase*. En su lugar **si toma en cuenta estas posibles relaciones.** Otra diferencia importante es que una red Bayesiana puede sacar ventaja al prescindir de variables irrelevantes para la clase, tomando en cuenta para la clasificación aquellas que le son importantes.

En esta red bayesiana podemos *observar las relaciones directas e indirectas entre las variables* correspondientes al dominio utilizado. Además utilizando métodos de inferencia para el cálculo de las probabilidades a posteriori un experto podrá analizar los cambios de estas probabilidades conforme vayan introduciéndose nuevas evidencias.

2.2.3.1. Algoritmo K2

Este algoritmo fue desarrollado por Cooper y Herskovits (1992). Se trata de un algoritmo de búsquedas, muy rápido por cierto, que o*ptimiza la probabilidad de la red dada la base de datos*. En realidad lo que hace este algoritmo es encontrar el conjunto de padres más probables, utilizando la métrica Bayesiana, que mide precisamente la probabilidad de la estructura dados los datos. La heurística de este algoritmo se basa en un ordenamiento topológico que tiene que ser especificado por el usuario.

De esta forma se reduce considerablemente el espacio de búsqueda, por que el ordenamiento hace que un nodo que esta después en el ordenamiento que otro no pueda ser su padre. En este trabajo uso la implementación hecha de K2 en el programa Elvira [Elvira, 2000].

Se puede considerar el primer algoritmo basado en búsqueda u optimización de una métrica bayesiana y ha sido fuente de inspiración para posteriores trabajos sobre el tema. Este algoritmo utiliza un esquema voraz en su búsqueda de soluciones candidatas cada vez mejores y parte de que las variables de entrada están ordenadas, de forma que los posibles padres de una variable aparecen en el orden antes que ella misma. El proporcionarle al algoritmo un orden entre las variables hace que éste tan “sólo” tenga que buscar el mejor conjunto de padres posibles de entre las variables predecesoras en el orden. La búsqueda de este conjunto se hace de forma voraz.

El algoritmo parte de que el conjunto de padres para cada variable es el conjunto vació. Posteriormente, y siguiendo el orden establecido, pasa a procesar cada variable Xi, calcula la ganancia que se produce en la medida utilizada al introducir una variable Xj como padre de Xi de entre todas sus predecesoras, esto es, para todo j<i y se queda con la que produce la mejor ganancia. Desde un enfoque bayesiano, esta ganancia se calcula como la razón de la métrica de calidad de la red bayesiana donde se le introduce un segundo padre no insertado previamente.

2.3. El procedimiento de descripción estadística[2]

2.3.1. Validación estadística

En todas las ramas de la ciencia, cuando un investigador hace una afirmación con respecto a un fenómeno (que puede estar basado en su intuición, o en algún desarrollo teórico que parece demostrarla); debe luego probar la misma mediante el *método empírico*.

La experimentación consiste en "armar un ambiente de prueba en el que ocurre el fenómeno" (o buscarlo en el ambiente real) y tomar *mediciones de las variables* involucradas, luego realizar el *procesamiento de los datos* y *análisis estadístico* de los resultados y determinar si los mismos *confirman la afirmación* realizada en forma de hipótesis.

La esencia de probar una hipótesis estadística es "decidir" si la afirmación se encuentra apoyada por la evidencia empírica que se obtiene a través de una muestra.

El proceso de *validación estadística* se resume cómo:

1. Si extraemos una, muestra al azar, de tamaño n y deseamos estimar sus parámetros, media (μ) y varianza (σ^2), se plantea si los parámetros estimados serán próximos a algún valor hipotético.
2. Ante esto surge la probabilidad de formular una "hipótesis nula" ($\mathbf{H_0}$) y de manera sucesiva formar una "hipótesis alternativa" ($\mathbf{H_1}$).
3. Luego, los procedimientos que permiten aceptar o rechazar hipótesis, o determinar si una muestra "difiere significativamente" de otra o de los resultados esperados.; se denomina "prueba de hipótesis o de significación".

2.3.2. Valor de prueba estadística

El investigador, en el análisis de datos, siempre está obligado a deducir inferencias de una muestra y mostrar su significación estadística.

Para interpretar las *inferencias estadísticas* debe especificar los "**niveles aceptables de error**". El modo de aproximación más común es determinar el nivel de error de **tipo I**, también conocido como α (alfa).

		Realidad	
		H_0 (cierta)	H_0 (falsa)
Decisión Estadística	H_0 (Aceptar)	**1 - α** **Significación**	**β** **Error Tipo II**
	H_0 (no aceptar)	**α** **Error Tipo I**	**1 - β** **Potencia**

Figura 2.3.2.1. .Decisión Estadística

[2] Monografía protegida en la Direccional Nacional del Derecho del autor – expediente nº 044117, formulario 14317. Autores Ricardo y Matilde Césari.

El error **Tipo I** es la probabilidad de rechazar la hipótesis nula $\mathbf{H_0}$ cuando es cierta; es decir, la posibilidad de que la prueba muestre significación estadística cuando en realidad no este presente (se debe al azar las diferencias).

Especificando un nivel **α**, el investigador fija los márgenes admisibles de error de rechazar la Ho siendo cierta. Consideramos **(1 - α)**, es la probabilidad de aceptar correctamente la hipótesis nula cuando debe ser aceptada; es el test más usado y denominado "prueba de significación". Al especificar el nivel de error Tipo I, el investigador también determina un error asociado denominado el error de Tipo II o β (beta).

El error **Tipo II** es la probabilidad de fallo en rechazar la hipótesis nula $\mathbf{H_0}$ cuando es realmente falsa. Una probabilidad más interesante es (1-β), es denominada la "potencia" del *test de inferencia estadística*. La potencia (1-β) es la probabilidad de rechazar correctamente la hipótesis nula cuando debe ser rechazada.

Aunque la especificación α, establece el nivel de significación estadística (error aceptable), es el nivel de potencia el que dicta la probabilidad de "éxito" en la búsqueda de las diferencias si es que realmente existen.

Los errores de Tipo I y de Tipo II están inversamente relacionados y a medida que el error de Tipo I se hace más restrictivo (se acerca a cero), el error Tipo II aumenta. Al disminuir el error de Tipo I también se reduce el poder de la prueba estadística.

El investigador tiene que conseguir un equilibrio entre el nivel α y la potencia resultante.

Se han examinado la potencia [Hair, Anderson, Tathem y Blas, 1999] para la mayor parte de las pruebas de inferencia estadística y se ha proporcionado pautas para niveles aceptables de potencia, sugiriéndose que los estudios deben diseñarse para conseguir niveles de al menos **α=0,05** con niveles de potencia (1-β)=80%.

La prueba o test debe tratar de minimizar los errores pues para una muestra de un tamaño dado si se desea *disminuir un tipo de error se aumenta el otro.*

En la práctica se toma el error de Tipo I y se le llama *nivel de significancia* α (*o probabilidad de cometer un error de Tipo I, de rechazar una hipótesis cuando debería ser aceptada*).

Se toma usualmente α=0,05 (o 0,01), o sea hay 5 (o 1) oportunidades en 100 de rechazar la hipótesis, cuando debería ser aceptada o 95% (o 99%) de confianza de que se toma la decisión adecuada".

Al **valor de Prueba Crítico** se lo llama también: *Valor Test* **(V-test).** El "valor test" es el criterio que evalúa estadísticamente la desviación entre la medida sobre el grupo y la medida sobre la población.

"*Evalúa de alguna manera la distancia entre la media general y la media en el grupo, en número de desviaciones tipo de una ley normal*".

2.3.2.1. Valor Test (V-test)

Es un valor de prueba estadístico que permite la clasificación y ordenación de los elementos característicos de una muestra [A. MORINEAUU, 1984]. Constituye una herramienta que se utiliza en la *aproximación exploratoria y descriptiva de grandes matrices de datos*. El paquete informático SDPAD [SDPAD, 2002], consagrado al tratamiento estadístico de grandes matrices de datos, hace un amplio uso de este concepto.

El *valor de test* evalúa el interés de un elemento para caracterizar una categoría de individuos a partir de un estadístico calculado sobre la muestra. Los elementos característicos se evalúan por V-test decrecientes, es decir por orden de interés decreciente. Cuanto mayor sea el V-test (y superior al valor umbral usual de 2 desviaciones típicas), mejor caracterizará significativamente el elemento la categoría de casos. [Césari, 2005].

2.3.2.1.1. Valor Test – Características Cuantitativas Continuas

1. *Variables continuas*, consideremos la utilización de *la media* para determinar las variables continuas que caracterizan a una clase de grupos de observaciones. "Si la desviación entre la media calculada en el grupo y su valor calculado sobre la población es atribuible al "azar", la variable no caracteriza el grupo. Si la media se desvía "significativamente" de la media general, se dirá que los individuos del grupo se caracterizan por esta variable".

Si consideramos que $\mathbf{n_{pk}}$ valores de la **p**-ésima variable continua, ha sido extraídos al azar sobre los **n** elementos de la población, la media *intra–clase* $\overline{\mathbf{X}}_{\mathbf{pk}}$ de la **p**-ésima variable (propia del grupo de elementos que presentaron la característica "**p**" de la **k**-ésima clase de la variable cualitativa), será muy similar a la media general $\overline{\mathbf{X}}_{\mathbf{p}}$ de la **p**-ésima variable continua sobre, los **n** elementos observados. La diferencia entre la media *intra-clase* y la media general será "tanto más importante" cuanto más inadmisible sea la hipótesis que los $\mathbf{n_{pk}}$ valores de la variable **p** continua extraídos al azar sobre los **n** elementos.

Para categorizar las relaciones existentes entre la **k**-ésima clase de la variable cualitativa y la **p**-ésima variable continua, podemos proceder de manera muy similar al caso de una prueba estadística clásica.

Si se plantea la siguiente hipótesis nula $\mathbf{H_0}$:

- $\mathbf{H_0}$: los $\mathbf{n_{pk}}$ valores de la **p**-ésima variable continua SI han sido extraídos al azar sobre los **n** elementos de la población.

  Es decir: $\mathbf{H_0 : \mu = \overline{X}_{pk} = \overline{X}_p}$

- $\mathbf{H_1}$: los $\mathbf{n_{pk}}$ valores de la **p**-ésima variable continua NO han sido extraídos al azar sobre los **n** elementos de la población.

  Es decir: $\mathbf{H_0 : \mu \neq \overline{X}_{pk} \neq \overline{X}_p}$

Si consideramos que esos valores constituyen una muestra aleatoria simple, sin restricción, de $\mathbf{n_{pk}}$ observaciones de **n** (puesto que cada elemento recibió una y sólo un valor de la **p**-ésima variable continua); podemos calcular un valor "medio de la muestra", que sea tan extremo como el observado en una clase "**k**" dada de elementos.

Llamamos $\boxed{\overline{\mathbf{X}}_\mathbf{p}}$ la *media general* de la **p**-ésima variable continua sobre los $\boxed{\mathbf{n}}$ elementos (media general observada)

Y llamamos $\boxed{\mathbf{S_p^2}}$, la *varianza* de esa variable **p**.

Además, $\boxed{\overline{\mathbf{X}}_{\mathbf{pk}}}$, es la *media intra-clase* de la **p**-ésima variable sobre los $\boxed{\mathbf{n_{pk}}}$ elementos que pertenecen a la **k**-ésima clase

Si repetimos un gran número de veces la operación de seleccionar aleatoriamente $\mathbf{n_{pk}}$ valores entre los **n** valores posibles de la **p**-ésima variable, observamos que esas diferentes muestras, presentan una gran variabilidad de valores de $\overline{\mathbf{X}}_{\mathbf{pk}}$ (medias intra-clase). Las series de esas "medias de muestra" constituyen una variable aleatoria $\mathbf{M_k}$, que se puede demostrar que es un estimador de $\overline{\mathbf{X}}_\mathbf{p}$. Cada una de las medidas de la muestra $\overline{\mathbf{X}}_{\mathbf{pk}}$ constituye una estimación de $\overline{\mathbf{X}}_\mathbf{p}$. Podemos demostrar, [Grosbras, 1987], la hipótesis que los valores de la clase que nos interesa caracterizar, han sido "seleccionados al azar".

La *media del estimador* de $\overline{\mathbf{X}}_\mathbf{p}$, tiende hacia $\mathbf{X : E_{H_0}(M_k) = \overline{X}_p}$.

La *varianza del estimador* de $\overline{\mathbf{X}}_\mathbf{p}$ la definimos como:

$$\mathbf{Var_{H_0}(M_k) = S_{pk}^2 = \left(\frac{S_p^2}{n_{pk}}\right)\cdot\left(\frac{n - n_{pk}}{n - 1}\right)} \quad \text{(Ecuación 2.3.2.1)}$$

En estas condiciones, la variable aleatoria $\mathbf{M_k}$, se ajusta a una variable normal de parámetros: $\mathbf{M_k \equiv N(\overline{X}_p, S_p)}$

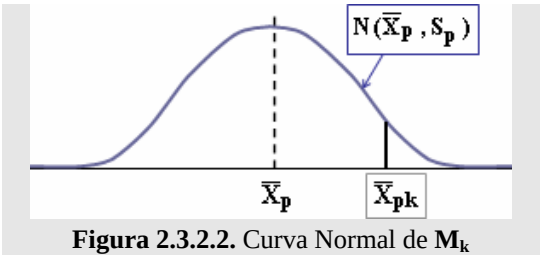


Figura 2.3.2.2. Curva Normal de $\mathbf{M_k}$

"Centrando y reduciendo" la variable aleatoria $\mathbf{M_k}$ (normalizando), podemos definir una variable **U**, que será una variable normal reducida:

$$\mathbf{U_k = \frac{M_k - \overline{X}_p}{S_{pk}}}$$

En estas condiciones, la variable normalizada $\mathbf{U_k}$, se ajusta a una variable normal de parámetros: $\mathbf{U_k \equiv N(0,1)}$

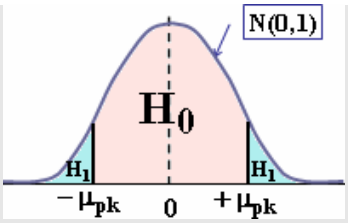


Figura 2.3.2.3. Curva Normal de $\mathbf{U_k}$

De modo que si **n** y $\mathbf{n_{pk}}$ son suficientemente grandes (mayores a 30), se puede demostrar que la probabilidad asociada a la ocurrencia de un valor dado de $\mathbf{M_k}$ es igual a la probabilidad que una *ley normal* sobrepase el valor $\boldsymbol{\mu_{pk}}$ calculado sobre la media $\overline{\mathbf{X}}_{\mathbf{pk}}$ de muestra, de los $\mathbf{n_{pk}}$ elementos de la clase **k**.

La **p**-ésima variable cuantitativa será "típica" de la **k**-ésima clase de una variable cualitativa, si la probabilidad crítica del valor $\boldsymbol{\mu_{pk}}$ es muy pequeña; o bien, si el valor $\boldsymbol{\mu_{pk}}$ es muy grande.

Se le llama Valor de Test al valor: $$\boldsymbol{\mu_{pk}} = \frac{\overline{\mathbf{X}}_{\mathbf{pk}} - \overline{\mathbf{X}}_{\mathbf{p}}}{\mathbf{S_{pk}}}$$ (Ecuación 2.3.2.2)

Donde $\mathbf{S_{pk}} = \sqrt{\mathbf{S}_{\mathbf{pk}}^{\mathbf{2}}}$

Señalemos que $\boldsymbol{\mu_{pk}} \cdot \mathbf{S_{pk}} = \overline{\mathbf{X}}_{\mathbf{jk}} - \overline{\mathbf{X}}$. La "distancia" entre la media de la **p**-ésima variable cuantitativa y *la media* de la misma en la **k**-ésima clase de la variable cualitativa, queda evaluada en términos de cantidad de desviaciones estándar de una variable normal.

Interpretación:

El valor de prueba $\boldsymbol{\mu_{pk}}$ es calculado con el signo correspondiente. Si es positivo (+), la clase k considerada se "caracteriza" por presentar valores fuertes (altos) de la variable cuantitativa ($\overline{\mathbf{X}}_{\mathbf{pk}} > \overline{\mathbf{X}}_{\mathbf{p}}$). Si el valor es negativo (-) se caracteriza por presentar valores bajos de la variable continua ($\overline{\mathbf{X}}_{\mathbf{pk}} < \overline{\mathbf{X}}_{\mathbf{p}}$).

Disponemos así de un índice que permite la comparación entre k clases de una variable cualitativa en función de una variable p cuantitativa. Podemos ordenar las p variables cuantitativas (continuas) según la "importancia" del valor de prueba para una clase dada. [Crivisky, 1997 y Droesbeke, 1992]

Veamos un ejemplo[3] *sencillo:*

Supongamos un problema de clasificación con la variable clase **C** (tomando valores *+* y -) y un atributo: **N** (*numérico*). Sea el siguiente conjunto de datos:

N	C
5	+
2.2	-
1.8	-
4	+
2	+
3	-

Tenemos las siguientes probabilidades **P(C)**, y **P(N | C)**, y sabemos que **n= 6**

npk	
+	3
-	3

Pk=P(C)	
+	3/6=**0.5**
-	3/6=**0.5**

N
Xp=3
Sp=1.27

P(N | C)	N(μ,σ)
+	N(μ=3.67, σ=1.53)
-	N(μ=2.33, σ=0.61)

S^2jk	+	-
N	$\frac{(1.27)^2}{3}.\frac{6-3}{5}=0.32$	$\frac{(1.27)^2}{3}.\frac{6-3}{5}=0.32$

V-test Mjk	+	-
N	$\frac{3.67-3}{\sqrt{0.32}}=1.18$	$\frac{2.33-3}{\sqrt{0.32}}=-1.18$

Podemos ver que para los casos de la clase "+", se caracterizan con una media de 3.67 por valores más altos variable continua N, con un nivel de significación bajo del 90%, por el contrario la clase "-", se caracteriza por los valores más bajos, con una media del 2.33 y el mismo nivel de significación del 90%

Ejemplo 2.3.2.1. Cálculo V-Test

2.3.2.1.2. Valor Test – Características Cualitativas Nominales

2. *Variables nominales*, el valor de test proporciona una ordenación de las modalidades a partir de un criterio estadístico el cual evalúa la importancia de la desviación entre dos proporciones, sea cual sea la importancia de estas proporciones.

En la caracterización de una variable nominal por otra, se calcula el estadístico de Chi^2 asociado al cruzamiento de las dos variables nominales, así como la probabilidad de "sobrepasar" el valor umbral calculado. El valor de la ley normal que tiene esta probabilidad se llama valor de test. Mientras mayor sea el valor de test, más "interesante" será la tabla de cruce, o sea más significativa la muestra.

Sobre una población de **n** individuos (casos), se han observados **q** variables discretas: $\mathbf{X_1}$, $\mathbf{X_2}$, $\mathbf{X_3}$,…, $\mathbf{X_q}$, con características definidas en las modalidades de una variable clasificadora.

[3] Ver Ejemplo 2.2.2.1. Estimación de parámetros, epígrafe 2.2.2. Clasificador NB

Centrando la atención por un grupo particular de $\mathbf{n_k}$ casos incluidos en la *clase* **k** y $\mathbf{n_j}$ *casos con una característica* **j.** ¿Cómo explicar por orden de importancia las variables que caracterizan mejor a la clase **k**? ¿Cómo describir las características **j**-ésima de las variables que sean más típicos de esa clase?

Una variable no presenta interés para caracterizar un grupo particular de $\mathbf{n_k}$ individuos, si los valores de proporción que encontramos de esta modalidad, parecen estar distribuidos aleatoriamente (al azar) entre los **n** valores observados. Cuanto "más dudosa" parezca la hipótesis de una distribución aleatoria, mejor caracterizará la variable en cuestión al grupo de $\mathbf{n_{jk}}$ individuos.

Se procede como para un test estadístico clásico, la hipótesis "nula" ($\mathbf{H_0}$) y la hipótesis de una extracción aleatoria (al azar) de los $\mathbf{n_k}$ valores entre las **n** observaciones, o los $\mathbf{n_{jk}}$ características entre las $\mathbf{n_k}$ de una clase. La extracción se supone sin reposición debido a que cada uno de los valores es una y solo una de las observaciones.

Con esta hipótesis de trabajo se calcula la probabilidad de observar una configuración de valores al menos tan extrema como la de la muestra. Es la probabilidad crítica asociada al test de la hipótesis nula ($\mathbf{H_0}$). Cuanto más baja sea esta probabilidad, menos viable será la hipótesis de distribución a la azar.

Para clasificar las variables y sus modalidades por su *significación*, se ponderan en función de las probabilidades críticas. La *variable o modalidad más significativa o típica es aquella que corresponde a la probabilidad más pequeña.* Cuanto mayor sea la abundancia de una modalidad $\mathbf{n_k}$, más débil es la probabilidad de extracción aleatoria y más dudosa es la hipótesis nula $\mathbf{H_0}$. Dada las características del estadístico a estos valores de probabilidad menores corresponden por lo tanto valores de prueba mayores.

Llamemos **k** al grupo (clase) de $\mathbf{n_k}$ individuos (casos) y **j** a una modalidad (posible valor) o característica de una de las variables nominales.

Para saber si esta modalidad **j** es una característica pertinente del grupo, se debe responder la siguiente cuestión: ¿Es la característica **j** significativamente más abundante en la clase **k**, que en la población de los **n** individuos?

Los elementos del problema se reúnen en la tabla de contingencia donde los efectivos no indicados se calculan por diferencia:

	En la clase **k**	Fuera de la clase **k**	*Marginales*
En la modalidad **j**	$\mathbf{n_{jk}}$	=	$\mathbf{n_j}$
Fuera de la modalidad **j**	=	=	=
Marginales	$\mathbf{n_k}$	=	**n**

Tabla 2.3.2.1. Tabla de contingencia para cálculo V-test

Se procede cómo para un test clásico:

1. La hipótesis nula [$\mathbf{H_0}$] es aquí la hipótesis de una extracción al azar (sin reposición) de los $\mathbf{n_k}$ casos en la clase, entre los **n** casos totales de la población. Esta hipótesis asegura la igualdad de proporciones:

$$H_0 : \frac{n_{jk}}{n_k} = \frac{n_j}{n}$$

(P_{jk} ; P_j)

Hipótesis nula. *No hay diferencia del grupo con la población*

2. La hipótesis alternativa [$\mathbf{H_1}$], especifica una proporción **j** anormalmente elevada (abundante) entre los $\mathbf{n_k}$ casos incluidos en la clase **k**, es decir:

$$H_1 : \frac{n_{jk}}{n_k} \neq \frac{n_j}{n}$$

Hipótesis alternativa. *Hay diferencia del grupo con la población*

Si llamamos "suceso" el hecho de pertenecer a la característica o modalidad "**j**", nos interesamos por el número de sucesos $\mathbf{n_{jk}}$, observados en una muestra de tamaño $\mathbf{n_k}$ extraída al azar, sin reposición, de entre los **n** casos de la población total. Si repetimos un gran número de veces la operación de seleccionar aleatoriamente de $\mathbf{n_{jk}}$ valores entre los $\mathbf{n_k}$ valores posibles, hemos de observar que esas diferentes muestras presentan una gran variedad de valores de $\mathbf{P_{jk}}$ (proporción media intra-clase **k** de la característica **j**):

$$p_{jk_1} = \frac{n_{jk_1}}{n_{k_1}}$$ luego otra $$p_{jk_2} = \frac{n_{jk_2}}{n_{k_2}}$$ y así sucesivamente en un gran número de extracciones al azar

Según la teoría de grandes números, puede demostrarse que la serie de esas "medidas de muestra" constituye una variable aleatoria $\mathbf{M_{jk}}$, que se distribuye al azar en la clase **j**.

Tomando como estimador la probabilidad de distribución de esta característica dada su proporción en la clase, entonces:

- La *media del estimador* tiende hacia **P**:

$$P : E_{H_0}(M_{jk}) \qquad \boxed{p = P_{jk} = \frac{n_{jk}}{n_k}}$$

(Ecuación 2.3.2.3.) Probabilidad condicional
Medida de la probabilidad de la distribución de la característica **j** en la clase **k**, bajo la hipótesis nula.

- La *varianza estimada* de la medida de probabilidad P_{jk} tiende hacia $\mathbf{S^2}$:

$$S^2_{jk} : Var_{H_0}(M_{jk}) \qquad \boxed{S^2_{jk} = \left(\frac{P_j \cdot (1 - P_j)}{n_k}\right) \cdot \left(\frac{n - n_k}{n - 1}\right)}$$

(Ecuación 2.3.2.4.)
Varianza estimada de la medida estimada de la probabilidad **pjk**..

Donde

n	número *total de observaciones* (casos)
Pj = nj/n	probabilidad a priori de pertenecer a una característica **j**
Pk= nk/n	probabilidad a priori de pertenecer a una clase **k**
Pjk = njk/nk	probabilidad posteriori de **j** dado **k**.
nj = Pj · n	número total de observaciones *con la característica* **j**
nk = Pk · n	número total de observaciones *de la clase* **k**
njk = Pjk · nk	número de casos *de la clase* **k**, que *presentan la característica* **j**.

Esta variable aleatoria M_{jk}, se ajusta a una variable normal de parámetros: $\mathbf{M_{jk} \equiv N(P,S)}$

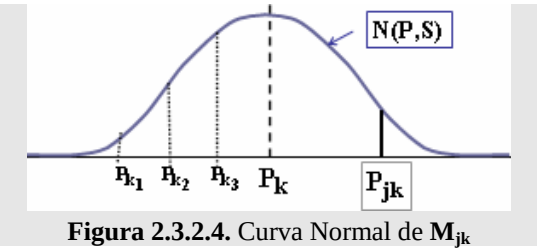


Figura 2.3.2.4. Curva Normal de $\mathbf{M_{jk}}$

Centrando y reduciendo un valor de muestra $\mathbf{P_{jk}}$, podemos definir otra variable **U**, que será también una variable normal reducida de parámetros:

$\mathbf{U = \dfrac{M_{jk} - P_j}{S_{jk}}}$ donde $\mathbf{U \equiv N(0,1)}$. Conteniendo un gran número de valores posibles de μ_{jk}

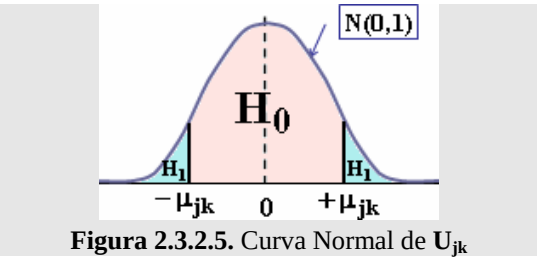


Figura 2.3.2.5. Curva Normal de $\mathbf{U_{jk}}$

Si conocemos $\mathbf{n_{jk}}$ y $\mathbf{n_k}$ se puede demostrar que la probabilidad asociada a la ocurrencia al azar de un valor dado de $\mathbf{M_{jk}=P_{jk}}$, e igual a la probabilidad que una ley normal sobrepase el valor de μ_{jk} calculado sobre la media $\mathbf{P_{jk}}$ de muestra de los $\mathbf{n_{jk}}$ característicos de la clase **k**.

Se le llama Valor de Test al valor: $\boxed{\mu_{jk} = \dfrac{P_{jk} - P_j}{S_{jk}}}$ Donde $S_{jk} = \sqrt{S_{jk}^2}$

(Ecuación 2.3.2.5.)

Interpretación

- Si el V-test es positivo, la característica **j** de la clase **k**, considera presente "valores elevados" de la probabilidad condicional (P_{jk}) de la característica dada la clase.
- Si el valor de test es negativo, se caracteriza por "valores muy bajos" de la probabilidad condicional (P_{jk}) de la característica dada la clase.
- Si el valor de prueba calculado es menor al valor crítico definido en la curva normal Z: $\mu_{jk} < \mu_{crítico}$, entonces la hipótesis nula H_0 es válida y no se puede rechazar.

- Si el valor de prueba calculado es mayor al valor crítico definido en la curva normal Z: $\mu_{jk} > \mu_{crítico}$, entonces la hipótesis nula H_1 es válida y se puede rechazar la H_0.
- Para V-test ≥ 2, se rechaza la H_0, al nivel de α=0.05 (95 de confianza).
- Para V-test ≥ 2.56, se rechaza la H_0, al nivel de α=0.01 (99 de confianza).

El valor de prueba permite entonces clasificar por orden "creciente" de importancia, los atributos característicos de una clase o grupo de individuos estadísticos. [Crivisky, 1997 y Droesbeke, 1992]

Ejemplo[4]*:*

Supongamos un problema de clasificación con la variable clase **C** (tomando valores *+* y -) y dos atributos: **D** (discreto tomando valores *a* y *b*). Sea el siguiente conjunto de casos nuestra tabla de datos:

D	C
a	+
a	-
a	-
b	+
b	+
a	-

n= 6, k=2, j=2

nk	
+	3
-	3

nj	
a	4
b	2

njk	+	-
a	1	3
b	2	0

Aplicando las ecuaciones 2.3.2.4. y 2.3.2.5. Obtenemos las siguientes estimaciones:

Pk=P(C)	
+	3/6=**0.5**
-	3/6=**0.5**

Pj=P(D)	
a	4/6=**0.67**
b	2/6=**0.33**

Pjk= P(D | C)	+	-
a	1/3=**0.33**	3/3=**1**
b	2/3=**0.67**	**0**

S^2jk	+	-
a	$\frac{0.67\cdot(1-0.67)}{3}\cdot\frac{6-3}{5}=0.04$	$\frac{0.67\cdot(1-0.67)}{3}\cdot\frac{6-3}{5}=0.04$
b	$\frac{0.33\cdot(1-0.33)}{3}\cdot\frac{6-3}{5}=0.04$	$\frac{0.33\cdot(1-0.33)}{3}\cdot\frac{6-3}{5}=0.04$

V-test Mjk	+	-
a	$\frac{0.33-0.67}{\sqrt{0.04}}=-1.7$	$\frac{1-0.67}{\sqrt{0.04}}=1.65$
b	$\frac{0.67-0.33}{\sqrt{0.04}}=1.7$	$\frac{0-0.33}{\sqrt{0.04}}=-1.65$

Podemos ver que la clase "+" **no** se caracteriza, con un nivel del 90% de certeza, por D="a", perteneciendo 33% del la clase "+" y si está caracterizada por la característica D=b, tambien con un nivel del 90%. Por el contrario, la clase "-" se caracteriza por la característica d=a y no por la D=b, esto tambien con un nivel de casi 90%. El nivel de significación es un poco bajo, menor al umbral de 2.

Ejemplo 2.3.2.2. Cálculo V-Test

[4] Ver Ejemplo 2.2.2.1. Estimación de parámetros, epígrafe 2.2.2. Clasificador NB

2.3.2.2. Criterios de significación estadística

Si la prueba del valor de prueba, tiene diferentes resultados para $\alpha = 0{,}01$, $\alpha = 0{,}05$ o $\alpha = 0{,}10$; se toman los siguientes criterios [Hair, Anderson, Tathem y Blas, 1999]:

- Cuando el V-test (absoluto) cumple con **α = 0,01** de significancia, el caso *es altamente significativo*. **(|V-test| ≥ que 2.58). (99% certeza)**
- Si el resultado se encuentra **entre α = 0,01** (*no lo cumple*) y **α = 0,05**(*si lo cumple*), el caso es *significativamente probable*. **(|V-test| entre 1.96 a 2.58). (95% certeza)**
- Si el resultado con **α = 0,05** (*no lo cumple*), y **α = 0,10** (*si lo cumple*) el caso es *significativamente poco probable*. **(|V-test| entre 1.96 a 1.65). (90% certeza)**
- Si el resultado con **α = 0,01** (*no lo cumple*)**, α = 0,05** (*no lo cumple*) y **α = 0,10** (*no lo cumple*) es *altamente no significativo*. **(|V-test| < de 1.65).**

El signo del valor de test determina, en las características continuas, si está relacionada con la clase en sus mayores valores (+) o valores más bajos (-). En el caso de las características cualitativas, que el V-test sea negativo significa que cuando aumenta la proporción de casos en la variable de clases disminuye la proporción en las variables descriptivas (relación opuesta).

2.3.3. Técnica para la descripción de variables cualitativas categóricas - DEMOOD

Se utilizan en la aproximación descriptiva de matrices de datos, [Morineau, 1994].

Son procedimientos de descripción de las observaciones y variables en una matriz de datos, con la finalidad de "*caracterizar significativamente*" grupos de una partición.

El objetivo es caracterizar en forma rápida y completa datos voluminosos. Se obtiene la caracterización estadística automáticamente, de las clases de una partición en función de todas las variables.

En este procedimiento de caracterización, la noción clave para la ordenación de los elementos característicos es el denominado *Valor de Test* (epígrafe 2.3.2.1.). Mediante éste se evalúa el interés de un elemento para caracterizar una categoría de individuos (casos) a partir de un estadístico calculado sobre la muestra.

Se caracteriza estadísticamente una variable cualitativa (nominal). Los *elementos característicos pueden ser otras variables nominales o también variables continuas.*

- Si los elementos característicos son las *modalidades de otras variables nominales se detectan las modalidades más significativas*. Los *valores test se calculan para todas las modalidades de las variables nominales, ordenándoselas, por tanto, en función de estos valores decrecientes para caracterizar cada modalidad*. La clasificación proporcionada por los valores test ordena las modalidades a partir de un criterio estadístico el cual evalúa la importancia de la desviación entre dos proporciones, la del grupo y la de la población general, es decir evalúa la abundancia de la modalidad en el grupo, frente a la abundancia de la modalidad en la población total.

- Si los elementos característicos *son variables continúas*, para clasificar las más características de la variable nominal, se efectúan todos los análisis de la variancia. El mejor *analizas de variancia es el que corresponde al estadístico de Fisher mas significativo y corresponde al parámetro continuo mas previsible con ayuda del factor. Para cada estadístico de Fisher se calcula la probabilidad de ser sobrepasado*. El valor test asociado es el valor de una variable nominal que tiene la misma probabilidad de ser sobrepasada. Entonces, se ordenan las variables características siguiendo el orden decreciente de los Valores Test.
- Si los elementos característicos son *otras variables también nominales*, se *calcula el estadístico de Chi cuadrado,* asociado al cruzamiento de dos variables nominales, así como la probabilidad de sobrepasar el valor calculado. El valor de la ley normal que tiene la misma probabilidad de ser sobrepasada es el valor test. *Mientras mayor sea el valor test, más interesante será la tabla de cruzamiento.(mayor asociación entre las variables.). Los (valores test< 2), determinan independencia entre ambas variables, o sea ausencia de significación estadística.*

Valor Test (V-test) en el DEMOD

En el procedimiento Demod los V-test, se calculan para todas las características (modalidades o valores) de todas las variables nominales. Lo que permite ordenarlas en función de sus valores de prueba V-test decrecientes, mostrando la importancia de la variable y característica en la clase. Se asegura con esto, que ninguna característica significativa estadísticamente se pueda escapar.

- Cuando las variables explicativas o descriptivas de la clase son nominales *cualitativas*:

  La ordenación que realiza el V-test se basa en un *criterio estadístico* que evalúa la importancia de la *desviación entre las proporciones*, sea cual sea la importancia de esas proporciones. Se puede obtener así, clasificación de las modalidades en función de su abundancia en la categoría a describir. En esta lista ordenada, la primera característica, es la modalidad mejor representada en la clase y tendrá asociado el valor de test más alto, o sea con mayor certeza o significación de pertenecer a esa clase y por lo tanto ser un "elemento característico de esa clase".

  Esta técnica permite también asociar con el V-test, el *porcentaje de casos de ejemplo que co-ocurre, para una clase dada, en una característica* y el *porcentaje de casos en total presentes en la clase* y el *porcentaje presente en la característica o modalidad*. Esto proporciona valores numéricos que representan la intensidad con que una característica está asociada a una clase, los mismos ordenados según su significancia estadística, (nivel de certeza).

  Entonces, el *porcentaje de casos de ejemplo que co-ocurre, para una clase dada, en una característica*, se denomina *porcentaje de la modalidad en la clase*. Por otro lado el *porcentaje casos de ejemplos totales* de una clase o de una modalidad característica, se lo denomina *porcentaje global* de la clase o modalidad en la población.

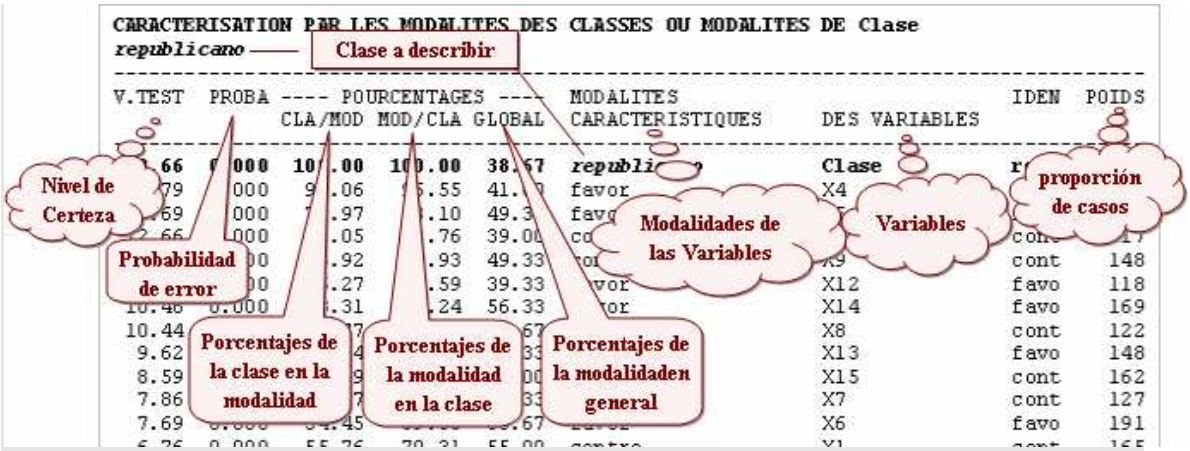


Figura 2.3.3.1. Demod – descripción de una variable cualitativa con otras cualitativas.

En este trabajo se utilizará la implementación del algoritmo Demod del paquete informático SDPAD [Sdpad, 2002].

- Cuando las variables explicativas o descriptivas de la clase son cuantitativas *numéricas*:

  Esta técnica permite también asociar la media y desviación estimada condicional de la clase, con el V-test. Con esto proporciona valores continuos que representan la media y desviación típica estimada que presentan las características en un clase dada, los mismos, ordenados según su significancia estadística, (nivel de certeza), tanto para los valores más altos de la variable explicativa numérica, cómo para sus valores bajos..

  La ordenación que realiza el V-test se basa en un criterio estadístico que evalúa la importancia de la desviación entre la medida (de la variable numérica descriptiva) sobre el grupo y la medida sobre la población. “Evalúa de alguna manera la distancia entre la media general y la media en el grupo, en número de desviaciones tipo de una ley normal”. [Morineau, 1994].

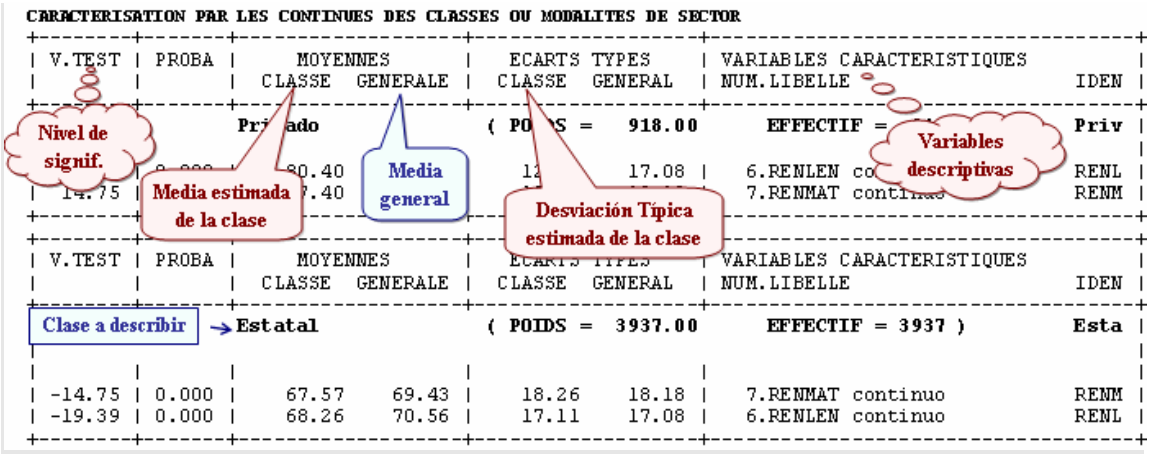


Figura 2.3.3.2. Demod – descripción de una variable cualitativa con otras numéricas.

3. HIPÓTESIS DE TRABAJO

3.1. Demod, clasificador bayesiano

Demod es un clasificador bayesiano, no grafico, que ordena la distribución de probabilidades según nivel de significación estadadística.

- Variables explicativas o descriptivas de la clase son nominales *cualitativas*: La ordenación que realiza el V-test se basa en un *criterio estadístico* que evalúa la importancia de la *desviación entre las proporciones* (abundancia) de la característica en la clase y en la población total, (epígrafe 2.3.2.2.2).

Esta técnica permite asociar al V-test, tres proporciones: el *porcentaje de casos de ejemplo que co-ocurren, para una clase dada, en una característica;* el *porcentaje de casos en total presentes en la clase;* y el *porcentaje presente en la característica o modalidad*.

- El *porcentaje de casos de ejemplo que co-ocurren, para una característica en una clase,* se denomina *porcentaje de la modalidad en la clase,* proporción que equivale a la *probabilidad condicional posteriori* de "dada como evidencia una clase, se de la característica".

  Siendo **k** la *clase*, X variable *descriptiva* y **j** un valor característico de la misma: P(**j** | **k**). Permite *caracterizar al clasificador*, indicando las proporciones de casos dados en cada característica.

- El *porcentaje de casos de ejemplo que co-ocurre para una clase en una característica*, se denomina *porcentaje de la clase en la modalidad*, proporción que equivale a la *probabilidad condicional posteriori* de "dada como evidencia una característica, se de la clase".

  Siendo **k** la *clase*, X variable *descriptiva* y **j** un valor de la misma: P(**k** | **j**). Permite *predecir* ante la evidencia de *una característica*, la probabilidad de pertenecer a una clase.

- El *porcentaje de casos de ejemplos totales* de una clase o el porcentaje total de una modalidad característica, se lo denomina *porcentaje global* de la clase o porcentaje global de la modalidad (respecto a la población); proporción que equivale a la *probabilidad a priori* de que se de la clase o de que se de esa modalidad, independientemente y según los casos tomados cómo ejemplo.

  Siendo **k** la *clase,* X variable *descriptiva* y **j** un valor de la misma: P (**k**) y P(**j**).

Sobre la base de la figura 2.3.3.1., podemos visualizar en la siguiente figura 3.1.1., La semejanza entre Naïve Bayes y Demod.

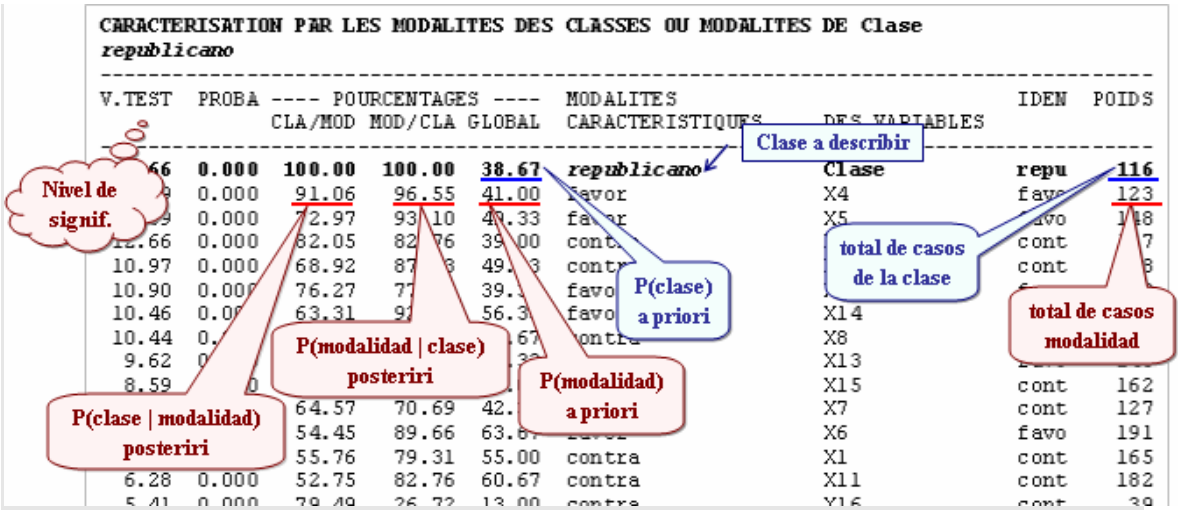


Figura 3.1.1. Descripción del Demod, equivalencia con las probabilidades de NB.

Ejemplo. Utilizaremos la base de datos "Votación" [5]. (En el anexo A, se adjunta las salidas completas de todos los programas utilizados para implementar los algoritmos correspondientes).

Esta base de datos posee registros de 300 individuos, donde para cada uno de ellos se relevaron la opinión *favorable*, *en contra* o *desconocida* de 16 aspectos y según su votación si son "republicanos" o "demócratas". En la tabla 3.1.1., podemos ver las variables.

Cod.	Variable	Valores o modalidades
X1	niños discapacitados	*a favor, en contra o desconocido*
X2	participación en el costo del proyecto del agua	*a favor, en contra o desconocido*
X3	adopción de la resolución sobre el presupuesto	*a favor, en contra o desconocido*
X4	congelamiento de los honorarios médicos	*a favor, en contra o desconocido*
X5	ayuda a El Salvador	*a favor, en contra o desconocido*
X6	grupos religiosos en las escuelas	*a favor, en contra o desconocido*
X7	prohibición de las pruebas anti satélistes	*a favor, en contra o desconocido*
X8	ayuda a los contras de Nicaragua	*a favor, en contra o desconocido*
X9	misil mx	*a favor, en contra o desconocido*
X10	inmigración	*a favor, en contra o desconocido*
X11	reducción a la corporación Synfuels	*a favor, en contra o desconocido*
X12	presupuesto de educación	*a favor, en contra o desconocido*
X13	derecho a demandar de la Superfund	*a favor, en contra o desconocido*
X14	crimen	*a favor, en contra o desconocido*
X15	exportaciones sin impuestos	*a favor, en contra o desconocido*
X16	acta sudafricana de administración de exportaciones	*a favor, en contra o desconocido*
Clase	Votación	*demócrata o republicano*

Tabla 3.1.1. Variables de la tabla "Votación"

Antes de realizar el Demod para caracterizar los grupos de la variable de Clase (Votación), construiremos el clasificador bayesiano para determinar la distribución de probabilidades y modelar el clasificador, nos valdremos de la ayuda de Elvira. [Elvira, 2000]

[5] Datos extraídos de extraídos de http://www.ics.uci.edu/~mlearn/MLSummary.html.

La red bayesiana construida por NB quedaría como sigue:

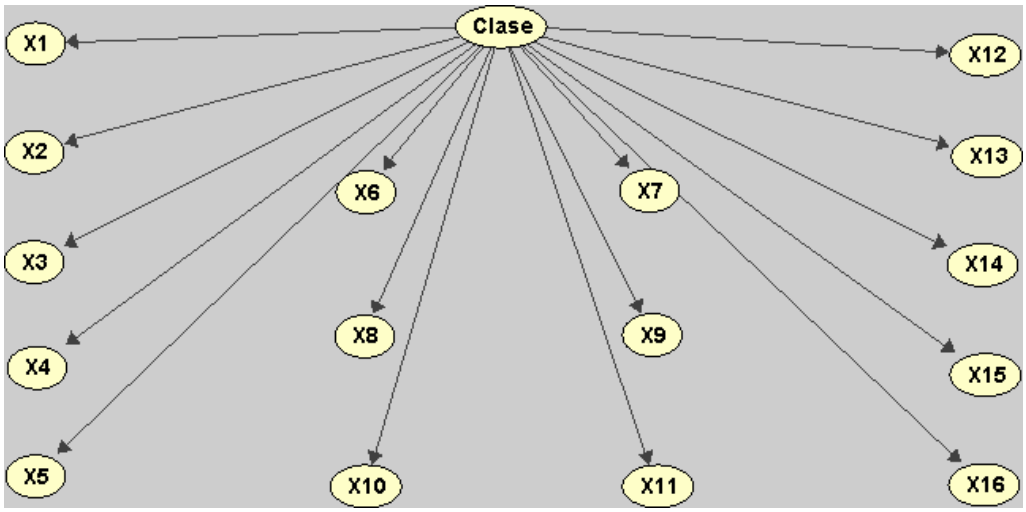


Figura 3.1.2. Red bayesiana construida con el clasificador NB.

Ahora, si incorporamos como evidencia a esta red las distintas clases, podremos inferir sus características. De esta manera en la figura 3.1.3 y 3.1.4., podremos ver la distribución de probabilidades una vez ingresado la evidencia "republicano" y "demócrata", respectivamente.

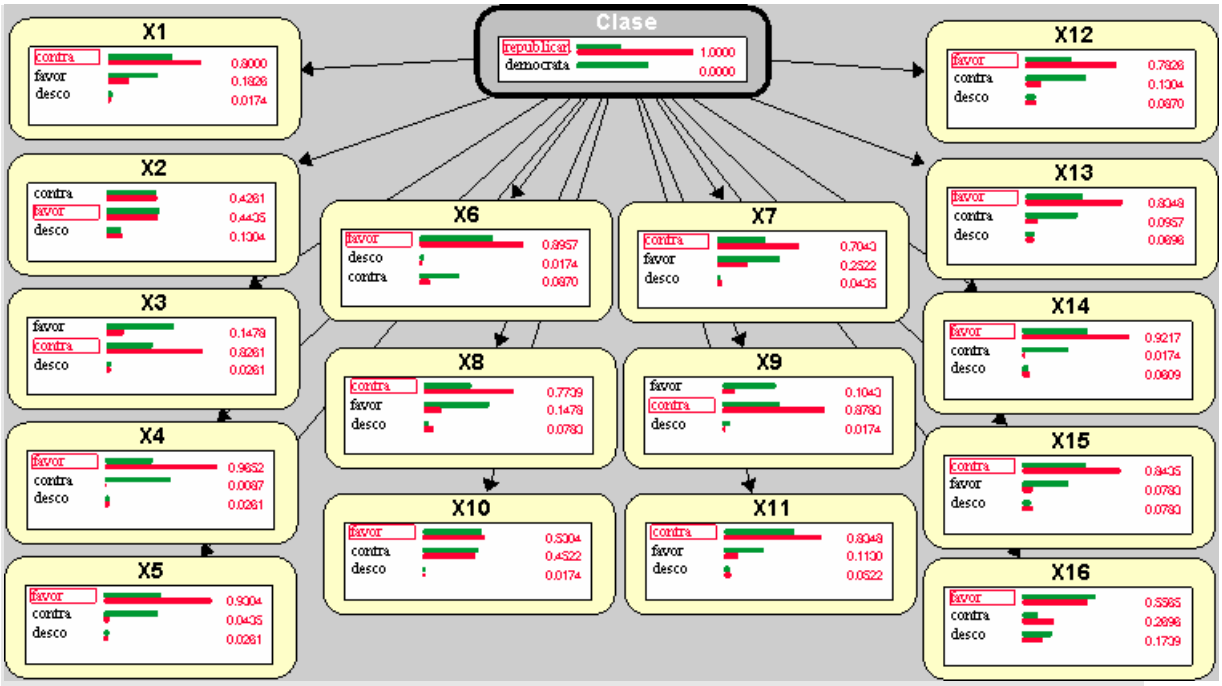


Figura 3.1.3. Distribución de probabilidad ingresada la evidencia "republicano"

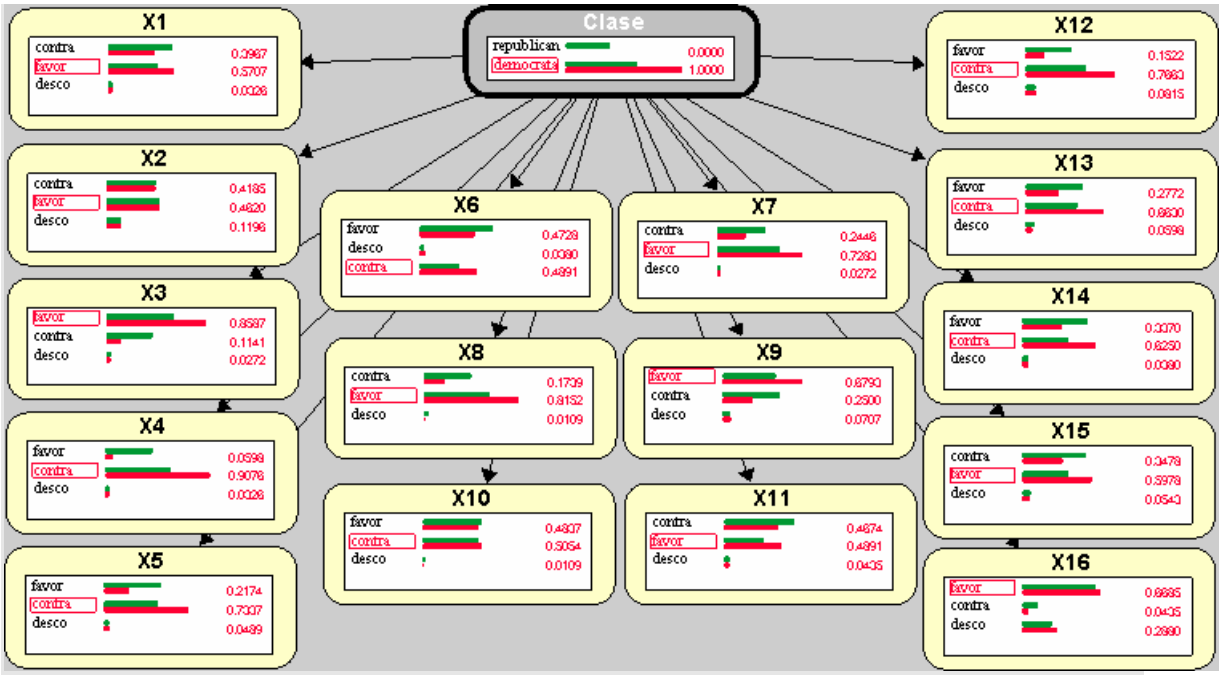


Figura 3.1.4. Distribución de probabilidad ingresada la evidencia "demócrata"

En la tabla 3.1.2., podemos analizar las probabilidades obtenidas para las variables X3, X4 y X5. Las mismas serán analizadas en el Demod.

probabilidades		**a priori**	**Evidencia 1**	**Evidencia 2**
Clase	**republicano**	**0.38**	**1**	**0**
	demócrata	**0.62**	**0**	**1**
X3	**"favor"**	0.59	0.15	**0.86**
	"contra"	0.39	**0.83**	0.11
	"desco"	0.03	0.03	0.03
X4	**"favor"**	0.41	**0.97**	0.06
	"contra"	0.56	0.01	**0.91**
	"desco"	0.03	0.03	0.03
X5	**"favor"**	0.49	**0.93**	0.22
	"contra"	0.47	0.04	**0.73**
	"desco"	0.04	0.03	0.05

Tabla 3.1.2. Probabilidades a priori y posteriori de X3, X4 y X5

Realizamos el Demod, para describir las clases "republicano" y "demócrata", obtenemos los siguientes resultados:

Caracterización de las modalidades de las clases de la variable Clase

Etiquetas de las variables	modalidades caracterísitcas	% de la modalidad en la Clase	% de la modalidad en General	Valor-Test	Probabilidad	Peso
Clase	**republicano**	**100.00**	**38.67**	**19.66**	**0.000**	**116**
X4	favor	96.55	41.00	16.79	0.000	123
X5	favor	93.10	49.33	12.69	0.000	148
X3	contra	82.76	39.00	12.66	0.000	117
X9	contra	87.93	49.33	10.97	0.000	148
X12	favor	77.59	39.33	10.90	0.000	118
X14	favor	92.24	56.33	10.46	0.000	169
X8	contra	77.59	40.67	10.44	0.000	122
X13	favor	83.62	49.33	9.62	0.000	148
X15	contra	84.48	54.00	8.59	0.000	162
X7	contra	70.69	42.33	7.86	0.000	127
X6	favor	89.66	63.67	7.69	0.000	191
X1	contra	79.31	55.00	6.76	0.000	165
X11	contra	82.76	60.67	6.28	0.000	182
X16	contra	26.72	13.00	5.41	0.000	39
X8	desco	7.76	3.67	2.66	0.004	11
X14	desco	6.03	4.67	0.62	0.267	14
X10	favor	52.59	50.00	0.59	0.277	150
X15	desco	7.76	6.33	0.57	0.284	19
X7	desco	4.31	3.33	0.44	0.332	10
X13	desco	6.90	6.33	0.09	0.464	19
X2	desco	12.93	12.33	0.08	0.468	37
X11	desco	5.17	4.67	0.07	0.473	14
X12	desco	8.62	8.33	0.06	0.477	25
X2	contra	42.24	42.00	0.05	0.480	126

Ejemplo 3.1.1. Caracterización de la clase “republicano”

Esta lista está ordenada según nivel de significación estadística (V-test ≥ 2), con un 99% de certeza (epígrafe 2.3.2.2.) podemos ver que los republicanos se caracterizan por estar a favor de “*congelamiento de los honorarios médicos*” (X4), (el 97% de los casos en la clase) y de “*ayuda a El Salvador*” (X5), (el 93% de los casos en la clase) y en contra de “*adopción de la resolución sobre el presupuesto*” (X3), (el 83% de los casos en la clase). Estas proporciones coinciden con las representadas por el clasificado bayesiano (tabla 3.2.1).

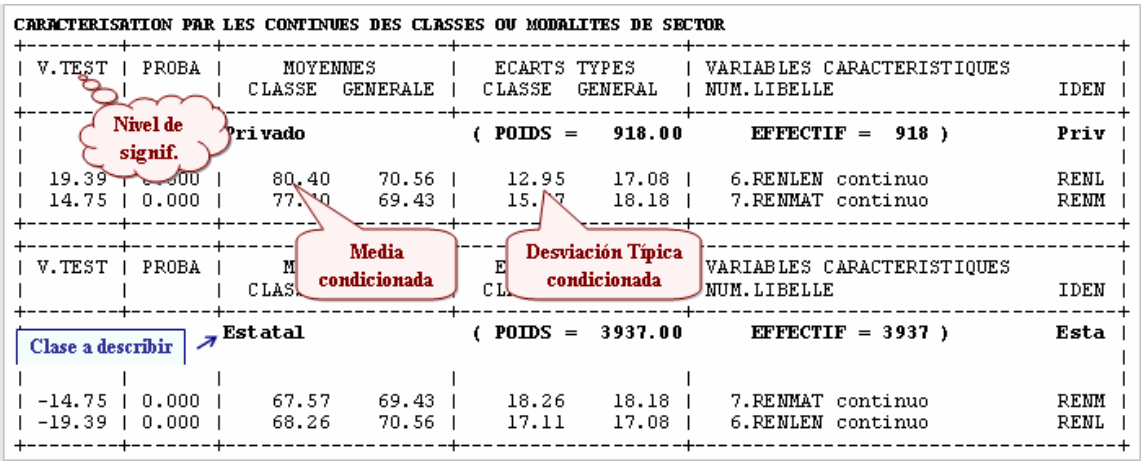
Etiquetas de las variables	modalidades caracterísitcas	% de la modalidad en la Clase	% de la modalidad en General	Valor-Test	Probabilidad	Peso
CLASE	**DEMOCRATA**	**100.00**	**61.33**	**19.66**	**0.000**	**184**
X4	contra	90.76	56.00	16.72	0.000	168
X3	favor	85.87	58.33	12.54	0.000	175
X5	contra	73.37	46.67	12.49	0.000	140
X8	favor	81.52	55.67	11.66	0.000	167
X14	contra	62.50	39.00	11.54	0.000	117
X12	contra	76.63	52.33	10.91	0.000	157
X9	favor	67.93	45.67	10.13	0.000	137
X13	contra	66.30	44.33	10.06	0.000	133
X15	favor	59.78	39.67	9.42	0.000	119
X7	favor	72.83	54.33	8.11	0.000	163
X6	contra	48.91	33.33	7.51	0.000	100
X11	favor	48.91	34.67	6.69	0.000	104
X1	favor	57.07	42.33	6.56	0.000	127
X16	desco	28.80	24.33	2.16	0.015	73
X9	desco	7.07	5.00	1.87	0.031	15
X16	favor	66.85	62.67	1.76	0.039	188
X10	contra	50.54	48.67	0.70	0.242	146
X5	desco	4.89	4.00	0.67	0.250	12
X6	desco	3.80	3.00	0.66	0.254	9
X1	desco	3.26	2.67	0.41	0.341	8
X3	desco	2.72	2.67	0.32	0.374	8
X2	favor	46.20	45.67	0.11	0.455	137
X4	desco	3.26	3.00	0.04	0.484	9

Ejemplo 3.1.2. Caracterización de la clase "republicano"

Esta lista está ordenada según nivel de significación estadística; con un 99% de certeza (epígrafe 2.3.2.2.) podemos ver que los republicanos se caracterizan por estar en contra de "*congelamiento de los honorarios médicos*" (X4), (el 90% de los casos en la clase) y de "*ayuda a El Salvador*" (X5), (el 74% de los casos en la clase) y a favor de "*adopción de la resolución sobre el presupuesto*" (X3), (el 85% de los casos en la clase). Proporciones que coinciden con el clasificado bayesiano (tabla 3.2.1).

- Variables explicativas o descriptivas de la clase son cuantitativas *numéricas*: Esta técnica permite también asociar la *media y desviación estimada condicional* de la clase, con el V-test. Proporciona valores continuos que representan la *media y desviación típica estimada* que presentan las características en una clase dada, los mismos, ordenados según su significancia estadística, (nivel de certeza), tanto para los valores más altos de la variable explicativa numérica, cómo para sus valores bajos (V-test positivo o negativo).

Sobre la base de la figura 2.3.3.2., podemos visualizar en la siguiente figura 3.1.5., la relación entre NB y Demod.



CARACTERISATION PAR LES CONTINUES DES CLASSES OU MODALITES DE SECTOR

V.TEST	PROBA	MOYENNES CLASSE	MOYENNES GENERALE	ECARTS TYPES CLASSE	ECARTS TYPES GENERAL	VARIABLES CARACTERISTIQUES NUM.LIBELLE	IDEN
		Privado	(POIDS = 918.00		EFFECTIF = 918)		Priv
19.39	0.000	80.40	70.56	12.95	17.08	6.RENLEN continuo	RENL
14.75	0.000	77.40	69.43	15.47	18.18	7.RENMAT continuo	RENM
V.TEST	PROBA	M CLAS		E CL		VARIABLES CARACTERISTIQUES NUM.LIBELLE	IDEN
		Estatal	(POIDS = 3937.00		EFFECTIF = 3937)		Esta
-14.75	0.000	67.57	69.43	18.26	18.18	7.RENMAT continuo	RENM
-19.39	0.000	68.26	70.56	17.11	17.08	6.RENLEN continuo	RENL

Figura 3.1.5.. Descripción del Demod, equivalencia con las probabilidades de NB.

Ejemplo. Se desea caracterizar los diferentes sectores "privado" y "estatal" de escuelas[6] en función de los rendimientos en lengua y matemática de 4855 niños.

El sector es la variable cualitativa que representa la clase, y las variables descriptivas continuas, son los rendimientos. Realizamos el demod y obtenemos los siguientes resultados:

Privado	(Peso = 918.00	Efectivos = 918)				
Variables caracterísitcas	**Medias estimada modalidad**	**Media General**	**Desvíos Típicos Modalidad**	**Desviación Típica General**	**Valor-Test**	**Probabilidad**
RENLEN	80.403	70.556	12.946	17.083	19.39	0.000
RENMAT	77.400	69.430	15.467	18.176	14.75	0.000
Estatal	**(Peso = 3937.00**	**Efectivos = 3937)**				
Variables caracterísitcas	**Medias estimada modalidad**	**Media General**	**Desvíos Típicos Modalidad**	**Desviación Típica General**	**Valor-Test**	**Probabilidad**
RENMAT	67.571	69.430	18.258	18.176	-14.75	0.000
RENLEN	68.259	70.556	17.115	17.083	-19.39	0.000

Ejemplo 3.1.3. Demod sobre Sector

Podemos ver que los 918 niños del sector privado se caracterizan por los *mayores promedios en lengua* (promedio de 80, V-test 19.39) y *mayores promedios* en matemática (promedio de 77, V-test = 14.75), ambos con un 99% de certeza (epígrafe 2.3.2.2.). Por otro lado los 3937 estudiantes del sector estatal, están con una certeza del 99%, caracterizados por los peores rendimientos (los más bajos valores de la variable, indicado por el valor de test negativo), es decir *bajos rendimiento* en lengua (promedio 68, V-test = -19.39) y *bajos rendimientos* en matemática (promedio de 67, V-test = -14.75).

Si construimos el clasificador NB (figura 3.1.6.), con la ayuda de Weka [Weka, 2003], obtenemos las medias y desviaciones condicionadas que coincide con el Demod (ejemplo 3.1.3).

[6] Fuente de Datos: ICEDE - MCE de la Nación

Naive Bayes (simple)
Scheme: weka.classifiers.bayes.NaiveBayesSimple
Instances: 4855
Attributes: 3 SECTOR, RENLEN, RENMAT
Class **Privado**: P(C) = 0.18921145
Attribute RENLEN Mean: 80.40295207 Standard Deviation: 12.95329977
Attribute RENMAT Mean: 77.40028322 Standard Deviation: 15.47491064
Class **Estatal**: P(C) = 0.81078855
Attribute RENLEN Mean: 68.25931166 Standard Deviation: 17.11667546
Attribute RENMAT Mean: 67.57092202 Standard Deviation: 18.26056256

Figura 3.1.6. Clasificador NB (Weka), sobre la clase Sector

3.2. Puntos débiles del enfoque bayesiano.

Los modelos bayesianos calculan el costo esperado asociado a cada una de las decisiones posibles, utilizando las probabilidades a posteriori y adoptan la decisión que tenga el menor costo o la mayor utilidad posible. Pero ¿Qué tan significativa es esta utilidad para ser tenida en cuenta?

Siguiendo con el ejemplo de "Votaciones" (ejemplo 3.1.2.), damos cómo evidencia al clasificador NB, la clase "demócrata" (probabilidad 1):

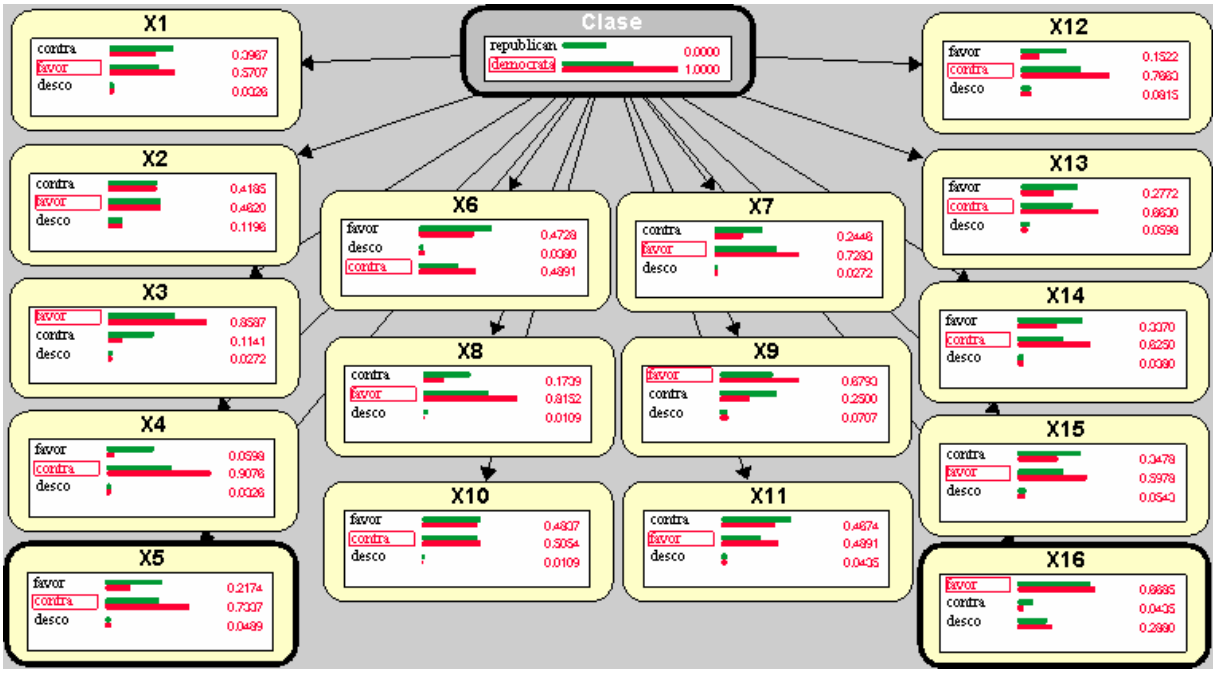


Figura 3.2.1. Red del clasificador NB para el caso "Votaciones"

Podemos ver como de los demócratas, el 46% están a favor de "*participación en el costo del proyecto del agua*" (X2) y parece más relevante, que el 28% que no saben sobre "*acta sudafricana de administración de exportaciones*" (X16). Aunque es más probable que los demócratas estén a favor de X2, esto no significa que sea la "característica típica" que permitirá diferenciarlo del republicano.

De hecho, si obtenemos un valor de prueba estadístico (epígrafe 2.3.2.2.1), comprobaremos que no es la variable X2 una característica significativa (ya sea que este a favor, en contra o no sepa respecto a esta). Lo mismo, ocurre con "*inmigración*"

(X10), hay un 50% de probabilidad de que los demócratas estén en contra, pero, ¿es válida esta afirmación?, ¿con qué precisión o certeza puede afirmarse?

Calculando el V-test (ecuación 2.3.2.5), para cada característica, en relación a la clase "demócrata", y ordenarlas en función de este nivel de significación, confirmamos que para ninguna de las modalidades de X2 y X10 existe significación estadística (V-test<1, altamente no significativo). Realizando el Demod para confirmarlo, se describen los demócratas teniendo en cuenta los criterios de significación (epígrafe 2.3.2.2.).

Etiquetas de las variables	modalidades caracterísitcas	% de la modalidad en la Clase	% de la modalidad en General	% de la Clase en la modalidad	Valor-Test	Probabilidad	Peso
CLASE	**DEMOCRATA**	**100.00**	**61.33**	**100.00**	**19.66**	**0.000**	**184**
X4	contra	90.76	56.00	99.40	16.72	0.000	168
X3	favor	85.87	58.33	90.29	12.54	0.000	175
X5	contra	73.37	46.67	96.43	12.49	0.000	140
X8	favor	81.52	55.67	89.82	11.66	0.000	167
X14	contra	62.50	39.00	98.29	11.54	0.000	117
X12	contra	76.63	52.33	89.81	10.91	0.000	157
X9	favor	67.93	45.67	91.24	10.13	0.000	137
X13	contra	66.30	44.33	91.73	10.06	0.000	133
X15	favor	59.78	39.67	92.44	9.42	0.000	119
X7	favor	72.83	54.33	82.21	8.11	0.000	163
X6	contra	48.91	33.33	90.00	7.51	0.000	100
X11	favor	48.91	34.67	86.54	6.69	0.000	104
X1	favor	57.07	42.33	82.68	6.56	0.000	127
X16	desco	28.80	24.33	72.60	2.16	0.015	73
X9	desco	7.07	5.00	86.67	1.87	0.031	15
X16	favor	66.85	62.67	65.43	1.76	0.039	188
X10	contra	50.54	48.67	63.70	0.70	0.242	146
X5	desco	4.89	4.00	75.00	0.67	0.250	12
X6	desco	3.80	3.00	77.78	0.66	0.254	9
X1	desco	3.26	2.67	75.00	0.41	0.341	8
X3	desco	2.72	2.67	62.50	0.32	0.374	8
X2	favor	46.20	45.67	62.04	0.11	0.455	137
X4	desco	3.26	3.00	66.67	0.04	0.484	9

Ejemplo 3.2.1. Demod clase "demócratas". V-test = 0, no figuran en el listado.

Podríamos pensar que es evidente que estas probabilidades no sean significativas, ya que son inferiores al 50% (0.5). Pero es curioso ver que con una confianza mayor del 90% (V-test =1.87) se puede afirmar que del total de los "demócratas", el 7% no saben respecto a "*misil mx*" (X9); esta es una característica de los demócratas aún siendo tan baja la proporción, pero si lo comparamos con la probabilidad global de esa característica (50% de casos), vemos que es probablemente significativo.

Por otro lado, que el 25% de los casos, esta en contra de X9, no es relevante por su significación (V-test = 0). Es una proporción mucho mayor que el 7% que desconoce X9, pero comparando con las probabilidades globales no es para nada significativo.

¿Cómo podemos saber cuales son las variables que están relacionadas con la clase de manera que permitan caracterizarla?, es decir ¿Cómo saber fácilmente cuáles son las relaciones que no se dan al zar para el conjunto de casos tomado para aprender?

El Demod (epígrafe 2.3.3.) me provee información muy útil respecto a esta cuestión. Un *listado ordenado de todas las variables* con el índice del [7]Chi^2 y el V-test (epígrafe 2.3.2.2.1.), medidas que me muestran la *dependencia / independencia de las variables características respecto a las variable de clase* y el *nivel de significación de esta relación*. Siguiendo el ejemplo 3.2.1., veamos cuales son las variables que realmente están relacionadas a la clase.

Etiqueta de la variable	Chi-2	grados de libertad	Valor-Test	Probabilidad
Clase	**295.80**	**1**	**99.99**	**0.000**
X4	245.14	2	99.99	0.000
X3	154.72	2	12.16	0.000
X5	147.10	2	11.85	0.000
X8	129.17	2	11.07	0.000
X12	124.06	2	10.84	0.000
X9	112.85	2	10.31	0.000
X14	111.43	2	10.24	0.000
X13	96.98	2	9.52	0.000
X15	81.70	2	8.69	0.000
X7	66.42	2	7.78	0.000
X6	55.74	2	7.07	0.000
X1	45.35	2	6.31	0.000
X11	43.18	2	6.14	0.000
X16	32.64	2	5.24	0.000
X10	0.81	2	-0.43	0.666
X2	0.09	2	-1.72	0.958

Ejemplo 3.2.2. Demod - Relación de dependencia con la variable clase

Podemos ver con mayor precisión que las variable X10 es independientes de la variable clase (V-test = -0.43), le sigue la variable X2, que presenta una baja dependencia negativa (probablemente significativo con V-test = -1.72). El V-test negativo, significa que cuando aumenta la proporción de casos en la variable de clases disminuye la proporción en las variables descriptivas (relación inversa). Podemos rescatar que la variable X2 tiene un nivel de significación con 90% confianza, esto puede que tenga que ver que hay alguna dependencia (negativa) en algunas de sus modalidades respecto algunas de las clases.

Las redes bayesianas pueden realizar la tarea de clasificación, la cual es un caso particular de la tarea de predicción, que se caracteriza por tener una sola de las variables de la base de datos (clasificador) que se desea predecir, mientras que todas las otras son los datos propios del caso que se desea clasificar. Pueden existir una gran cantidad de variables en la base de datos, algunas de las cuales estarán *directamente relacionadas* con la variable clasificadora que se quiere predecir, pero también pueden existir otras variables que no lo estén.

Las capacidades predictivas de las redes bayesianas están orientadas a pronosticar el valor de cualquiera de las variables pertenecientes al dominio de aplicación, en lugar de intentar maximizar el poder clasificatorio. Se ha visto como se puede seleccionar las variables que repercuten directamente en la clasificación, tomando aquellas que están más relacionadas o depende directamente de la clase y eliminado las que no.

[7] *Estadístico de Chi cuadrado*, asociado al cruzamiento de dos variables nominales

Esto último muestra una herramienta para seleccionar atributos relevantes en función de su dependencia con la clase, esta reflejada en el grado de significación de la asociación.

También se puede utilizar esta técnica, para establecer las restricciones de orden para el algoritmo K2 (epígrafe 2.2.3.) utilizado cómo clasificador: "el proporcionarle al algoritmo un orden entre las variables hace que éste tan sólo tenga que buscar el mejor conjunto de padres posibles de entre las variables predecesoras en el orden".

3.3. Naïve Bayes, validación de nuevas clasificaciones

Naïve Bayes es un clasificador bayesiano, gráficamente caracteriza las clases indicando la intensidad con que las variables se asocian con la misma; y provee la base para calcular un valor de prueba que determinará el nivel de significación estadística de estas relaciones y permitiendo validar la predicción de nuevas clasificaciones.

Siguiendo el ejemplo de "Votaciones" (ejemplo 3.1.2), si tenemos la opinión de un nuevo individuo (caso nuevo) sobre el "*congelamiento de los honorarios del medico* "(X4) y sobre la "*ayuda al El Salvado"* (X5). ¿Cómo será su votación?, ¿en qué clase lo apuntamos?

Sobre el clasificador bayesiano creado con Elvira [Elvira, 2000], introducimos estas nuevas evidencias para las variables X4 y X5 y vemos la clase y probabilidad pronosticada (epígrafe 2.2.1).

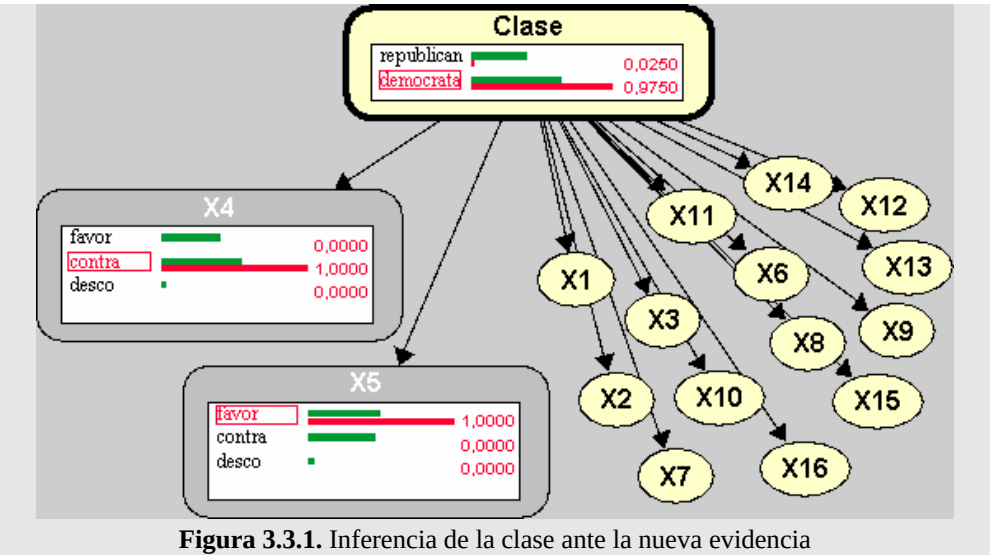


Figura 3.3.1. Inferencia de la clase ante la nueva evidencia

Puede verse en la figura 3.3.1., que la probabilidad condicional de "dado que está a favor de la variable X5 y en contra de la variable X4", sea "demócrata", es de **0.9750**. ¿Es estadísticamente significativo? Si combinamos las variables X4 y X5 y con el Demod describimos las clases con esta nueva variable combinada, podremos obtener la certeza de esta predicción.

Etiquetas de las variables	modalidades caracterísitcas	Valor-Test	% de la Clase en la modalidad	% de la modalidad en General	% de la modalidad en la Clase	Probabilidad	Peso
Clase	**democrata**	**19.66**	**100**	**61**	**100**	**0.000**	**184**
X4 * X5	X4CX5C	13.43	100	43	71	0.000	130
X4 * X5	**X4CX5F**	**4.59**	**97**	**10**	**16**	**0.000**	**31**
X4 * X5	X4CX5D	1.86	100	2	4	0.031	7
X4 * X5	X4FX5C	-1.00	33	2	1	0.159	6
X4 * X5	X4F-X5F	-16.11	8	39	5	0.000	116
Clase	**republicano**	**-19.66**	**0**	**39**	**0**	**0.000**	**116**

Ejemplo 3.3.1. Caracterización de los "demócratas" en función de la combinación de X4 y X5

Podemos observar en el ejemplo 3.3.1., que "cuando está en contra de X4 y a favor de X5" (X4CX5F), la probabilidad es de 0.97 con un nivel muy alto de significación (V-test = 4.59), es decir 99% de certeza. Cabe destacar, que ahora nos interesa para la predicción, ver el porcentaje de la clase en la modalidad y no de la modalidad en la clase, como hacemos para describir al clasificador.

Etiquetas de las variables	modalidades caracterísitcas	Valor-Test	% de la Clase en la modalidad	% de la modalidad en General	% de la modalidad en la Clase	Probabilidad	Peso
Clase	**republicano**	**19.66**	**100.00**	**38.67**	**100.00**	**0.000**	**116**
X4 * X5	X4F-X5F	16.11	92.24	38.67	92.24	0.000	116
X4 * X5	X4FX5C	1.00	66.67	2.00	3.45	0.159	6
X4 * X5	X4CX5D	-1.86	0.00	2.33	0.00	0.031	7
X4 * X5	**X4CX5F**	**-4.59**	**3**	**10**	**1**	**0.000**	**31**
X4 * X5	X4CX5C	-13.43	0.00	43.33	0.00	0.000	130
Clase	**democrata**	**-19.66**	**0.00**	**61.33**	**0.00**	**0.000**	**184**

Ejemplo 3.3.2. Caracterización de los "republicano" en función de la combinación de X4 y X5

Si analizamos lo que pasa con la clase "republicamos" (ejemplo 3.3.2.), observamos que con el mismo nivel de significancia (V-test = -4.59), se puede afirmar que la probabilidad de que sea republicano es de 3% de los casos, muy baja. El V-test negativo me indica justamente que esta no es una característica típica de los republicanos, esto con una certeza del 99%.

Es importante destacar, en este caso, que la probabilidad posteriori que se obtiene con el Demod es aproximada, ya que al combinar ambas variables estamos aportando al modelo una nueva variable que cambia la distribución de probabilidades. Por lo tanto con el Demod (porcentaje de la clase en la modalidad) solo podemos predecir la clase ante una única evidencia de una sola variable, por vez. Cuando las evidencias son más (lo más usual) será necesario calcular el V-test (ecuación 2.3.2.2/5), en base a la probabilidad condicional posteriori de la clase dada las evidencias y la probabilidad a priori de las evidencias, ambos parámetros obtenidos en el modelo bayesiano.

3.4. Red Bayesiana K2 – Asociación entre variables, obtención y validación de hipótesis

Red Bayesiana es una técnica descriptiva estadística que muestra la asociación entre variables. Las inferencias realizadas sobre ella pueden validarse estadísticamente, calculando un valor de prueba en función de las probabilidades.

La aplicación de Redes bayesianas (K2) para el análisis y comprensión de los datos y las relaciones entre las variables es más que aparente, la capacidad descriptiva y explicativa las hace muy adecuadas para analizar las relaciones entre atributos de una manera gráfica y mucho más sofisticada que las reglas de asociación o los estudios correlacionarles.

Con el Demod podremos analizar cada nodo, como si fuera una variable de grupo o de clasificación y describirla indicando las probabilidades ordenadas en función de su importancia (de los V-test).

Tomamos como *ejemplo* la base de datos de "[8]Estudio de las prácticas de lectura del niño y las opiniones en relación con características socioculturales del mismo".

Se observo sobre una muestra de 743 niños, 14 característica (Tabla 3.4.1.).

El objetivo es el estudio de las *relaciones entre las diferentes prácticas de lectura* y *contexto sociocultural de los niños*, para determinar los *factores* que afectan la frecuencia en la lectura de los mismos. ¿Cómo puedo analizar las relaciones entre variables, de manera que facilite la obtención de hipótesis relacionadas a la práctica de lectura de los niños, a partir de las cuales surgirán nuevas investigaciones y estrategias de acción?. ¿Son estadísticamente válidas estas inferencias?

Variables		Valores posibles
Frecuencia Lectura Escuela	"en la escuela leemos"	"poco", "bastante", "mucho".
Libros en Casa	"en casa tenemos"	"pocos libros", "bastantes libros", "muchos libros"
Frecuencia Lectura	"yo leo"	"poco", "bastante", "mucho"
Dificultad Lectura	"leo con"	"mucha dificultad", "alguna dificultad", "facilidad"
Libros dados Maestro	"libros escuela dados por maestro"	"me gustan", "no me gustan",
Cuando Lee	"leo cuando"	"hago trabajo", "me apetece", "las dos cosas"
Cómo Lee	"pref. leer"	"en silencio", "en voz alta", "1+2"
Gusto Textos Escuela	"leer textos escuela"	"no me gustan", "me gustan", "gust. a veces si, a"
Gusto Escuela	"gustar escuela"	"si", "no", "si i no"
Tipo Escuela	"tipo de escuela"	"publica", "privada",
Calificación	"califica. globales"	"suspenso", "suficiente", "bien", "notable", "sobresaliente"
Sexo	"sexo"	"niño", "niña"
Trabajo Padre	"trabajo padre"	"adm. banc. emp.", "funcionario", "pr. liberales", "pr. industria", "agr. gan. min.", "comercio", "trans. comunic", "oficios", "ama de casa", "parado", "jubilado"
Inteligencia Global	"MI inteligencia global-promedio de las otras"	MB[1a13.5], B[13.5a28.5], MeB[28.5a45], MeA[45a65], A[65a83.5], MA[83.5a98]

Tabla 3.4.1. Variables tabla de datos "Práctica de lectura"

[8] Estudio realizado con datos aportados por la Dirección de Evaluación Educativa - Dirección General de Escuelas

Mediante la Red Bayesiana, que se muestra en la figura 3.4.1., podemos observar las *relaciones directas e indirectas entre las variables correspondientes al dominio utilizado.*

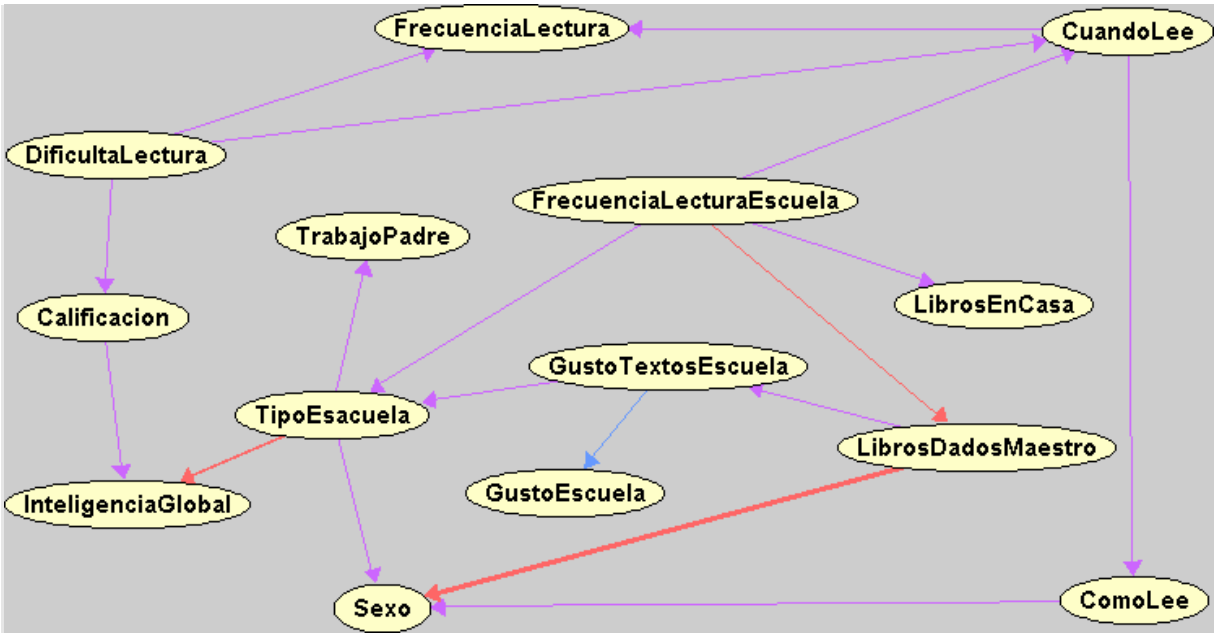


Figura 3.4.1. Red bayesiana K2, sin restricciones. Las influencias **positivas** se colorean en rojo; las **negativas**, en azul; las **nulas,** en negro; y las **indefinidas**, en violeta[9].

Se puede interpretar lo siguiente:

- La calificación global está influenciada por el *grado de dificultad con que lee*, esto a su vez influye en *la frecuencia con qué lee* y con *cuándo lee.*
- *Cuando lee* influye en la *frecuencia con qué lee* y sus *preferencias en cómo lee*; esto último está relacionado con el *sexo*.
- La *calificación global* influye sobre la *inteligencia global*. El *tipo de escuela* (privada, pública) influye sobre los *coeficientes intelectuales* y sobre el *tipo de trabajo* del padre.
- El *gusto de los libros* dados por el maestro influye sobre el *gusto por la escuela* y por el *sexo*. Y depende de la *frecuencia con que leen en la escuela*.
- El *gusto por los textos en la escuela* están relacionadas con el *tipo de escuela* y el *gusta* por esta.

Con el Demod puedo saber, por ejemplo (ejemplo 3.4.1), cuales son las variables que están relacionadas a la frecuencia con que lee, lo cuál se corresponde con lo visto en la red (figura 3.4.1) y se valida con el valor de prueba.

[9] A influye positivamente en su hijo B si a mayores valores de A aumenta la probabilidad de que B tome valores mayores. Influencias indefinidas: si no se puede establecer un orden entre las distribuciones

FrecuenciaLectura ("yo leo")

Etiqueta de la variable	Khi-2	Valor-Test	Probabilidad
FrecuenciaLectura	**1484.00**	**99.99**	**0.000**
DificultaLectura ["leo con"]	106.92	9.62	0.000
CuandoLee	58.18	6.76	0.000
Calificación	45.11	4.96	0.000
LibrosEnCasa	29.33	4.35	0.000
InteligenciaGlobal	40.84	4.22	0.000
LibrosDadosMaestro	17.91	3.65	0.000
GustoEscuela	19.98	3.29	0.001
FrecuenciaLecturaEscuela	15.81	2.72	0.003
GustoTextosEscuela	12.98	2.28	0.011
Sexo	4.66	1.30	0.097
TrabajoPadre	24.90	1.14	0.128
TipoEsacuela	1.14	-0.16	0.565
CómoLee	2.55	-0.34	0.635

Ejemplo 3.4.1. Listado de variables asociadas a "yo leo"

Si se analiza, nodo por nodo, se puede, además de los factores asociados a la frecuencia de lectura, saber mucho más, con lo que podemos sintetizar: *"La dificultad en la lectura y el tener libros en casa es un factor que afecta la frecuencia con que los niños leen y esta depende de la inteligencia que se ve reflejado en las calificaciones. Otro factor es la frecuencia con que leen libros en la escuela, que depende de si le gustan los libros dados por el maestro y del tipo de escuela privada o pública"*.

El problema está en que para llegar a esta conclusión se debe tomar en la red, nodo por nodo, valor, por valor, e indicarlo como evidencia para obtener las inferencias que permitirán describir las proporciones de cada variable; luego será necesario ordenar y dejar lo más relevante en las relaciones, según el nivel de confianza. Un procedimiento rápido, es realizar un Demod de cada variable y utilizar estos listados para escribir el informe con las conclusiones, además de las probabilidades tendrá el nivel de significación estadística. Esto último permitirá que cualquier hipótesis que se obtenga sea conocimiento estadísticamente válido, respecto a la muestra de ejemplos tomada.

Por ejemplo, se ontiene la siguiente hipótesis: *"Los que leen mucho se caracterizan por tener facilidad para la lectura, tiene muchos libros en casa, tienen calificaciones sobresalientes, inteligencia muy alta, leen cuando quieren, leen mucho en la escuela, les gusta la escuela y los libros dados por el maestro"*. ¿Estadísticamente es significativo?, ¿con qué confianza puedo afirmarlo?

En el ejemplo 3.4.2., podemos ver las probabilidades condicionales y el nivel altamente significativo (V-test >2), lo cual valida la hipótesis.

Etiquetas de las variables	modalidades caracterísitcas	% de la modalidad en la Clase	% de la modalidad en General	Valor-Test	Probabilidad	Peso
FrecuenciaLectura	**mucho**	**100.00**	**27.36**	**29.24**	**0.000**	**203**
DificultaLectura	Facilidad	84.73	58.76	9.13	0.000	436
LibrosEnCasa	muchos	79.80	68.60	4.05	0.000	509
Calificación	sobresaliente	19.70	11.73	3.86	0.000	87
InteligenciaGlobal	MA[83.5a98]	20.69	12.94	3.61	0.000	96
CuandoLee	Apetece	59.61	51.48	2.64	0.004	382
LibrosDadosMaestro	Gustan	90.64	85.58	2.36	0.009	635
FrecuenciaLecturaEscuela	mucho	37.93	31.40	2.25	0.012	233
TrabajoPadre	parado	8.37	5.12	2.20	0.014	38
GustoTextosEscuela	Gustan	84.24	78.71	2.20	0.014	584
GustoEscuela	si	90.64	86.39	2.00	0.023	641
TrabajoPadre	funcionario	6.90	4.58	1.62	0.053	34
FrecuenciaLectura	**bastante**	**100.00**	**56.74**	**31.62**	**0.000**	**421**
DificultaLectura	AlgunaDificultad	45.37	38.14	4.60	0.000	283
LibrosEnCasa	bastantes	33.02	27.76	3.61	0.000	206
CuandoLee	Ambas	29.69	26.82	1.94	0.026	199
FrecuenciaLectura	**poco**	**100.00**	**15.90**	**25.19**	**0.000**	**118**
CuandoLee	Trabajo	47.46	21.70	6.79	0.000	161
InteligenciaGlobal	MB[1a13.5]	39.83	22.24	4.65	0.000	165
DificultaLectura	MuchaDificultad	11.02	3.10	4.37	0.000	23
GustoEscuela	no	23.73	11.86	3.89	0.000	88
LibrosDadosMaestro	NoGustan	26.27	14.42	3.62	0.000	107
DificultaLectura	AlgunaDificultad	52.54	38.14	3.37	0.000	283
FrecuenciaLecturaEscuela	poco	18.64	10.78	2.69	0.004	80
LibrosEnCasa	pocos	8.47	3.64	2.55	0.005	27
Calificación	suspenso	12.71	6.60	2.53	0.006	49
GustoTextosEscuela	NoGustan	25.42	17.39	2.31	0.011	129
Sexo	niño	60.17	51.08	2.06	0.020	379
Calificación	bien	36.44	29.11	1.78	0.037	216
Calificación	suficiente	29.66	22.91	1.76	0.040	170

Ejemplo 3.4.2. Demod variable “Frecuencia de lctura”

En el anexo B.2., podemos ver el análisis detallado de cada nodo de la red, con el Demod.

4. PROPUESTA

La propuesta es muy simple, complementar las ventajas gráficas de un modelo bayesiano, con las ventajas del Demod en la ordenación de asociaciones según importancia jerárquica del V-test.

La *metodología de trabajo* es la siguiente:

1. Construir el modelo, con el clasificador bayesiano (en este caso Naïve Bayes)
2. Describir las clases mediante un listado ordenado según importancia (Demod)
3. Clasificar un nuevo caso con el modelo creado. (NB para inferir la probabilidad y cálculo del V-test para validarla).

Caso práctico. Análisis de 80 encuestas para medir los sentimientos que se perciben de los colores (8 colores):

Existe cierta evidencia empírica de que en la predisposición individual una estrategia de relación con uno u otro color, intervienen variables vinculadas a las características psicológicas estables (grado de ansiedad, tipo de personalidad, etc.) de los individuos. La relevancia de la Comunicación Visual, ante la compleja realidad actual, requiere de estudios orientados hacia diversas especialidades. En nuestro enfoque la exploración de los estados de ánimo (EA) de un determinado individuo a partir de su percepción de un color, permite medir los significados connotativos del mismo. La intensidad de las emociones revela cómo piensa la gente que está manejando el medio que la rodea. No obstante la escasa la investigación sobre el rol que de los estados de ánimo individuales, permanentes o transitorios, desempeñan en la elección de uno u otro color, el objetivo principal del presente estudio apunta, precisamente, a "explorar las posibles vinculaciones entre los colores y los estados afectivos, evaluando una tabla normalizada de colores aplicada en un grupo de individuos Normal Tipo, mayores de 18 años de edad de ambos sexos". [Césari R., Correa M. T., 1999]

Una vez modelado este conocimiento mediante un clasificador bayesiano, este modelo, establecerá una herramienta de diseñadores y profesionales que utilicen el efecto del color en el estado de animó para desarrollar productos o proporcionar un servicio más acorde a las necesidades. Sabiendo las emociones que se espera elicitar, se podrá elegir los colores adecuados y significativos, en función de las preediciones hecha sobre el modelo aprendido.

Instrumentos de recolección de datos

Diferencial Semántico. Instrumento de medición del significado Connotativo, adaptado al Estudio del Estado de Animo (EA) - IDDA-EA. Explora la *Depresión, Alegría, Exaltación, Irritabilidad, Confusión, Apatía y Deterioro.* Consta de veintiún pares dobles de adjetivos bipolares separados por siete posibilidades de respuesta, evaluando la reacción frente a seis colores de la tabla normalizada, además el blanco y negro. El Diferencial Semántico adaptado, permite evaluar cada una de las dimensiones correspondientes a los estados de ánimo licitados por éstos colores.

Tabla de Colores Normalizada. [10]Se compone de los colores primarios y secundarios definidos: ROJO, VERDE, AZUL-VIOLETA, MAGENTA, CIAN, AMARILLO, y además, el NEGRO y el BLANCO.

Mediciones:

Se evaluaron 80 personas normales tipo definidas por su disposición voluntaria de contestar los inventarios, (se han considerado estudiantes de la Universidad)

Con datos obtenidos del Diferencial Semántico, se medio la connotación de los adjetivos asociados al estado de ánimo para cada color. Los valores cuantitativos de las escalas de adjetivos, se "Discretizan" en clases, transformando a esos valores, en variables cualitativas (nominales), en siete modalidades cada una.

El criterio de Discretización lo vemos en la tabla 4.1.:

REFERENCIAS PARA LA INTERPRETACIÓN de LOS RESULTADOS

	ALTO+	ALTO-	MEDIO+	nsnc	MEDIO-	BAJO-	BAJO+
	7	6	5	4	3	2	1
	Alta connotación de	*Media connotación de*	*Baja connotación de*		*Baja connotación de*	*Media connotación de*	*Alta connotación de*
X1	Iniciativa / Interés			INDIFERENTE	Apatía / Desgano		
X2	Pudor / Recato				Lujuria / Exhibicionismo		
X3	Nobleza / Amor				Rencor / Odio		
X4	Fortaleza / Audacia				Timidez / Debilidad		
X5	Calma / Descanso				Agresividad / Irritabilidad		
X6	Fortaleza / Poder				Debilidad / Indefensión		
X7	Justicia				Injusticia		
X8	Diversión / Placer				Aburrimiento / Amargura		
X9	Sumisión / Docilidad				Rebeldía / Soberbia		
X10	En Paz / En Armonía				Culpa / Falta		
X11	Confianza / Certeza				Sospecha / Duda		
X12	Control / Equilibrio				Descontrol / Desequilibrio		
X13	Alegría / Seguridad				Tristeza / Desamparo		
X14	Calma / Seguridad				Agitación / Pánico		
X15	Vida / Satisfacción				Muerte / Hartazgo		
X16	Éxito / Optimismo				Pesimismo / Fracaso		
X17	Tranquilidad / Serenidad				Ansiedad / Angustia		
X18	Dicha / Esperanza				Desdicha / Desesperanza		
X19	Creación / Serenidad				Destrucción / Furia		
X20	Gozo / Placer				Malestar / Displacer		
X21	Reposo / Serenidad				Excitación / Exaltación		

Tabla 4.1. Referencia de las 21 variables cualitativas medidas para 8 colores.

La tabla de análisis queda estructurada de la siguiente manera: *21 variables nominales de 7 valores posibles* y *"80 (encuestados) x 8 (colores)" casos de ejemplos (648 observaciones).*

[10] Estudio Integral del Color, CESARI, CORREA y Otros. Laboratorio Color. Facultad de Artes. UNCuyo 1995

4.1. Clasificador bayesiano. Probabilidades, gráfico

En primer lugar se construye el modelo probabilístico a través del clasificador bayesiano Naïve Bayes. El mismo se aplicará para clasificar nuevos casos en base a las evidencias. Usamos el programa Elvira, para crear la red clasificadora, la misma podemos verla en la figura 4.1.1.

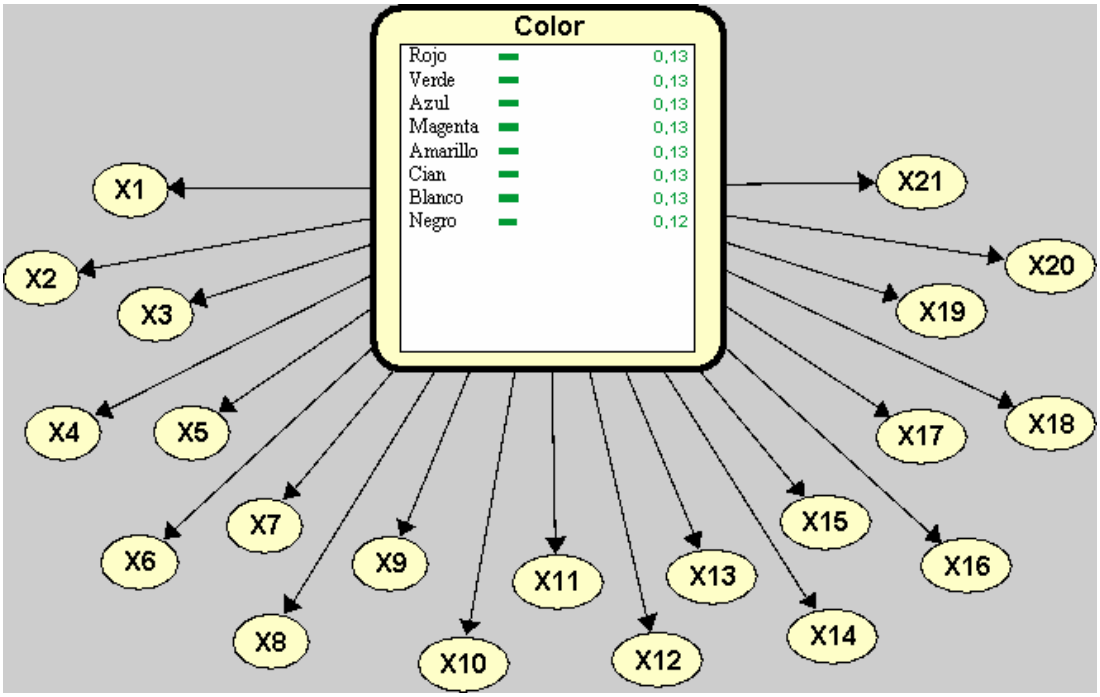


Figura 4.1.1 Modelo NB del Color

4.2. Descripción óptima del modelo

Mediante la aplicación del Demod, se genera un listado ordenado por color, indicando las proporciones de casos (de cada color) que están asociados a una connotación de un par de adjetivos.

El criterio de ordenación es el V-test (epígrafe 2.3.2.2), que determina la significación estadística de las relaciones entre las variables y el clasificador (Color), es decir mide la calidad de los datos de entrenamiento con qué construimos el modelo para clasificar.

Teniendo en cuenta epígrafe 3.1., el porcentaje de casos de una modalidad en la clase (% Modalidad en la clase), es la proporción de casos de una característica dada la clase (probabilidad condicional posteriori) El porcentaje total de la modalidad, determina la probabilidad a priori de la característica.

A continuación, se presentan los listados que proporcionan la descripción óptima de cada color. (En el anexo C, se puede ver el análisis de la resonancia afectiva de los colores en función de los adjetivos que connotan)

Color ROJO

	connotación de adjetivos	P(X | Color)	P(X)	P(Color | X)	Signif.	error	N
Etiquetas de las variables	**modalidades caracterísitcas**	**% de la modalidad en la Clase**	**% de la modalidad en General**	**% de la Clase en la modalidad**	**Valor-Test**	**Probabilidad**	**Peso**
Color	**ROJO**	**100.00**	**12.50**	**100.00**	**21.75**	**0.000**	**81**
X2	Bajo+ Lujuria / Exhibicionismo	50.62	13.27	47.67	8.86	0.000	86
X21	Bajo+ Excitación / Exaltación	46.91	17.90	32.76	6.43	0.000	116
X5	Bajo+ Agresividad / Irritabilidad	37.04	11.88	38.96	6.33	0.000	77
X8	Alto+ Diversión / Placer	45.68	17.90	31.90	6.17	0.000	116
X21	Bajo- Excitación / Exaltación	34.57	11.42	37.84	5.93	0.000	74
X14	Bajo- Agitación / Pánico	37.04	14.35	32.26	5.42	0.000	93
X20	Alto+ Gozo / Placer	37.04	15.12	30.61	5.16	0.000	98
X19	Bajo+ Destrucción / Furia	29.63	11.11	33.33	4.86	0.000	72
X4	Alto+ Fortaleza / Audacia	40.74	18.83	27.05	4.84	0.000	122
X17	Bajo- Ansiedad / Angustia	29.63	12.19	30.38	4.45	0.000	79
X6	Alto+ Fortaleza / Poder	37.04	17.44	26.55	4.44	0.000	113
X19	Bajo- Destrucción / Furia	24.69	9.26	33.33	4.35	0.000	60
X10	NsNc En Paz / Falta	40.74	21.30	23.91	4.16	0.000	138
X17	Medio- Ansiedad / Angustia	30.86	13.89	27.78	4.16	0.000	90
X11	NsNc Confianza / Duda	41.98	22.69	23.13	4.06	0.000	147
X10	Medio- Culpa / Falta	24.69	10.19	30.30	3.97	0.000	66
X12	Medio- Descontrol / Desequilibrio	25.93	11.73	27.63	3.70	0.000	76
X5	Bajo- Agresividad / Irritabilidad	28.40	13.58	26.14	3.67	0.000	88
X5	Medio- Agresividad / Irritabilidad	24.69	11.27	27.40	3.56	0.000	73
X6	Alto- Fortaleza / Poder	33.33	18.06	23.08	3.45	0.000	117
X12	Bajo- Descontrol / Desequilibrio	23.46	10.80	27.14	3.41	0.000	70
X2	Bajo- Lujuria / Exhibicionismo	23.46	10.80	27.14	3.41	0.000	70
X19	Medio- Destrucción / Furia	20.99	9.26	28.33	3.35	0.000	60
X1	Alto- Iniciativa / Interés	37.04	22.22	20.83	3.14	0.001	144
X18	NsNc Dicha / Desesperanza	40.74	25.46	20.00	3.12	0.001	165
X1	Alto+ Iniciativa / Interés	37.04	22.69	20.41	3.03	0.001	147
X8	Alto- Diversión / Placer	30.86	17.90	21.55	2.94	0.002	116
X4	Alto- Fortaleza / Audacia	28.40	16.05	22.12	2.90	0.002	104
X14	Bajo+ Agitación / Pánico	20.99	10.80	24.29	2.76	0.003	70
X14	Medio- Agitación / Pánico	22.22	11.73	23.68	2.76	0.003	76
X16	Alto- Éxito / Optimismo	32.10	20.68	19.40	2.47	0.007	134
X3	Alto+ Nobleza / Amor	32.10	20.99	19.12	2.39	0.008	136
X11	Medio- Sospecha / Duda	19.75	10.96	22.54	2.37	0.009	71

Figura 4.2.1. Demod color "rojo"

Alta connotación de Lujuria / Exhibicionismo, Excitación / Exaltación, Agresividad / Irritabilidad, Diversión / Placer, Gozo / Placer, Destrucción / Furia, Fortaleza / Audacia, Fortaleza / Poder, etc.

Media connotación de Excitación / Exaltación, Agitación / Pánico, Ansiedad / Angustia, Destrucción / Furia, Agresividad / Irritabilidad, Fortaleza / Poder, Descontrol / Desequilibrio, Lujuria / Exhibicionismo, etc.

Baja connotación de Ansiedad / Angustia, Culpa / Falta, Descontrol / Desequilibrio, Agresividad / Irritabilidad, Destrucción / Furia, etc.

Indiferente connotación de En Paz / Falta y de Confianza / Duda, etc.

Color VERDE

Etiquetas de las variables	modalidades caracterísitcas	% de la modalidad en la Clase	% de la modalidad en General	% de la Clase en la modalidad	Valor-Test	Probabilidad	Peso
Color	Verde	100.00	12.50	100.00	21.75	0.000	81
X18	Alto+ Dicha / Esperanza	60.49	20.52	36.84	8.45	0.000	133
X15	Alto+ Vida / Satisfacción	51.85	25.00	25.93	5.48	0.000	162
X12	Alto+ Control / Equilibrio	44.44	20.22	27.48	5.23	0.000	131
X17	Alto+ Tranquilidad / Serenidad	40.74	18.21	27.97	5.02	0.000	118
X19	Alto+ Creación / Serenidad	40.74	18.36	27.73	4.98	0.000	119
X14	Alto- Calma / Seguridad	40.74	18.67	27.27	4.89	0.000	121
X5	Alto+ Calma / Descanso	41.98	20.52	25.56	4.63	0.000	133
X10	Alto- En Paz / En Armonía	34.57	15.90	27.18	4.36	0.000	103
X10	Alto+ En Paz / En Armonía	44.44	24.23	22.93	4.18	0.000	157
X17	Alto- Tranquilidad / Serenidad	35.80	17.44	25.66	4.17	0.000	113
X21	Alto- Reposo / Serenidad	33.33	16.82	24.77	3.81	0.000	109
X21	Alto+ Reposo / Serenidad	35.80	19.14	23.39	3.69	0.000	124
X11	Alto+ Confianza / Certeza	33.33	17.44	23.89	3.63	0.000	113
X19	Alto- Creación / Serenidad	35.80	19.75	22.66	3.52	0.000	128
X6	NsNc Fortaleza / Indefensión	43.21	26.23	20.59	3.44	0.000	170
X3	Alto- Nobleza / Amor	34.57	19.75	21.88	3.25	0.001	128
X5	Alto- Calma / Descanso	30.86	17.28	22.32	3.11	0.001	112
X11	Alto- Confianza / Certeza	29.63	16.82	22.02	2.97	0.002	109
X2	Alto- Pudor / Recato	24.69	13.27	23.26	2.87	0.002	86
X14	Alto+ Calma / Seguridad	32.10	19.60	20.47	2.75	0.003	127
X2	NsNc Pudor / Exhibicionismo	37.04	24.07	19.23	2.69	0.004	156
X8	Medio+ Diversión / Placer	24.69	13.89	22.22	2.67	0.004	90
X7	Alto+ Justicia	32.10	20.37	19.70	2.55	0.005	132
X13	Alto- Alegría / Seguridad	29.63	18.52	20.00	2.49	0.006	120
X3	Medio+ Nobleza / Amor	24.69	14.97	20.62	2.34	0.010	97

Figura 4.2.2. Demod color "verde"

Alta connotación de Dicha / Esperanza, Vida / Satisfacción, Control / Equilibrio, Tranquilidad / Serenidad, Creación / Serenidad, Calma / Descanso, En Paz / En Armonía Reposo / Serenidad, Confianza / Certeza, Calma / Seguridad y Justicia.

Media connotación de Calma / Seguridad, En Paz / En Armonía, Tranquilidad / Serenidad, Reposo / Serenidad, Creación / Serenidad, Nobleza / Amor, Calma / Descanso, Confianza / Certeza, Pudor / Recato y Alegría / Seguridad.

Baja connotación de Diversión / Placer y Nobleza / Amor.

Indiferente connotación de Pudor / Exhibicionismo y Fortaleza / Indefensión.

Color AZUL

Etiquetas de las variables	modalidades caracterísitcas	% de la modalidad en la Clase	% de la modalidad en General	% de la Clase en la modalidad	Valor-Test	Probabilidad	Peso
Color	Azul	100.00	12.50	100.00	21.75	0.000	81
X15	NsNc Vida / Muerte	39.51	18.98	26.02	4.53	0.000	123
X21	Medio+ Reposo / Serenidad	24.69	11.42	27.03	3.50	0.000	74
X12	Alto- Control / Equilibrio	29.63	15.74	23.53	3.28	0.001	102
X9	NsNc Sumisión / Soberbia	38.27	23.30	20.53	3.13	0.001	151
X7	Alto- Justicia	19.75	9.88	25.00	2.77	0.003	64
X14	Alto- Calma / Seguridad	30.86	18.67	20.66	2.73	0.003	121
X20	Medio- Malestar / Displacer	16.05	7.41	27.08	2.70	0.003	48
X21	Alto- Reposo / Serenidad	28.40	16.82	21.10	2.68	0.004	109
X17	Medio+ Tranquilidad / Serenidad	22.22	12.19	22.78	2.60	0.005	79
X5	Medio+ Calma / Descanso	20.99	11.27	23.29	2.60	0.005	73
X5	Alto- Calma / Descanso	28.40	17.28	20.54	2.55	0.005	112
X8	Bajo- Aburrimiento / Amargura	19.75	10.80	22.86	2.43	0.008	70

Figura 4.2.3. Demod color "azul"

Indiferente connotación de Vida / Muerte y Sumisión / Soberbia.

Baja connotación de Reposo / Serenidad, Malestar / Displacer, Tranquilidad / Serenidad y Calma / Descanso.

Alta connotación de Control / Equilibrio, Justicia, Calma / Seguridad, Reposo / Serenidad y Calma / Descanso

Media connotación de Aburrimiento / Amargura

Color AMARILLO

Etiquetas de las variables	modalidades caracterísitcas	% de la modalidad en la Clase	% de la modalidad en General	% de la Clase en la modalidad	Valor-Test	Probabilidad	Peso
Color	Amarillo	100.00	12.50	100.00	21.75	0.000	81
X19	Medio+ Creación / Serenidad	22.22	12.65	21.95	2.45	0.007	82
X4	Medio- Timidez / Debilidad	17.28	9.10	23.73	2.37	0.009	59
X3	Medio+ Nobleza / Amor	24.69	14.97	20.62	2.34	0.010	97

Figura 4.2.4. Demod color "amarillo"

Baja connotación de Creación / Serenidad, Timidez / Debilidad y Nobleza / amor

Color MAGENTA

Etiquetas de las variables	modalidades caracterísitcas	% de la modalidad en la Clase	% de la modalidad en General	% de la Clase en la modalidad	Valor-Test	Probabilidad	Peso
Color	Magenta	100.00	12.50	100.00	21.75	0.000	81
X10	NsNc En Paz / Falta	49.38	21.30	28.99	5.96	0.000	138
X7	NsNc Justicia / Injusticia	66.67	38.12	21.86	5.45	0.000	247
X21	Bajo+ Excitación / Exaltación	37.04	17.90	25.86	4.31	0.000	116
X2	Bajo- Lujuria / Exhibicionismo	25.93	10.80	30.00	4.05	0.000	70
X17	Medio- Ansiedad / Angustia	29.63	13.89	26.67	3.86	0.000	90
X12	Bajo- Descontrol / Desequilibrio	24.69	10.80	28.57	3.73	0.000	70
X21	Bajo- Excitación / Exaltación	24.69	11.42	27.03	3.50	0.000	74
X8	Alto+ Diversión / Placer	33.33	17.90	23.28	3.49	0.000	116
X6	Medio+ Fortaleza / Poder	28.40	14.20	25.00	3.47	0.000	92
X14	Bajo+ Agitación / Pánico	23.46	10.80	27.14	3.41	0.000	70
X19	NsNc Creación / Serenidad	34.57	19.60	22.05	3.30	0.000	127
X4	Medio+ Fortaleza / Audacia	25.93	13.89	23.33	2.98	0.001	90
X8	Alto- Diversión / Placer	30.86	17.90	21.55	2.94	0.002	116
X13	Alto- Alegría / Seguridad	30.86	18.52	20.83	2.77	0.003	120
X2	Medio- Lujuria / Exhibicionismo	19.75	9.88	25.00	2.77	0.003	64
X12	NsNc Control / Desequilibrio	35.80	22.69	19.73	2.77	0.003	147
X11	Medio- Sospecha / Duda	20.99	10.96	23.94	2.70	0.003	71
X19	Bajo- Destrucción / Furia	18.52	9.26	25.00	2.66	0.004	60
X20	Alto- Gozo / Placer	32.10	20.06	20.00	2.63	0.004	130
X5	NsNc Calma / Irritabilidad	24.69	14.20	21.74	2.58	0.005	92
X17	NsNc Tranquilidad / Angustia	28.40	17.28	20.54	2.55	0.005	112
X5	Bajo- Agresividad / Irritabilidad	23.46	13.58	21.59	2.47	0.007	88
X9	Alto- Sumisión / Docilidad	24.69	14.97	20.62	2.34	0.010	97

Figura 4.2.5. Demod color “magenta”

Alta connotación de Excitación / Exaltación, Diversión / Placer y Agitación / Pánico.

Media connotación de Lujuria / Exhibicionismo, Descontrol / Desequilibrio, Excitación / Exaltación, Diversión / Placer, Alegría / Seguridad, Destrucción / Furia y Gozo / Placer

Baja connotación de Ansiedad / Angustia, Fortaleza / Poder, Fortaleza / Audacia, Lujuria / Exhibicionismo y Sospecha / Duda.

Indiferente connotación de En Paz / Falta, Justicia / Injusticia, Creación / Serenidad, Control / Desequilibrio y Calma / Irritabilidad.

Color CIAN

Etiquetas de las variables	modalidades caracterísitcas	% de la modalidad en la Clase	% de la modalidad en General	% de la Clase en la modalidad	Valor-Test	Probabilidad	Peso
Color	**Cian**	**100.00**	**12.50**	**100.00**	**21.75**	**0.000**	**81**
X11	Medio+ Confianza / Certeza	27.16	13.27	25.58	3.48	0.000	86
X14	Alto+ Calma / Seguridad	34.57	19.60	22.05	3.30	0.000	127
X1	NsNc Iniciativa / Desgano	29.63	15.74	23.53	3.28	0.001	102
X3	Alto- Nobleza / Amor	34.57	19.75	21.88	3.25	0.001	128
X4	Bajo+ Timidez / Debilidad	23.46	11.73	25.00	3.08	0.001	76
X18	Medio+ Dicha / Esperanza	30.86	17.44	22.12	3.07	0.001	113
X14	Medio+ Calma / Seguridad	22.22	11.11	25.00	2.98	0.001	72
X10	Alto- En Paz / En Armonía	28.40	15.90	22.33	2.95	0.002	103
X17	Alto+ Tranquilidad / Serenidad	30.86	18.21	21.19	2.86	0.002	118
X5	Alto- Calma / Descanso	29.63	17.28	21.43	2.83	0.002	112
X10	Medio+ En Paz / En Armonía	22.22	11.73	23.68	2.76	0.003	76
X7	Medio+ Justicia	20.99	10.96	23.94	2.70	0.003	71
X6	Bajo- Debilidad / Indefensión	16.05	7.56	26.53	2.63	0.004	49
X5	Medio+ Calma / Descanso	20.99	11.27	23.29	2.60	0.005	73
X19	Alto+ Creación / Serenidad	29.63	18.36	20.17	2.53	0.006	119
X19	Medio+ Creación / Serenidad	22.22	12.65	21.95	2.45	0.007	82
X12	Alto- Control / Equilibrio	25.93	15.74	20.59	2.41	0.008	102
X21	Alto- Reposo / Serenidad	27.16	16.82	20.18	2.39	0.008	109
X21	Alto+ Reposo / Serenidad	29.63	19.14	19.35	2.33	0.010	124

Figura 4.2.6. Demod color "cian"

Alta connotación de Calma / Seguridad, Creación / Serenidad, Reposo / Serenidad, Timidez / Debilidad y Tranquilidad / Serenidad.

Media connotación de Calma / Descanso, Control / Equilibrio, Debilidad / Indefensión, En Paz / En Armonía, Nobleza / Amor y Reposo / Serenidad.

Baja connotación de Calma / Descanso, Calma / Seguridad, Confianza / Certeza, Creación / Serenidad, Dicha / Esperanza, En Paz / En Armonía y Justicia.

Indiferente connotación de Iniciativa / Desgano.

Color BLANCO

Etiquetas de las variables	modalidades caracterísitcas	% de la modalidad en la Clase	% de la modalidad en General	% de la Clase en la modalidad	Valor-Test	Probabilidad	Peso
Color	**Blanco**	**100.00**	**12.50**	**100.00**	**21.75**	**0.000**	**81**
X5	Alto+ Calma / Descanso	59.26	20.52	36.09	8.20	0.000	133
X10	Alto+ En Paz / En Armonía	64.20	24.23	33.12	8.16	0.000	157
X2	Alto+ Pudor / Recato	50.62	16.20	39.05	7.76	0.000	105
X7	Alto+ Justicia	53.09	20.37	32.58	6.97	0.000	132
X12	Alto+ Control / Equilibrio	50.62	20.22	31.30	6.51	0.000	131
X21	Alto+ Reposo / Serenidad	45.68	19.14	29.84	5.80	0.000	124
X3	Alto+ Nobleza / Amor	48.15	20.99	28.68	5.79	0.000	136
X14	Alto+ Calma / Seguridad	43.21	19.60	27.56	5.15	0.000	127
X4	Bajo+ Timidez / Debilidad	30.86	11.73	32.89	4.92	0.000	76
X11	Alto+ Confianza / Certeza	38.27	17.44	27.43	4.71	0.000	113
X19	Alto+ Creación / Serenidad	39.51	18.36	26.89	4.71	0.000	119
X8	NsNc Diversión / Amargura	40.74	20.83	24.44	4.29	0.000	135
X17	Alto+ Tranquilidad / Serenidad	37.04	18.21	25.42	4.22	0.000	118
X20	NsNc Gozo / Displacer	46.91	27.62	21.23	3.86	0.000	179
X14	Alto- Calma / Seguridad	34.57	18.67	23.14	3.55	0.000	121
X6	Bajo+ Debilidad / Indefensión	18.52	7.41	31.25	3.44	0.000	48
X19	Alto- Creación / Serenidad	33.33	19.75	21.09	2.99	0.001	128
X17	Alto- Tranquilidad / Serenidad	28.40	17.44	20.35	2.51	0.006	113
X9	Bajo+ Rebeldía / Soberbia	19.75	10.65	23.19	2.48	0.006	69

Figura 4.2.7. Demod color “blanco”

Alta connotación de *Calma / Descanso, Calma / Seguridad, Confianza / Certeza, Control / Equilibrio, Creación / Serenidad, Debilidad / Indefensión, En Paz / En Armonía, Justicia, Nobleza / Amor, Pudor / Recato, Rebeldía / Soberbia, Reposo / Serenidad, Timidez / Debilidad y Tranquilidad / Serenidad.*

Media connotación de *Calma / Seguridad, Creación / Serenidad y Tranquilidad / Serenidad.*

Indiferente connotación de *Diversión / Amargura y Gozo / Displacer*

Color NEGRO

Etiquetas de las variables	modalidades caracterísitcas	% de la modalidad en la Clase	% de la modalidad en General	% de la Clase en la modalidad	Valor-Test	Probabilidad	Peso
Color	**Negro**	**100.00**	**12.50**	**100.00**	**21.75**	**0.000**	**81**
X15	Bajo+ Muerte / Hartazgo	66.67	12.04	69.23	13.07	0.000	78
X16	Bajo+ Pesimismo / Fracaso	40.74	6.48	78.57	10.27	0.000	42
X7	Bajo+ Injusticia	49.38	10.19	60.61	10.02	0.000	66
X18	Bajo+ Desdicha / Desesperanza	40.74	6.94	73.33	9.89	0.000	45
X3	Bajo+ Rencor / Odio	41.98	7.87	66.67	9.55	0.000	51
X19	Bajo+ Destrucción / Furia	46.91	11.11	52.78	8.97	0.000	72
X20	Bajo+ Malestar / Displacer	33.33	6.48	64.29	8.17	0.000	42
X10	Bajo+ Culpa / Falta	32.10	6.33	63.41	7.93	0.000	41
X13	Bajo+ Tristeza / Desamparo	39.51	10.03	49.23	7.74	0.000	65
X1	Bajo+ Apatía / Desgano	39.51	10.96	45.07	7.31	0.000	71
X11	Bajo+ Sospecha / Duda	30.86	7.56	51.02	6.81	0.000	49
X10	Bajo- Culpa / Falta	34.57	10.34	41.79	6.39	0.000	67
X11	Bajo- Sospecha / Duda	35.80	11.27	39.73	6.29	0.000	73
X17	Bajo+ Ansiedad / Angustia	30.86	8.80	43.86	6.17	0.000	57
X16	Bajo- Pesimismo / Fracaso	27.16	7.41	45.83	5.90	0.000	48
X3	Bajo- Rencor / Odio	27.16	7.72	44.00	5.73	0.000	50
X20	Bajo- Malestar / Displacer	24.69	6.48	47.62	5.72	0.000	42
X8	Bajo+ Aburrimiento / Amargura	27.16	8.02	42.31	5.57	0.000	52
X21	NsNc Reposo / Exaltación	38.27	16.51	28.97	5.00	0.000	107
X18	Medio- Desdicha / Desesperanza	22.22	7.25	38.30	4.59	0.000	47
X12	Bajo+ Descontrol / Desequilibrio	22.22	7.56	36.73	4.44	0.000	49
X13	Bajo- Tristeza / Desamparo	22.22	7.56	36.73	4.44	0.000	49
X8	Bajo- Aburrimiento / Amargura	25.93	10.80	30.00	4.05	0.000	70
X7	Bajo- Injusticia	16.05	5.56	36.11	3.62	0.000	36
X14	Bajo+ Agitación / Pánico	23.46	10.80	27.14	3.41	0.000	70
X14	Medio- Agitación / Pánico	24.69	11.73	26.32	3.39	0.000	76
X5	Bajo+ Agresividad / Irritabilidad	24.69	11.88	25.97	3.34	0.000	77
X4	NsNc Fortaleza / Debilidad	33.33	18.83	22.13	3.24	0.001	122
X12	Medio- Descontrol / Desequilibrio	23.46	11.73	25.00	3.08	0.001	76
X6	Alto+ Fortaleza / Poder	29.63	17.44	21.24	2.79	0.003	113
X19	Medio- Destrucción / Furia	18.52	9.26	25.00	2.66	0.004	60
X19	Bajo- Destrucción / Furia	18.52	9.26	25.00	2.66	0.004	60
X5	NsNc Calma / Irritabilidad	24.69	14.20	21.74	2.58	0.005	92
X9	Bajo+ Rebeldía / Soberbia	19.75	10.65	23.19	2.48	0.006	69
X5	Bajo- Agresividad / Irritabilidad	23.46	13.58	21.59	2.47	0.007	88
X18	Bajo- Desdicha / Desesperanza	13.58	6.48	26.19	2.34	0.010	42

Figura 4.2.8. Demod color "negro"

Alta connotación de *Rencor / Odio, Aburrimiento / Amargura, Agitación / Pánico, Agresividad / Irritabilidad, Ansiedad / Angustia, Apatía / Desgano, Culpa / Falta, Descontrol / Desequilibrio, Desdicha / Desesperanza, Destrucción / Furia, Fortaleza / Poder, Injusticia, Malestar / Displacer, Muerte / Hartazgo, Pesimismo / Fracaso, Sospecha / Duda y Tristeza / Desamparo.*

Media connotación de *Aburrimiento / Amargura, Culpa / Falta, Injusticia, Malestar / Displacer, Pesimismo / Fracaso, Rencor / Odio y Sospecha / Duda.*

Indiferente connotación de *Fortaleza / Debilidad y Reposo / Exaltación*

4.3. Predicción de la clase para nuevos casos y nivel de significación de la inferencia.

Se necesita elegir un color que elicite un estado de ánimo "competitivo", por ejemplo para elegir el color de una remera deportiva para un equipo. Se utiliza el modelo bayesiano construido en el apartado 4.1., para determinar el color y la probabilidad ante las siguientes evidencias:

Evidencias propuestas: connotación "baja" (medio-) de *destrucción / furia* (X19) y connotación alta (alto+) de *fortaleza / audacia* (X4)

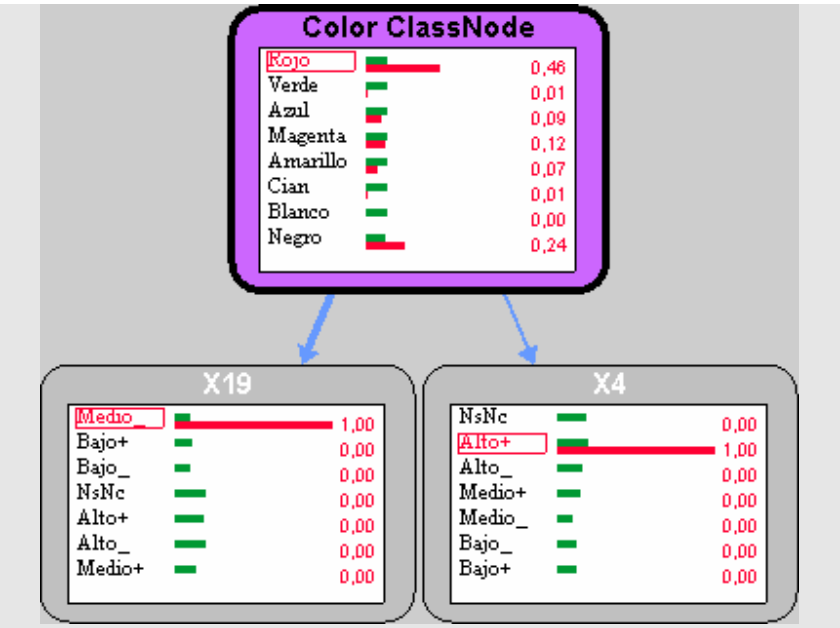


Figura 4.3.1. NB - Inferencia del color ante la evidencia 1

Validación estadística mediante el V-test (ecuación 2.3.2.3 / 4 / 5)

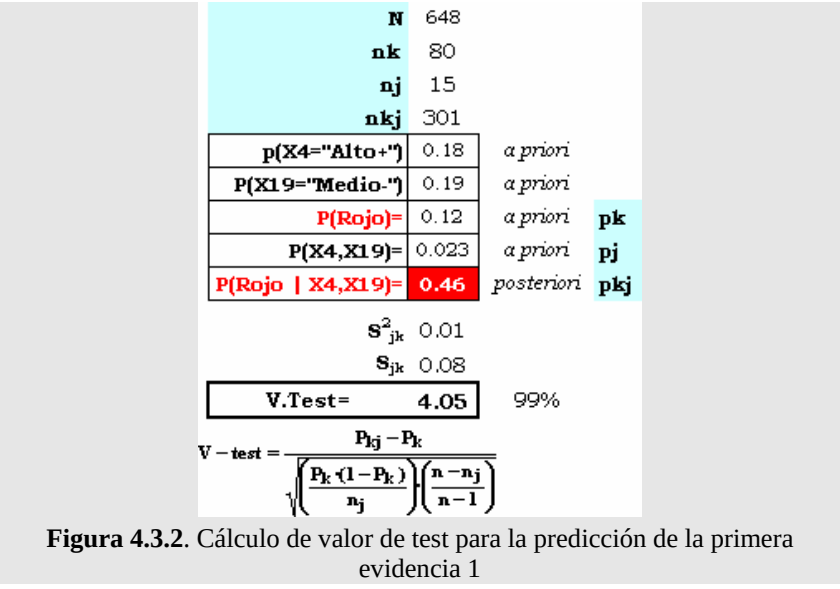


Figura 4.3.2. Cálculo de valor de test para la predicción de la primera evidencia 1

Dada las evidencias, se puede decir, con una probabilidad del 46% (figura 4.3.1.), que el color conveniente es el rojo (la gama de los rojos), esto con un 99% de certeza (figura 4.3.2)

5. CONCLUSIONES

Después de los análisis que se realizaron en el capítulo 3, podemos sintetizar las siguientes conclusiones:

- *Demod es un clasificador bayesiano*, no grafico, que ordena la distribución de probabilidades según nivel de significación estadadística.

  Esta técnica permite asociar al V-test, la *proporción de casos de ejemplo que co-ocurren, para una clase dada, en una característica*, y el *porcentaje de casos en total presentes en la clase* y el *total presente en la característica;* cuando las variables explicativas o descriptivas de la clase son nominales *cualitativas.*

  También asocia la *media y desviación estimada condicional* de la clase. Proporciona la *media y desviación típica estimada* que presentan las características en una clase dada, los mismos, ordenados según su significancia estadística, (nivel de certeza), tanto para los valores más altos de la variable explicativa numérica, cómo para sus valores bajos (V-test positivo o negativo); cuando las variables explicativas o descriptivas de la clase son cuantitativas *numéricas.*

- Los *modelos bayesianos* calculan el costo esperado asociado a cada una de las decisiones posibles, utilizando las probabilidades a posteriori y adoptan la decisión que tenga el menor costo o la mayor utilidad posible. *¿Qué tan significativa es esta utilidad para ser tenida en cuenta?* Calculamos el V-test para cada característica, en relación a la clase y los ordenamos en función de este nivel de significación, de esta manera podemos determinar cuáles son las proporciones representadas en el modelo bayesiano, que realmente determinan una *relación de dependencia* (positiva o negativa) con la clase, y así caracterizar y validar estadísticamente el modelo de aprendizaje.

  Para analizar e interpretar las inferencias obtenidas del modelo bayesiano, es necesario estudiar las probabilidades a priori (globales) y posteriori (condicionales); suele cometerse errores en la obtención de conclusiones a partir de estos resultados, por lo que no se miran ambas probabilidades en conjunto, cuando comparo proporciones de diferentes variables (para una clase dada), debo tener en cuenta la proporción total de las características y sólo los más experimentados logran hacerlo con facilidad. El V-test me permite obtener inferencias relevantes, con mucha facilidad sin necesidad de mucha experiencia y uniendo la información de ambas proporciones en una medida o valor de prueba que simplifica y valida el proceso de educción de conocimiento sobre el modelo bayesiano.

- Las *capacidades predictivas de las redes bayesianas* están orientadas a pronosticar el valor de cualquiera de las variables pertenecientes al dominio de aplicación, en lugar de intentar *maximizar el poder clasificatorio*. Hemos visto, con el Demod, cómo podemos *seleccionar las variables que dependen directamente en la clasificación*, tomando aquellas que están más relacionadas o depende directamente de la clase y eliminado las que no.

  Es de rescatar que sólo el paquete informático que implementa la técnica Demod y los conceptos del V-test, es el programa SPAD [SPAD, 2002], por lo menos que se conoce hasta el momento y ha difundido información al respecto. Esto crea una dependencia con el mismo que hace que sea difícil acceder a estos algoritmos, por ser un programa comercial de mucho valor. Queda abierta la posibilidad de incorporar estos métodos en las herramientas inteligentes de acceso libre a los investigadores.

- *Naïve Bayes es un clasificador bayesiano*, gráficamente caracteriza los grupos homogéneos indicando la intensidad con que las variables se asocian con la clase y provee la base para calcular un *valor de prueba que determinará el nivel de significación estadística* de estas asociaciones, lo que permite posteriormente, validar la predicción de nuevas clasificaciones, realizadas sobre este modelo.

  Para validar estadísticamente las predicciones hechas con un clasificador, así como cualquier inferencia realizada sobre una red bayesiana, calculamos el V-test sobre la base de la probabilidad condicional de "dada la evidencia se de la característica o la clase". Con el Demod sólo podemos validar las predicciones sobre una única evidencia, cuando las evidencias son de más de una variable, usamos la información del modelo para calcular a mano el V-test.

- *Red Bayesiana es una técnica descriptiva estadística que muestra la asociación entre variables.* Las inferencias realizadas sobre ella pueden validarse estadísticamente, calculando un valor de prueba para validar la distribución de las probabilidades.

Resumiendo:

El V-test nos permite *validar las predicciones que realizamos en la clasificación*, determinando si la probabilidad de (ante las evidencias) pertenezca a una clase no se al azar (en base a los datos de aprendizaje). Lo mismo sobre cualquier inferencia que realicemos sobre una red Bayesiana.

Este valor constituye una medida de la calidad de las relaciones representadas en un modelo bayesiano, es decir la confianza de los datos de aprendizaje. En consecuencia valida cualquier inferencia que se realice sobre el modelo.

Esta medida constituye una herramienta importante de nexo entre las técnicas clásicas (estadística multivariada) y las técnicas inteligentes (data mining), para la explotación de datos.

6. FUTURAS LÍNEAS DE INVESTIGACIÓN

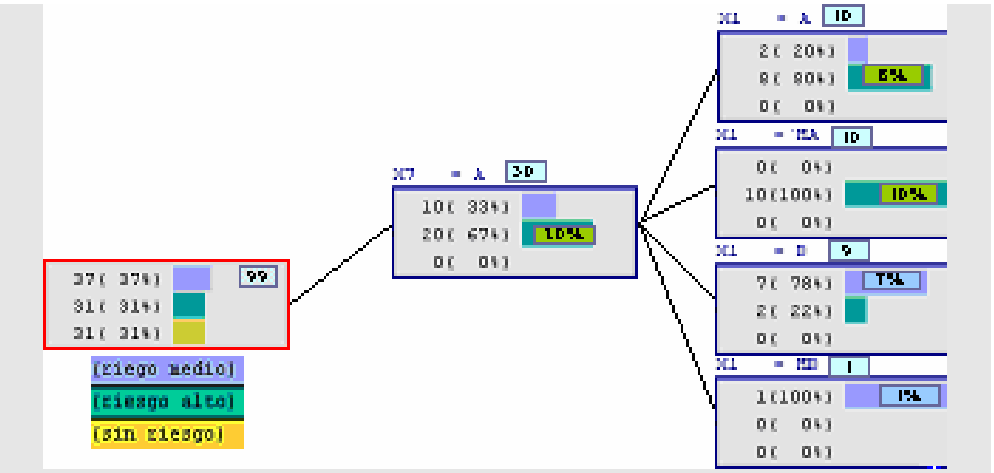
6.1. Nivel de significación estadística para las reglas

El uso de reglas es una de las formas más popular de representación del conocimiento debido, entre otras razones, a su sencillez, capacidad de expresión y escalabilidad. Dependiendo de la naturaleza del conocimiento que almacenan, se ha establecido una tipología informal para es tipo de estructuras. Así, se habla de reglas de decisión, asociación, clasificación, predicción, causalidad, optimización etc.

La tarea del descubrimiento de reglas es una tarea bien definida y determinista cuyo objetivo es extraer todas las reglas que tienen unas medidas de *soporte* y *confianza* mayor o igual a un umbral especificado por el usuario.

Podemos demostrar que el *soporte* representa la proporción de casos que presentan una o varias características (probabilidad global evidencias – antecedente de la regla) y la *confianza* es la probabilidad condicional posteirori de "dada las evidencia se da una clase determinada, consecuente de la regla" (regla completa).

Por ejemplo, veamos la figura 6.1 una rama de un árbol aplicado para clasificar de donde extraigo la siguiente regla: Si X7 es "alto" y X1 es "alto" entonces la clase es "riesgo alto". Las variables X7 y X1 son las evidencias y el riesgo es la clase.



Si X7=A y X1=A entonces riesgo alto

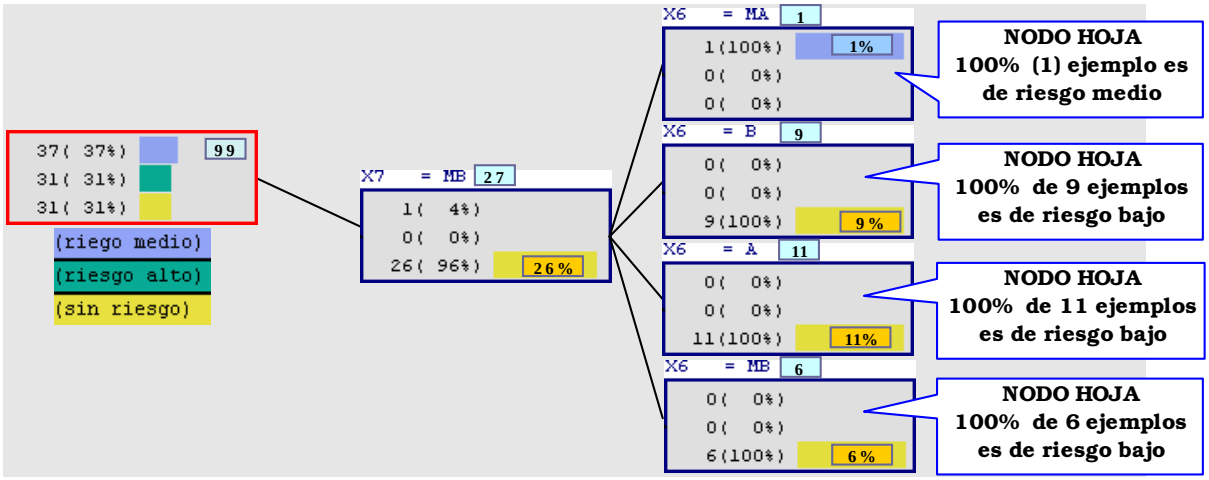
n=	99
nj=	10
nkj=	8
nk=	31
confianza	0.80
soporte	0.10
pk=	0.31
S^2_{kj}=	0.02
S_{kj}=	0.14
V-test	3.48

99%

pkj	pj
Confianza 0.8	**soporte 0.08**
8 de 10	8 de 99

Figura 6.1. Regla de clasificación.

En la figura 6.1., puede observarse la correspondencia entre el soporte, la confianza y las probabilidades a priori y condicional. Con las ecuaciones del V-test, vistas en el epígrafe 2.3.2.2., podemos calcular el nivel de significación de esta regla. En este caso con una confianza del 100% y soporte del 8% es muy significativa la asociación entre las evidencias y la clase, validando esta regla con un V-test de 3.48 mayor al 99% de certeza.



Si X7=MB y X6=B entonces riesgo bajo

n=	99
nj=	1
nkj=	1
nk=	37
confianza	1.00
soporte	0.01
pk=	0.37
S^2_{kj}=	0.23
S_{kj}=	0.48
V-test	1.29

no es significativo

pkj	pj
Confianza 1	**soporte 0..1**
1 de 1	1 de 99

Figura 6.2. Regla de clasificación.

En la figura 6.2., podemos otra regla con confianza de 100% y soporte de 1%. En este caso no es significativa, estadísticamente, la relación entre el antecedente y el consecuente de la regla, por cuanto el V-test es 1.29, menor al 90% de certeza.

Este valor de prueba permite ordenar las reglas según importancia real, constituyendo una métrica de calidad de cualquier tipo de regla, además del soporte y la confianza. El trabajo futuro consiste en incorporar el V-test cómo métrica en la generación de reglas.

6.2. Nivel de significación estadística para Árboles de Regresión.

En los árboles de regresión son los atributos numéricos continuos los que pueden ser clasificados. Para dicha clasificación se parte de un árbol de decisión de profundidad 1 en el que se aplican funciones de regresión a los nodos finales en lugar de colocar clases. Los resultados generados suelen ser más fiables que los obtenidos mediante los árboles de decisión o clasificación.

En este trabajo hemos demostrado la utilidad del algoritmo Demod para caracterizar variables cualitativas. Del mismo modo se puede establecer la aplicación de otro algoritmo, denominado "Desco", para caracterizar variables continuas (cuantitativas) en función de otras también continuas o cualitativas. Esta técnica también incorpora el V-test asociado al test de nulidad de correlaciones y del Chi^2.

El trabajo futuro consiste en incorporar el V-test como métrica para validar las predicciones realizadas sobre la base de un modelo de clasificación de variables continuas.

7. BIBLIOGRAFÍA

[Césari R., Correa M. T., 1999] Césari Ricardo Manuel, Correa María Teresa; *El color en la comunicación social- medición de su valor connotativo en personalidades psicopáticas, nivel de agresividad y su incidencia en la elección del color* .(primera etapa), Resolución Nº 631 -Rectorado UNCuyo, Proyecto Tipo C, Agosto 1999

[Césari, 2005] Césari R., Césari M.; Material científico y pedagógico del curso de posgrado: *Estadistica Multivariada, Métodos Iconográficos de Observación, Información y Comunicación en Investigación en Ingeniería.*, Posgrado Regional Cooperativo en Alimentos, Universidad Nacional de Cuyo, de San Juan, de La Rioja y de San Luis, Sede Mendoza: Secretaría de Posgrado, Facultad de Ciencias Agrarias http://www.fca.uncu.edu.ar

[Cooper & HersKovits, 1992] Cooper G.F., y Herskovits E., *A Bayesian Method for the induction of probabilistic networks from data*. Machine Learning, 9 (pp. 309-347), 1992.

[Cooper, 1999] Cooper G. F., *An Overview of the Representation and Discovery of Causal Relationships using Bayesian Networks.* Computation, Causation & Discovery. C. Glymour and G. F. Cooper, AAAI Press / MIT Press: 3-62. (1999).

[Cowell., Dawid., Lauritzen. y Spiegelhalter, 1999] Cowell R.G., Dawid A.P., Lauritzen S.L. y Spiegelhalter D.J.; *Probabilistic Networks and Expert Systems*. Springer-Verlag, New York, 1999.

[Crivisky, 1997] Crivisky Eduardo, *Material Científico y Pedagógico de los Seminarios PRESTA*. Universidad Libre de Bruselas - Unión Europea – Universidad de Concepción Chile1997.

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

[Díez, 2004] Díez, F. J., *Introducción al Razonamiento Aproximado.* Dpto. Inteligencia Artificial, UNED, Primera edición: Octubre 1998, Revisión: Mayo 2004

[Droesbeke, 1992] Droesbeke, J. .J.; *Méthodes statisques apliques aux sciences sociales, publication du Laboratoire de Méthodologie du Traitement des Données*, U.L.B., Bruxelles, 1992.

[Elvira, 2000] Proyecto Elvira, *Entorno de investigación de métodos y algoritmos de razonamiento probabilístico*. Universidades: Granada, Almería, País Vasco y UNED, 1997-2000. Dirección electrónica: http://www.ia.uned.es/~elvira/

[Felgaer, P, 2005] Felgaer, P; García Martínez, R.; Britos P, 2005; *Tesis de grado en ingeniería informática. Optimización de redes bayesianas basado en técnicas de aprendizaje por inducción*. Facultad de Ingeniería. Universidad de Buenos Aires.

[Felgaer, P., 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.

[Freidman., 1997] Freidman N., Geiger D., Goldszmidt S., *Bayesian Networks classifiers*. *Machine Learning*, 29 (pp 131-161), (1997).

[Gámez y Puerta, 1998] Gámez J. A. y Puerta J. M. (eds.). *Sistemas Expertos Probabilísitcos*. Universidad de Castilla-La Mancha, Cuenca, 1998.

[Grosbras, 1987] Grosbras, J. M.; *Méthodes statisques des sondages*, Ed. Econo-mica, París, 1987

[Hair, Anderson, Tathem y Blas, 1999] Hair Jr., Anderson R., Tatham R., Blak W. "*Significación estadística*". Análisis Multivariante- 5 Edición- PRETENCI HALL. Madrid.- Vol 1, p 8-10. 1999.

[Heckerman D., 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.

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

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

[Jensen, 2001] Jensen F. V., *Bayesian Networks and Decision Graphs*. Springer-Verlag, New York, 2001.

[Jiménez, Martínez, Cruz, 2003] Jiménez Andrade, J. L.; Martínez Morales, M., Cruz Ramírez, N *BayesN: Un Algoritmo para Aprender Redes Bayesianas Clasificadoras a partir de datos* Tesis maestría en Inteligencia Artificial Facultad de Física e Inteligencia Artificial. Universidad Veracruzana. Xalapa~Enríquez, Ver., 2003

[Larrañaga, P., 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

[Lebart, Morineau, Piron, 1984] Lebart L., Morineau A., Piron M. ; *Multivariate Descriptive Statistical Analysis*. John Wiley. Nueva Cork, 1984

[Mitchell, 1997] Mitchell T.; *Machine Learning*. McGraw-Hill, 1997.

[Morineau, 1994] Morineau A., *Note sur la Caractérisation Statique d'une Classe et les Valeurs-tests*, Bulletin Technique Centre Statistique Informática Appliquées, France, Vol 2, nº1 1-2.p 20-27. 1994).

[Pearl. J, 1991] Pearl. J. *Probabilistic Reasoning in Intelligent Systems*: *Networks of Plausible Inference*.Morgan Kaufmann, San Mateo, CA, 1988. Reimpreso con correcciones en 1991.

[Peral, Geiger y Verma , 1989] Pearl J., Geiger D. y Verma T. *Conditional independence and its representations*. *Kybernetika*, 25:33–44, 1989.

[Pérez, Larrañaga, 2005] Pérez A., Larrañaga, P, *Modelos gráficos probabilísticos para la clasificación supervisada empleando la estimación basada en kernels Gaussianos esféricos*; Departamento de Ciencias de la Computación e Inteligencia Artificial; Universidad del País Vasco-Euskal Herreiko Unibertsitatea; 2005

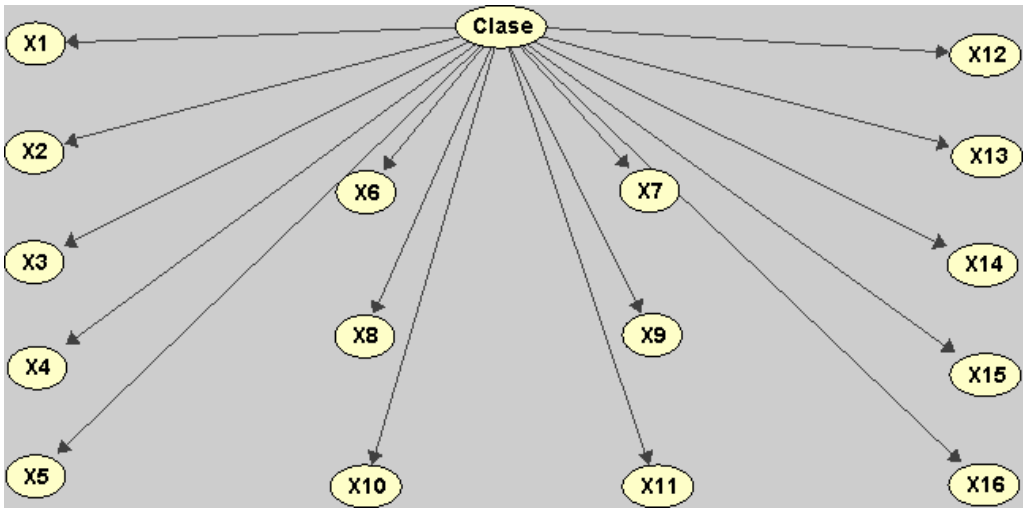
[SAPD, 2002] SPAD (*Système Portable pour l'Analyse de Données*), paquete estadístico consagrado al tratamiento estadístico de grandes matrices de datos, desarrollado por DECISIA, Francia, 1996-2002. www.decisia.fr

[Weka, 2003] Weka (*Waikato Enviroment for Knowledge Analysis*), Biblioteca de clases de aprendizaje en Java, desarrollada en la universidad de Waikato, Nueva Zelanda. (1999-2003)

ANEXOS

A. Salida de las aplicaciones informáticas, algoritmos aplicados sobre la tabla de ejemplo "*Votaciones*"[11].

A. 1. Construcción del clasificador bayesiano Naïve Bayes con Elvira y Weka.



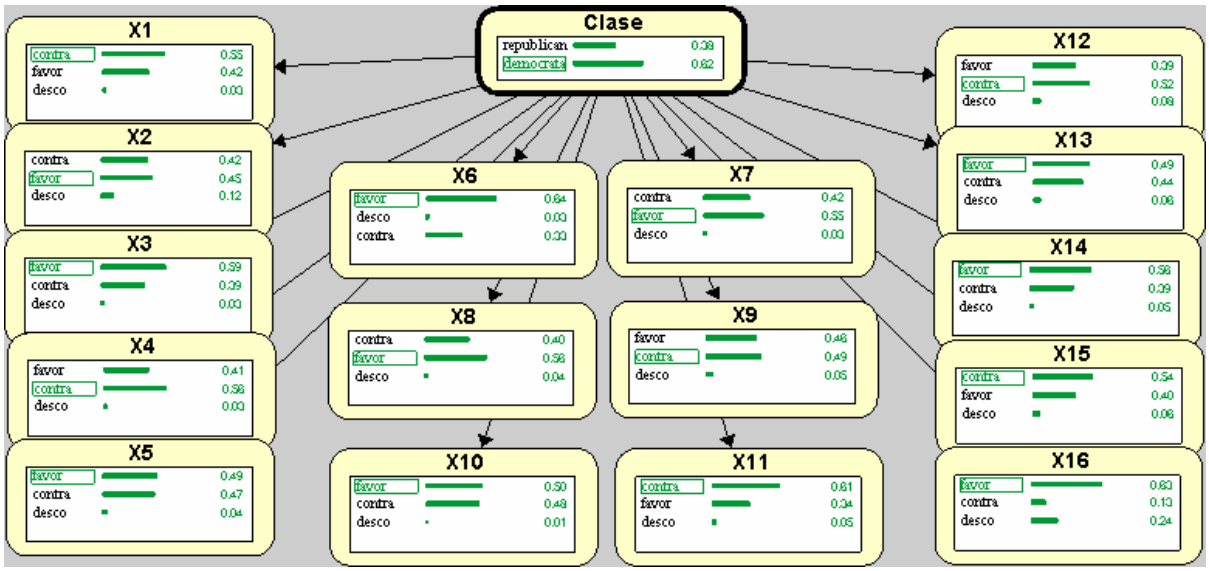
Red bayesiana con Elvira (edición)

Referencias:

X1	Niños discapacitados
X2	Participación en el costo del proyecto del agua
X3	Adopción de la resolución sobre el presupuesto
X4	Congelamiento de los honorarios médicos
X5	Ayuda a El Salvador
X6	Grupos religiosos en las escuelas
X7	Prohibición de las pruebas anti satélistes
X8	Ayuda a los contras de Nicaragua
X9	Misil mx
X10	Inmigración
X11	Reducción a la corporación Synfuels
X12	Presupuesto de educación
X13	Derecho a demandar de la Superfund
X14	Crimen
X15	Exportaciones sin impuestos
X16	Acta sudafricana de administración de exportaciones
Clase	Tipo Republicano o Tipo demócrata

[11] Los siguientes datos fueron extraídos de http://www.ics.uci.edu/~mlearn/MLSummary.html

A.1.1. Probabilidades a priori de cada nodo



Red bayesiana con Elvira (inferencia)

probabilidades		**a priori**
Clase	**republicano**	0.38
	democrata	0.62
X1	**"contra"**	0.55
	"favor"	0.42
	"desco"	0.03
X2	**"contra"**	0.42
	"favor"	0.45
	"desco"	0.12
X3	**"favor"**	0.59
	"contra"	0.39
	"desco"	0.03
X4	**"favor"**	0.41
	"contra"	0.56
	"desco"	0.03
X5	**"favor"**	0.49
	"contra"	0.47
	"desco"	0.04
X6	**"favor"**	0.64
	"desco"	0.03
	"contra"	0.33
X7	**"contra"**	0.42
	"favor"	0.55
	"desco"	0.03
X8	**"contra"**	0.40
	"favor"	0.58
	"desco"	0.04

X9	**"favor"**	0.46
	"contra"	0.49
	"desco"	0.05
X10	**"favor"**	0.50
	"contra"	0.48
	"desco"	0.01
X11	**"contra"**	0.61
	"favor"	0.34
	"desco"	0.05
X12	**"favor"**	0.39
	"contra"	0.52
	"desco"	0.08
X13	**"favor"**	0.49
	"contra"	0.44
	"desco"	0.06
X14	**"favor"**	0.56
	"contra"	0.39
	"desco"	0.05
X15	**"contra"**	0.54
	"favor"	0.40
	"desco"	0.06
X16	**"favor"**	0.63
	"contra"	0.13
	"desco"	0.24

A.1.2. Probabilidades condicionales posteriori

Scheme: weka.classifiers.bayes.NaiveBayesSimple
Relation: Bayesianas_Votacion.csv
Instances: 300
Attributes: 17

Test mode: evaluate on training data
=== Classifier model (full training set) ===

Naive Bayes (simple)

Class **republicano**: P(C) = 0.38741722

Attribute X1

contra	favor	desco
0.78151261	0.19327731	0.02521008

Attribute X2

contra	favor	desco
0.42016807	0.44537815	0.13445378

Attribute X3

favor	contra	desco
0.1512605	0.81512605	0.03361345

Attribute X4

favor	contra	desco
0.94957983	0.01680672	0.03361345

Attribute X5

favor	contra	desco
0.91596639	0.05042017	0.03361345

Attribute X6

favor	desco	contra
0.88235294	0.02521008	0.09243697

Attribute X7

contra	favor	desco
0.69747899	0.25210084	0.05042017

Attribute X8

contra	favor	desco
0.76470588	0.1512605	0.08403361

Attribute X9

favor	contra	desco
0.1092437	0.86554622	0.02521008

Attribute X10

favor	contra	desco
0.5210084	0.45378151	0.02521008

Attribute X11

contra	favor	desco
0.81512605	0.12605042	0.05882353

Attribute X12

favor	contra	desco
0.76470588	0.14285714	0.09243697

Attribute X13
favor contra desco
0.82352941 0.10084034 0.07563025

Attribute X14
favor contra desco
0.90756303 0.02521008 0.06722689

Attribute X15
contra avor desco
0.83193277 0.08403361 0.08403361

Attribute X16
favor contra desco
0.55462185 0.26890756 0.17647059

Class **democrata**: P(C) = 0.61258278

Attribute X1
contra favor desco
0.39572193 0.56684492 0.03743316

Attribute X2
contra f avor desco
0.4171123 0.45989305 0.12299465

Attribute X3
favor contra desco
0.85026738 0.11764706 0.03208556

Attribute X4
favor contra desco
0.06417112 0.89839572 0.03743316

Attribute X5
favor contra desco
0.21925134 0.72727273 0.05347594

Attribute X6
favor desco contra
0.47058824 0.04278075 0.48663102

Attribute X7
contra favor desco
0.2459893 0.72192513 0.03208556

Attribute X8
contra favor desco
0.17647059 0.80748663 0.01604278

Attribute X9
favor contra desco
0.67379679 0.2513369 0.07486631

Attribute X10
favor contra desco
0.48128342 0.5026738 0.01604278

Attribute X11
contra favor desco
0.46524064 0.48663102 0.04812834

Attribute X12
favor contra desco
0.15508021 0.75935829 0.0855615

Attribute X13
favor contra desco
0.27807487 0.65775401 0.06417112

Attribute X14
favor contra desco
0.3368984 0.62032086 0.04278075

Attribute X15
contra favor desco
0.34759358 0.59358289 0.05882353

Attribute X16
favor contra desco
0.6631016 0.04812834 0.28877005

A. 2. Demod de la variable de clase.

A.2.1. Listado de variables relacionadas con la clase

CARACTERISATION PAR LES QUESTIONS DE Clase

V.TEST	PROBA	NUM . LIBELLE DE LA QUESTION	KHI-2	DEG.LIB	INF.A 5
99.99	**0.000**	**4 . X4**	**245.14**	**2**	**1**
99.99	0.000	17 . Clase	295.80	1	0
12.16	0.000	3 . X3	154.72	2	2
11.85	0.000	5 . X5	147.10	2	1
11.07	0.000	8 . X8	129.17	2	1
10.84	0.000	12 . X12	124.06	2	0
10.31	0.000	9 . X9	112.84	2	0
10.24	0.000	14 . X14	111.43	2	0
9.52	0.000	13 . X13	96.98	2	0
8.69	0.000	15 . X15	81.70	2	0
7.78	0.000	7 . X7	66.42	2	1
7.07	0.000	6 . X6	55.74	2	1
6.31	0.000	1 . X1	45.35	2	2
6.14	0.000	11 . X11	43.18	2	0
5.24	0.000	16 . X16	32.64	2	0
-1.72	0.958	2 . X2	0.09	2	0

A.2.2. Caracterización de las clases

CARACTERISATION PAR LES MODALITES DES CLASSES OU MODALITES DE Clase
republicano

V.TEST	PROBA	---- POURCENTAGES ---- CLA/MOD	MOD/CLA	GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
19.66	**0.000**	**100.00**	**100.00**	**38.67**	**republicano**	**Clase**	**repu**	**116**
16.79	0.000	91.06	96.55	41.00	favor	X4	favo	123
12.69	0.000	72.97	93.10	49.33	favor	X5	favo	148
12.66	0.000	82.05	82.76	39.00	contra	X3	cont	117
10.97	0.000	68.92	87.93	49.33	contra	X9	cont	148
10.90	0.000	76.27	77.59	39.33	favor	X12	favo	118
10.46	0.000	63.31	92.24	56.33	favor	X14	favo	169
10.44	0.000	73.77	77.59	40.67	contra	X8	cont	122
9.62	0.000	65.54	83.62	49.33	favor	X13	favo	148
8.59	0.000	60.49	84.48	54.00	contra	X15	cont	162
7.86	0.000	64.57	70.69	42.33	contra	X7	cont	127
7.69	0.000	54.45	89.66	63.67	favor	X6	favo	191
6.76	0.000	55.76	79.31	55.00	contra	X1	cont	165
6.28	0.000	52.75	82.76	60.67	contra	X11	cont	182
5.41	0.000	79.49	26.72	13.00	contra	X16	cont	39
2.66	0.004	81.82	7.76	3.67	desco	X8	desc	11

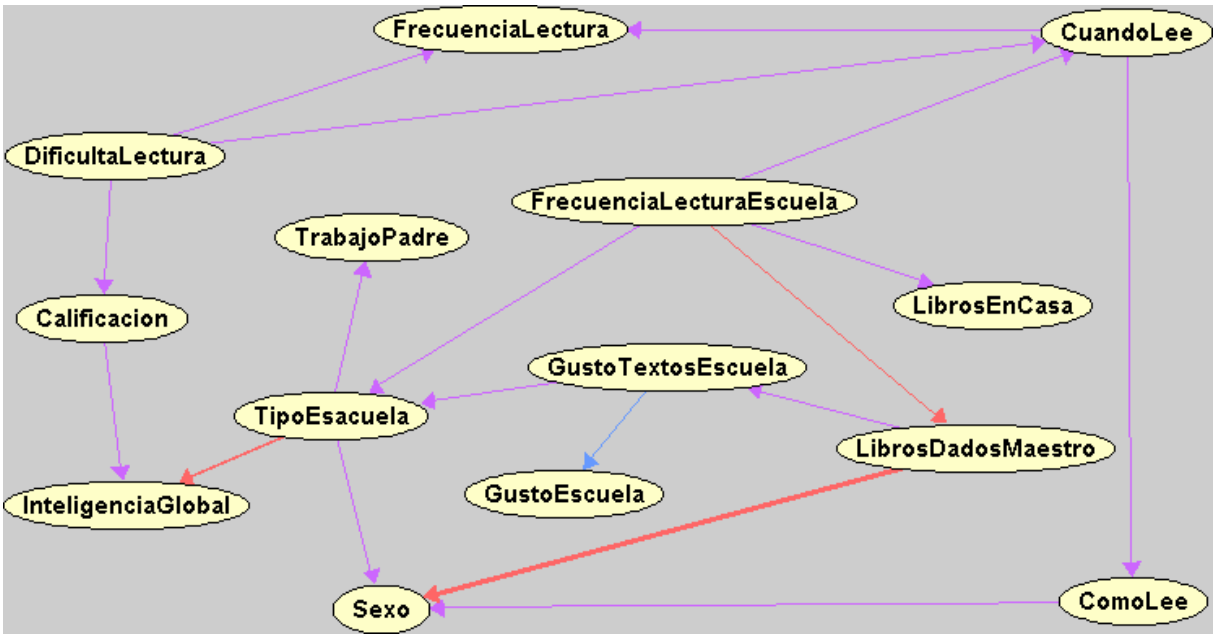
democrata

V.TEST	PROBA	CLA/MOD	MOD/CLA	GLOBAL	CARACTERISTIQUES	DES VARIABLES	IDEN	POIDS
19.66	**0.000**	**100.00**	**100.00**	**61.33**	**democrata**	**Clase**	**demo**	**184**
16.72	0.000	99.40	90.76	56.00	contra	X4	cont	168
12.54	0.000	90.29	85.87	58.33	favor	X3	favo	175
12.49	0.000	96.43	73.37	46.67	contra	X5	cont	140
11.66	0.000	89.82	81.52	55.67	favor	X8	favo	167
11.54	0.000	98.29	62.50	39.00	contra	X14	cont	117
10.91	0.000	89.81	76.63	52.33	contra	X12	cont	157
10.13	0.000	91.24	67.93	45.67	favor	X9	favo	137
10.06	0.000	91.73	66.30	44.33	contra	X13	cont	133
9.42	0.000	92.44	59.78	39.67	favor	X15	favo	119
8.11	0.000	82.21	72.83	54.33	favor	X7	favo	163
7.51	0.000	90.00	48.91	33.33	contra	X6	cont	100
6.69	0.000	86.54	48.91	34.67	favor	X11	favo	104
6.56	0.000	82.68	57.07	42.33	favor	X1	favo	127
2.16	0.015	72.60	28.80	24.33	desco	X16	desc	73
1.87	0.031	86.67	7.07	5.00	desco	X9	desc	15
1.76	0.039	65.43	66.85	62.67	favor	X16	favo	188

B. Salida de las aplicaciones informáticas, algoritmos aplicados sobre la tabla de ejemplo "Práctica de Lectura"[12].

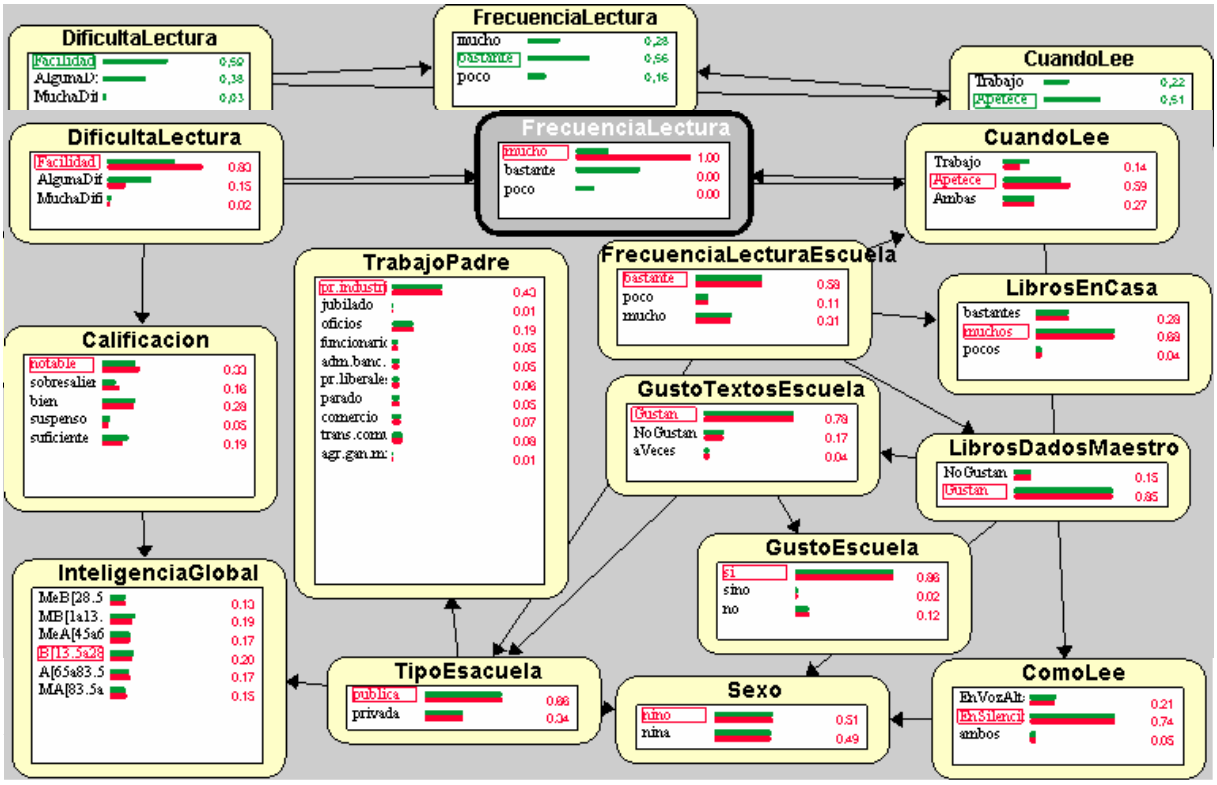
B.1. Red Bayesiana K2

B.1.1. Red bayesiana obtenida por el algoritmo de aprendizaje K2, sin restricciones. (Elvira)



[12] Estudio realizado con datos aportados por la Dirección de Evaluación Educativa - Dirección General de Escuelas

B.1.2. Probabilidades a priori de cada nodo



B.2. Demod para cada nodo

B.2.1. Descripción de Frecuencia de Lectura

CARACTERISATION PAR LES QUESTIONS DE FrecuenciaLectura

V.TEST	PROBA	NUM . LIBELLE DE LA QUESTION	KHI-2	DEG.LIB	INF.A 5
99.99	**0.000**	**16 . FrecuenciaLectura**	**1484.00**	4	0
9.62	0.000	3 . DificultaLectura	106.92	4	1
6.76	0.000	5 . CuandoLee	58.18	4	0
4.96	0.000	10 . Calificación	45.11	8	0
4.35	0.000	2 . LibrosEnCasa	29.33	4	1
4.22	0.000	13 . InteligenciaGlobal	40.84	10	0
3.65	0.000	4 . LibrosDadosMaestro	17.91	2	0
3.29	0.001	8 . GustoEscuela	19.98	4	2
2.72	0.003	1 . FrecuenciaLecturaEscuela	15.81	4	0
2.28	0.011	7 . GustoTextosEscuela	12.98	4	1

CARACTERISATION PAR LES MODALITES DES CLASSES OU MODALITES DE FrecuenciaLectura

V.TEST	PROBA	POURCENTAGES CLA/MOD	POURCENTAGES MOD/CLA	POURCENTAGES GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	DEN POIDS
29.24	**0.000**	**100.00**	**100.00**	**27.36**	**MUCHO**	**FrecuenciaLectura**	**uch 203**
9.13	0.000	39.45	84.73	58.76	Facilidad	DificultaLectura	aci 436
4.05	0.000	31.83	79.80	68.60	muchos	LibrosEnCasa	uch 509
3.86	0.000	45.98	19.70	11.73	sobresaliente	Calificación	obr 87
3.61	0.000	43.75	20.69	12.94	MA[83.5a98]	InteligenciaGlobal	A[8 96
2.64	0.004	31.68	59.61	51.48	Apetece	CuandoLee	pet 382
2.36	0.009	28.98	90.64	85.58	Gustan	LibrosDadosMaestro	ust 635
2.25	0.012	33.05	37.93	31.40	mucho	FrecuenciaLecturaEscuela	uch 233
2.20	0.014	44.74	8.37	5.12	parado	TrabajoPadre	ara 38
2.20	0.014	29.28	84.24	78.71	Gustan	GustoTextosEscuela	ust 584
2.00	0.023	28.71	90.64	86.39	si	GustoEscuela	i 641
1.62	0.053	41.18	6.90	4.58	funcionario	TrabajoPadre	unc 34

V.TEST	PROBA	POURCENTAGES CLA/MOD	POURCENTAGES MOD/CLA	POURCENTAGES GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	DEN POIDS
31.62	**0.000**	**100.00**	**100.00**	**56.74**	**BASTANTE**	**FrecuenciaLectura**	**ast 421**
4.60	0.000	67.49	45.37	38.14	AlgunaDificultad	DificultaLectura	lgu 283
3.61	0.000	67.48	33.02	27.76	bastantes	LibrosEnCasa	ast 206
1.94	0.026	62.81	29.69	26.82	Ambas	CuandoLee	mba 199

V.TEST	PROBA	POURCENTAGES CLA/MOD	POURCENTAGES MOD/CLA	POURCENTAGES GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	DEN POIDS

```
25.19  0.000  100.00  100.00  15.90  poco                 FrecuenciaLectura          oco
118
  6.79  0.000   34.78   47.46  21.70  Trabajo              CuandoLee                  rab
161
  4.65  0.000   28.48   39.83  22.24  MB[1a13.5]           InteligenciaGlobal         B[1
165
  4.37  0.000   56.52   11.02   3.10  MuchaDificultad      DificultaLectura           uch
23
  3.89  0.000   31.82   23.73  11.86  no                   GustoEscuela               o
88
  3.62  0.000   28.97   26.27  14.42  NoGustan             LibrosDadosMaestro         oGu
107
  3.37  0.000   21.91   52.54  38.14  AlgunaDificultad     DificultaLectura           lgu
283
  2.69  0.004   27.50   18.64  10.78  poco                 FrecuenciaLecturaEscuela   oco
80
  2.55  0.005   37.04    8.47   3.64  pocos                LibrosEnCasa               oco
27
  2.53  0.006   30.61   12.71   6.60  suspenso             Calificación               usp
49
  2.31  0.011   23.26   25.42  17.39  NoGustan             GustoTextosEscuela         oGu
129
  2.06  0.020   18.73   60.17  51.08  niño                 Sexo                       iño
379
  1.78  0.037   19.91   36.44  29.11  bien                 Calificación               ien
216
  1.76  0.040   20.59   29.66  22.91  suficiente           Calificación               ufi
170
-------------------------------------------------------------------------------------------
---
```

B.2.2. Descripción de Frecuencia de Lectura en la escuela

CARACTERISATION PAR LES QUESTIONS DE FrecuenciaLecturaEscuela

V.TEST	PROBA	NUM . LIBELLE DE LA QUESTION	KHI-2	DEG.LIB	INF.A 5
99.99	**0.000**	**1 . FrecuenciaLecturaEscuela**	**1484.00**	4	0
5.56	0.000	9 . TipoEsacuela	36.22	2	0
3.69	0.000	7 . GustoTextosEscuela	23.23	4	1
3.58	0.000	5 . CuandoLee	22.33	4	0
3.49	0.000	2 . LibrosEnCasa	21.56	4	1
2.96	0.002	4 . LibrosDadosMaestro	12.94	2	0
2.72	0.003	16 . FrecuenciaLectura	15.81	4	0
2.72	0.003	3 . DificultaLectura	15.80	4	1
-1.93	0.973	10 . Calificación	2.22	8	0

CARACTERISATION PAR LES MODALITES DES CLASSES OU MODALITES DE FrecuenciaLecturaEscuela

```
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                    DEN
POIDS
                CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
 31.54  0.000  100.00  100.00  57.82  BASTANTE             FrecuenciaLecturaEscuela  ast
429
  3.08  0.001   66.99   32.17  27.76  bastantes            LibrosEnCasa              ast
206
  2.93  0.002   65.46   38.00  33.56  privada              TipoEsacuela              riv
249
  2.70  0.004   70.83   15.85  12.94  MA[83.5a98]          InteligenciaGlobal        A[8
96
  2.44  0.007   65.33   30.30  26.82  Ambas                CuandoLee                 mba
199
  2.25  0.012   78.13    5.83   4.31  ambos                CómoLee                   mbo
32
  1.97  0.024   59.37   87.88  85.58  Gustan               LibrosDadosMaestro        ust
635
-------------------------------------------------------------------------------------------
---
-------------------------------------------------------------------------------------------
---
```

```
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                   DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
 22.18  0.000  100.00  100.00  10.78  POCO                 FrecuenciaLecturaEscuela  oco
80
  3.94  0.000   21.71   35.00  17.39  NoGustan             GustoTextosEscuela        oGu
129
  3.34  0.000   16.47   51.25  33.56  privada              TipoEsacuela              riv
249
  3.13  0.001   20.56   27.50  14.42  NoGustan             LibrosDadosMaestro        oGu
107
  2.94  0.002   34.78   10.00   3.10  MuchaDificultad      DificultaLectura          uch
23
  2.69  0.004   18.64   27.50  15.90  poco                 FrecuenciaLectura         oco
118
  2.56  0.005   29.63   10.00   3.64  pocos                LibrosEnCasa              oco
27
  2.46  0.007   13.61   65.00  51.48  Apetece              CuandoLee                 pet
382
-------------------------------------------------------------------------------------------
---
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                   DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
 30.13  0.000  100.00  100.00  31.40  MUCHO                FrecuenciaLecturaEscuela  uch
233
  5.63  0.000   38.13   80.69  66.44  publica              TipoEsacuela              ubl
493
  3.94  0.000   44.72   30.90  21.70  Trabajo              CuandoLee                 rab
161
  2.97  0.002   55.26    9.01   5.12  parado               TrabajoPadre              ara
38
  2.88  0.002   34.77   75.97  68.60  muchos               LibrosEnCasa              uch
509
  2.78  0.003   33.90   84.98  78.71  Gustan               GustoTextosEscuela        ust
584
  2.25  0.012   37.93   33.05  27.36  mucho                FrecuenciaLectura         uch
203
-------------------------------------------------------------------------------------------
---
```

B.2.3. Descripción de Cantidad de libros en casa

CARACTERISATION PAR LES QUESTIONS DE LibrosEnCasa

V.TEST	PROBA	NUM . LIBELLE DE LA QUESTION	KHI-2	DEG.LIB	INF.A 5
99.99	**0.000**	**2 . LibrosEnCasa**	**1484.00**	4	1
4.35	0.000	16 . FrecuenciaLectura	29.33	4	1
3.49	0.000	1 . FrecuenciaLecturaEscuela	21.56	4	1
2.35	0.009	10 . Calificación	20.26	8	2
1.90	0.029	3 . DificultaLectura	10.81	4	1
1.69	0.046	8 . GustoEscuela	9.70	4	3
1.50	0.066	13 . InteligenciaGlobal	17.39	10	4
1.49	0.068	4 . LibrosDadosMaestro	5.37	2	1
-1.03	0.848	12 . TrabajoPadre	11.98	18	12

CARACTERISATION PAR LES MODALITES DES CLASSES OU MODALITES DE LibrosEnCasa

```
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                   DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
```

```
29.34  0.000  100.00  100.00  27.76  bastantes            LibrosEnCasa              ast
206
  3.61  0.000   33.02   67.48  56.74  bastante             FrecuenciaLectura         ast
421
  3.08  0.001   32.17   66.99  57.82  bastante             FrecuenciaLecturaEscuela  ast
429
  1.80  0.036   33.53   27.67  22.91  suficiente           Calificación              ufi
170
  1.67  0.047   31.45   43.20  38.14  AlgunaDificultad     DificultaLectura          lgu
283
-------------------------------------------------------------------------------------------
---
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                    DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
 30.13  0.000  100.00  100.00  68.60  muchos               LibrosEnCasa              uch
509
  4.05  0.000   79.80   31.83  27.36  mucho                FrecuenciaLectura         uch
203
  2.88  0.002   75.97   34.77  31.40  mucho                FrecuenciaLecturaEscuela  uch
233
  2.17  0.015   74.54   31.63  29.11  bien                 Calificación              ien
216
  2.08  0.019   78.13   14.73  12.94  MA[83.5a98]          InteligenciaGlobal        A[8
96
-------------------------------------------------------------------------------------------
---
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                    DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
 99.99  0.000  100.00  100.00   3.64  pocos                LibrosEnCasa              oco
27
  2.56  0.005   10.00   29.63  10.78  poco                 FrecuenciaLecturaEscuela  oco
80
  2.55  0.005    8.47   37.04  15.90  poco                 FrecuenciaLectura         oco
118
  2.35  0.009    9.09   29.63  11.86  no                   GustoEscuela              o
88
  2.28  0.011    6.36   51.85  29.65  notable              Calificación              ota
220
  2.10  0.018    6.83   40.74  21.70  Trabajo              CuandoLee                 rab
161
  1.89  0.030    7.48   29.63  14.42  NoGustan             LibrosDadosMaestro        oGu
107
  1.73  0.041    7.53   25.93  12.53  MeB[28.5a45]         InteligenciaGlobal        eB[
93
  1.68  0.047   13.04   11.11   3.10  MuchaDificultad      DificultaLectura          uch
23
  1.60  0.055    6.06   37.04  22.24  MB[1a13.5]           InteligenciaGlobal        B[1
165
-------------------------------------------------------------------------------------------
---
```

B.2.4. Descripción de Grado de dificultad en la lectura

CARACTERISATION PAR LES QUESTIONS DE DificultaLectura

V.TEST	PROBA	NUM . LIBELLE DE LA QUESTION	KHI-2	DEG.LIB	INF.A 5
99.99	**0.000**	**3 . DificultaLectura**	**1484.00**	4	1
9.62	0.000	16 . FrecuenciaLectura	106.92	4	1
8.22	0.000	10 . Calificación	93.25	8	2
5.45	0.000	13 . InteligenciaGlobal	55.57	10	5
3.90	0.000	5 . CuandoLee	25.08	4	1
2.72	0.003	1 . FrecuenciaLecturaEscuela	15.80	4	1
2.60	0.005	6 . CómoLee	15.01	4	2
2.31	0.010	4 . LibrosDadosMaestro	9.14	2	1
1.90	0.029	2 . LibrosEnCasa	10.81	4	1

CARACTERISATION PAR LES MODALITES DES CLASSES OU MODALITES DE DificultaLectura

V.TEST	PROBA	POURCENTAGES CLA/MOD	POURCENTAGES MOD/CLA	POURCENTAGES GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	DEN POIDS
31.47	**0.000**	**100.00**	**100.00**	**58.76**	**Facilidad**	**DificultaLectura**	**aci 436**
9.13	0.000	84.73	39.45	27.36	mucho	FrecuenciaLectura	uch 203
6.88	0.000	90.80	18.12	11.73	sobresaliente	Calificación	obr 87
4.17	0.000	70.45	35.55	29.65	notable	Calificación	ota 220
3.43	0.000	75.00	16.51	12.94	MA[83.5a98]	InteligenciaGlobal	A[8 96
3.03	0.001	62.11	78.21	73.99	EnSilencio	CómoLee	nSi 549
2.30	0.011	65.83	30.05	26.82	Ambas	CuandoLee	mba 199
2.13	0.017	60.37	88.76	86.39	si	GustoEscuela	i 641
1.71	0.044	66.12	18.35	16.31	MeA[45a65]	InteligenciaGlobal	eA[121
1.70	0.045	60.45	80.96	78.71	Gustan	GustoTextosEscuela	ust 584
1.65	0.050	61.78	54.13	51.48	Apetece	CuandoLee	pet 382
1.64	0.051	73.53	5.73	4.58	funcionario	TrabajoPadre	unc 34

V.TEST	PROBA	POURCENTAGES CLA/MOD	POURCENTAGES MOD/CLA	POURCENTAGES GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	DEN POIDS
31.16	**0.000**	**100.00**	**100.00**	**38.14**	**AlgunaDificultad**	**DificultaLectura**	**lgu 283**
5.67	0.000	57.58	33.57	22.24	MB[1a13.5]	InteligenciaGlobal	B[1 165
5.45	0.000	56.47	33.92	22.91	suficiente	Calificación	ufi 170
4.60	0.000	45.37	67.49	56.74	bastante	FrecuenciaLectura	ast 421
3.83	0.000	51.55	29.33	21.70	Trabajo	CuandoLee	rab 161
3.37	0.000	52.54	21.91	15.90	poco	FrecuenciaLectura	oco 118
2.35	0.009	55.10	9.54	6.60	suspenso	Calificación	usp 49
2.30	0.011	50.00	15.55	11.86	no	GustoEscuela	o 88
1.84	0.033	44.72	25.44	21.70	EnVozAlta	CómoLee	nVo 161
1.67	0.047	43.20	31.45	27.76	bastantes	LibrosEnCasa	ast 206
1.64	0.051	44.60	21.91	18.73	oficios	TrabajoPadre	fic 139

```
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                    DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
 13.90  0.000  100.00  100.00   3.10  MuchaDificultad      DificultaLectura          uch
23
  4.37  0.000   11.02   56.52  15.90  poco                 FrecuenciaLectura         oco
118
  2.94  0.002   10.00   34.78  10.78  poco                 FrecuenciaLecturaEscuela  oco
80
  2.30  0.011    7.48   34.78  14.42  NoGustan             LibrosDadosMaestro        oGu
107
  2.20  0.014   10.20   21.74   6.60  suspenso             Calificación              usp
49
  2.18  0.014    6.21   43.48  21.70  EnVozAlta            CómoLee                   nVo
161
  2.11  0.017    6.06   43.48  22.24  MB[1a13.5]           InteligenciaGlobal        B[1
165
  1.73  0.042    5.59   39.13  21.70  Trabajo              CuandoLee                 rab
161
  1.68  0.047   11.11   13.04   3.64  pocos                LibrosEnCasa              oco
27
  1.66  0.049    5.76   34.78  18.73  oficios              TrabajoPadre              fic
139
-------------------------------------------------------------------------------------------
---
```

B.2.5. Descripción de Gusto por los libros dados por el maestro

CARACTERISATION PAR LES QUESTIONS DE LibrosDadosMaestro

V.TEST	PROBA	NUM . LIBELLE DE LA QUESTION	KHI-2	DEG.LIB	INF.A 5
99.99	**0.000**	**4 . LibrosDadosMaestro**	**733.92**	1	0
8.91	0.000	7 . GustoTextosEscuela	85.56	2	1
4.85	0.000	11 . Sexo	24.85	1	0
3.65	0.000	16 . FrecuenciaLectura	17.91	2	0
3.28	0.001	8 . GustoEscuela	15.09	2	1
2.96	0.002	1 . FrecuenciaLecturaEscuela	12.94	2	0
2.55	0.005	9 . TipoEsacuela	7.77	1	0
2.31	0.010	3 . DificultaLectura	9.14	2	1

CARACTERISATION PAR LES MODALITES DES CLASSES OU MODALITES DE LibrosDadosMaestro

```
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                    DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
 24.43  0.000  100.00  100.00  14.42  NoGustan             LibrosDadosMaestro        oGu
107
  8.15  0.000   40.31   48.60  17.39  NoGustan             GustoTextosEscuela        oGu
129
  5.07  0.000   20.84   73.83  51.08  niño                 Sexo                      iño
379
  3.62  0.000   26.27   28.97  15.90  poco                 FrecuenciaLectura         oco
118
  3.26  0.001   27.27   22.43  11.86  no                   GustoEscuela              o
88
  3.13  0.001   27.50   20.56  10.78  poco                 FrecuenciaLecturaEscuela  oco
80
  2.74  0.003   19.68   45.79  33.56  privada              TipoEsacuela              riv
249
  2.30  0.011   34.78    7.48   3.10  MuchaDificultad      DificultaLectura          uch
23
  1.89  0.030   29.63    7.48   3.64  pocos                LibrosEnCasa              oco
27
  1.76  0.039   16.75   59.81  51.48  Apetece              CuandoLee                 pet
382
  1.71  0.044   18.82   29.91  22.91  suficiente           Calificación              ufi
170
-------------------------------------------------------------------------------------------
---
-------------------------------------------------------------------------------------------
---
```

```
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                   DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------
---
 24.43  0.000  100.00  100.00  85.58  Gustan               LibrosDadosMaestro       ust
635
  7.01  0.000   90.75   83.46  78.71  Gustan               GustoTextosEscuela       ust
584
  5.07  0.000   92.29   52.76  48.92  niña                 Sexo                     iña
363
  2.74  0.003   88.24   68.50  66.44  publica              TipoEsacuela             ubl
493
  2.59  0.005   87.05   87.87  86.39  si                   GustoEscuela             i
641
  2.36  0.009   90.64   28.98  27.36  mucho                FrecuenciaLectura        uch
203
  1.97  0.024   87.88   59.37  57.82  bastante             FrecuenciaLecturaEscuela ast
429
  1.74  0.041   96.88    4.88   4.31  ambos                CómoLee                  mbo
32
  1.73  0.042   89.45   28.03  26.82  Ambas                CuandoLee                mba
199
  1.59  0.055   89.93   21.10  20.08  B[13.5a28.5]         InteligenciaGlobal       [13
149
-------------------------------------------------------------------------------------
---
```

B.2.6. Descripción de cuándo leen

CARACTERISATION PAR LES QUESTIONS DE CuandoLee

V.TEST	PROBA	NUM . LIBELLE DE LA QUESTION	KHI-2	DEG.LIB	INF.A 5
99.99	**0.000**	**5 . CuandoLee**	**1484.00**	4	0
6.76	0.000	16 . FrecuenciaLectura	58.18	4	0
4.89	0.000	13 . InteligenciaGlobal	48.44	10	0
4.34	0.000	6 . CómoLee	29.18	4	0
3.90	0.000	3 . DificultaLectura	25.08	4	1
3.72	0.000	10 . Calificación	31.87	8	0
3.58	0.000	1 . FrecuenciaLecturaEscuela	22.33	4	0
3.37	0.000	9 . TipoEsacuela	15.77	2	0
2.93	0.002	11 . Sexo	12.77	2	0
2.78	0.003	7 . GustoTextosEscuela	16.24	4	0
1.67	0.047	8 . GustoEscuela	9.62	4	2

CARACTERISATION PAR LES MODALITES DES CLASSES OU MODALITES DE CuandoLee

```
-------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                   DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------
---
 27.58  0.000  100.00  100.00  21.70  Trabajo              CuandoLee                rab
161
  6.79  0.000   47.46   34.78  15.90  poco                 FrecuenciaLectura        oco
118
  5.68  0.000   38.79   39.75  22.24  MB[1a13.5]           InteligenciaGlobal       B[1
165
  3.99  0.000   25.96   79.50  66.44  publica              TipoEsacuela             ubl
493
  3.94  0.000   30.90   44.72  31.40  mucho                FrecuenciaLecturaEscuela uch
233
  3.83  0.000   29.33   51.55  38.14  AlgunaDificultad     DificultaLectura         lgu
283
  3.27  0.001   26.65   62.73  51.08  niño                 Sexo                     iño
379
  2.68  0.004   38.78   11.80   6.60  suspenso             Calificación             usp
49
  2.62  0.004   29.41   31.06  22.91  suficiente           Calificación             ufi
170
  2.50  0.006   32.95   18.01  11.86  no                   GustoEscuela             o
88
  2.10  0.018   40.74    6.83   3.64  pocos                LibrosEnCasa             oco
27
  2.04  0.021   27.95   27.95  21.70  EnVozAlta            CómoLee                  nVo
161
  1.73  0.042   39.13    5.59   3.10  MuchaDificultad      DificultaLectura         uch
23
```

```
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                   DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------
---
 31.82  0.000  100.00  100.00  51.48  Apetece              CuandoLee                pet
382
  2.83  0.002   54.64   78.53  73.99  EnSilencio           CómoLee                  nSi
549
  2.64  0.004   59.61   31.68  27.36  mucho                FrecuenciaLectura        uch
203
  2.55  0.005   62.02   20.94  17.39  NoGustan             GustoTextosEscuela       oGu
129
  2.46  0.007   65.00   13.61  10.78  poco                 FrecuenciaLecturaEscuela oco
80
  1.92  0.028   56.63   36.91  33.56  privada              TipoEsacuela             riv
249
  1.76  0.039   59.81   16.75  14.42  NoGustan             LibrosDadosMaestro       oGu
107
  1.65  0.050   54.13   61.78  58.76  Facilidad            DificultaLectura         aci
436
-------------------------------------------------------------------------------------
---
-------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                   DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------
---
 29.11  0.000  100.00  100.00  26.82  Ambas                CuandoLee                mba
199
  4.16  0.000   62.50   10.05   4.31  ambos                CómoLee                  mbo
32
  2.73  0.003   51.72    7.54   3.91  aVeces               GustoTextosEscuela       Vec
29
  2.60  0.005   33.64   37.19  29.65  notable              Calificación             ota
220
  2.44  0.007   30.30   65.33  57.82  bastante             FrecuenciaLecturaEscuela ast
429
  2.30  0.011   30.05   65.83  58.76  Facilidad            DificultaLectura         aci
436
  2.18  0.015   30.58   55.78  48.92  niña                 Sexo                     iña
363
  2.13  0.017   28.24   90.95  86.39  si                   GustoEscuela             i
641
  2.12  0.017   36.46   17.59  12.94  MA[83.5a98]          InteligenciaGlobal       A[8
96
  1.94  0.026   29.69   62.81  56.74  bastante             FrecuenciaLectura        ast
421
-------------------------------------------------------------------------------------
---
```

B.2.7. Descripción de Cómo prefieren leen

CARACTERISATION PAR LES QUESTIONS DE CómoLee

V.TEST	PROBA	NUM . LIBELLE DE LA QUESTION	KHI-2	DEG.LIB	INF.A 5
99.99	**0.000**	**6 . CómoLee**	**1484.00**	4	1
7.41	0.000	11 . Sexo	60.83	2	0
4.34	0.000	5 . CuandoLee	29.18	4	0
2.60	0.005	3 . DificultaLectura	15.01	4	2
2.54	0.006	10 . Calificación	21.64	8	2
2.24	0.013	7 . GustoTextosEscuela	12.75	4	1
1.99	0.023	13 . InteligenciaGlobal	20.71	10	2

CARACTERISATION PAR LES MODALITES DES CLASSES OU MODALITES DE CómoLee

```
-------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                   DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------
---
```

```
27.58  0.000  100.00  100.00  21.70  EnVozAlta            CómoLee                  nVo
161
  7.00  0.000   32.51   73.29  48.92  niña                 Sexo                     iña
363
  2.68  0.004   38.78   11.80   6.60  suspenso             Calificación             usp
49
  2.18  0.014   43.48    6.21   3.10  MuchaDificultad      DificultaLectura         uch
23
  2.05  0.020   27.88   28.57  22.24  MB[1a13.5]           InteligenciaGlobal       B[1
165
  2.04  0.021   27.95   27.95  21.70  Trabajo              CuandoLee                rab
161
  1.87  0.031   26.39   35.40  29.11  bien                 Calificación             ien
216
  1.84  0.033   25.44   44.72  38.14  AlgunaDificultad     DificultaLectura         lgu
283
  1.65  0.049   27.34   23.60  18.73  oficios              TrabajoPadre             fic
139
-------------------------------------------------------------------------------------------
---
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                    DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
 28.90  0.000  100.00  100.00  73.99  EnSilencio           CómoLee                  nSi
549
  7.82  0.000   86.28   59.56  51.08  niño                 Sexo                     iño
379
  3.03  0.001   78.21   62.11  58.76  Facilidad            DificultaLectura         aci
436
  2.83  0.002   78.53   54.64  51.48  Apetece              CuandoLee                pet
382
  2.76  0.003   86.21   13.66  11.73  sobresaliente        Calificación             obr
87
  2.55  0.005   84.95   14.39  12.53  MeB[28.5a45]         InteligenciaGlobal       eB[
93
  1.91  0.028   82.29   14.39  12.94  MA[83.5a98]          InteligenciaGlobal       A[8
96
  1.72  0.043   76.06   68.31  66.44  publica              TipoEsacuela             ubl
493
-------------------------------------------------------------------------------------------
---
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                    DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
 99.99  0.000  100.00  100.00   4.31  ambos                CómoLee                  mbo
32
  4.16  0.000   10.05   62.50  26.82  Ambas                CuandoLee                mba
199
  2.50  0.006    6.34   71.88  48.92  niña                 Sexo                     iña
363
  2.25  0.012    5.83   78.13  57.82  bastante             FrecuenciaLecturaEscuela ast
429
  1.86  0.032   13.79   12.50   3.91  aVeces               GustoTextosEscuela       Vec
29
  1.74  0.041    4.88   96.88  85.58  Gustan               LibrosDadosMaestro       ust
635
  1.62  0.053   11.76   12.50   4.58  funcionario          TrabajoPadre             unc
34
-------------------------------------------------------------------------------------------
---
```

B.2.8. Descripción del gusto por los textos escolares

CARACTERISATION PAR LES QUESTIONS DE GustoTextosEscuela

V.TEST	PROBA	NUM . LIBELLE DE LA QUESTION	KHI-2	DEG.LIB	INF.A 5
99.99	**0.000**	**7 . GustoTextosEscuela**	**1484.00**	4	1
8.91	0.000	4 . LibrosDadosMaestro	85.56	2	1
4.33	0.000	9 . TipoEsacuela	23.57	2	0
4.17	0.000	8 . GustoEscuela	27.54	4	3
3.69	0.000	1 . FrecuenciaLecturaEscuela	23.23	4	1
3.20	0.001	11 . Sexo	14.58	2	0
2.78	0.003	5 . CuandoLee	16.24	4	0
2.28	0.011	16 . FrecuenciaLectura	12.98	4	1
2.24	0.013	6 . CómoLee	12.75	4	1

CARACTERISATION PAR LES MODALITES DES CLASSES OU MODALITES DE GustoTextosEscuela

V.TEST	PROBA	POURCENTAGES CLA/MOD	POURCENTAGES MOD/CLA	POURCENTAGES GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	DEN POIDS
27.44	**0.000**	**100.00**	**100.00**	**78.71**	**Gustan**	**GustoTextosEscuela**	**ust 584**
7.01	0.000	83.46	90.75	85.58	Gustan	LibrosDadosMaestro	ust 635
4.38	0.000	83.57	70.55	66.44	publica	TipoEsacuela	ubl 493
2.78	0.003	84.98	33.90	31.40	mucho	FrecuenciaLecturaEscuela	uch 233
2.53	0.006	80.34	88.18	86.39	si	GustoEscuela	i 641
2.30	0.011	82.37	51.20	48.92	niña	Sexo	iña 363
2.20	0.014	84.24	29.28	27.36	mucho	FrecuenciaLectura	uch 203
1.70	0.045	80.96	60.45	58.76	Facilidad	DificultaLectura	aci 436

V.TEST	PROBA	POURCENTAGES CLA/MOD	POURCENTAGES MOD/CLA	POURCENTAGES GLOBAL	MODALITES CARACTERISTIQUES	DES VARIABLES	DEN POIDS
25.89	**0.000**	**100.00**	**100.00**	**17.39**	**NoGustan**	**GustoTextosEscuela**	**oGu 129**
8.15	0.000	48.60	40.31	14.42	NoGustan	LibrosDadosMaestro	oGu 107
3.94	0.000	35.00	21.71	10.78	poco	FrecuenciaLecturaEscuela	oco 80
3.43	0.000	22.16	65.12	51.08	niño	Sexo	iño 379
3.17	0.001	30.68	20.93	11.86	no	GustoEscuela	o 88
3.07	0.001	23.69	45.74	33.56	privada	TipoEsacuela	riv 249
2.55	0.005	20.94	62.02	51.48	Apetece	CuandoLee	pet 382
2.31	0.011	25.42	23.26	15.90	poco	FrecuenciaLectura	oco 118
2.02	0.022	23.74	25.58	18.73	oficios	TrabajoPadre	fic 139

```
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                    DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
 99.99  0.000  100.00  100.00   3.91  aVeces               GustoTextosEscuela        Vec
29
  3.02  0.001    7.23   62.07  33.56  privada             TipoEsacuela              riv
249
  2.73  0.003    7.54   51.72  26.82  Ambas               CuandoLee                 mba
199
  1.86  0.032   12.50   13.79   4.31  ambos               CómoLee                   mbo
32
  1.64  0.051    5.23   65.52  48.92  niña                Sexo                      iña
363
-------------------------------------------------------------------------------------------
---
```

B.2.9. Descripción del gusto por la escuela

CARACTERISATION PAR LES QUESTIONS DE GustoEscuela

V.TEST	PROBA	NUM . LIBELLE DE LA QUESTION	KHI-2	DEG.LIB	INF.A 5
99.99	**0.000**	**8 . GustoEscuela**	**1484.00**	4	3
4.17	0.000	7 . GustoTextosEscuela	27.54	4	3
3.29	0.001	16 . FrecuenciaLectura	19.98	4	2
3.28	0.001	4 . LibrosDadosMaestro	15.09	2	1
3.14	0.001	11 . Sexo	14.18	2	0
1.69	0.046	2 . LibrosEnCasa	9.70	4	3
1.67	0.047	5 . CuandoLee	9.62	4	2
1.65	0.049	9 . TipoEsacuela	6.01	2	1

CARACTERISATION PAR LES MODALITES DES CLASSES OU MODALITES   DE GustoEscuela

```
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                    DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
 23.98  0.000  100.00  100.00  86.39  si                  GustoEscuela              i
641
  3.66  0.000   91.18   51.64  48.92  niña                Sexo                      iña
363
  2.59  0.005   87.87   87.05  85.58  Gustan              LibrosDadosMaestro        ust
635
  2.53  0.006   88.18   80.34  78.71  Gustan              GustoTextosEscuela        ust
584
  2.13  0.017   90.95   28.24  26.82  Ambas               CuandoLee                 mba
199
  2.13  0.017   88.76   60.37  58.76  Facilidad           DificultaLectura          aci
436
  2.00  0.023   90.64   28.71  27.36  mucho               FrecuenciaLectura         uch
203
  1.93  0.027   89.96   34.95  33.56  privada             TipoEsacuela              riv
249
-------------------------------------------------------------------------------------------
---
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                    DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
 2.26  0.012   10.34   23.08   3.91  aVeces               GustoTextosEscuela        Vec
29
  1.80  0.036    2.43   92.31  66.44  publica             TipoEsacuela              ubl
493
  1.58  0.057    3.18   53.85  29.65  notable             Calificación              ota
220
-------------------------------------------------------------------------------------------
---
-------------------------------------------------------------------------------------------
---
```

```
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                   DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------
---
 22.91  0.000  100.00  100.00  11.86  no                   GustoEscuela             o
88
  3.89  0.000   23.73   31.82  15.90  poco                 FrecuenciaLectura        oco
118
  3.57  0.000   16.09   69.32  51.08  niño                 Sexo                     iño
379
  3.26  0.001   22.43   27.27  14.42  NoGustan             LibrosDadosMaestro       oGu
107
  3.17  0.001   20.93   30.68  17.39  NoGustan             GustoTextosEscuela       oGu
129
  2.50  0.006   18.01   32.95  21.70  Trabajo              CuandoLee                rab
161
  2.40  0.008   24.49   13.64   6.60  suspenso             Calificación             usp
49
  2.35  0.009   29.63    9.09   3.64  pocos                LibrosEnCasa             oco
27
  2.30  0.011   15.55   50.00  38.14  AlgunaDificultad     DificultaLectura         lgu
283
  1.85  0.032   16.36   30.68  22.24  MB[1a13.5]           InteligenciaGlobal       B[1
165
-------------------------------------------------------------------------------------
---
```

B.2.10. Descripción del tipo de escuela

CARACTERISATION PAR LES QUESTIONS DE TipoEsacuela

V.TEST	PROBA	NUM . LIBELLE DE LA QUESTION	KHI-2	DEG.LIB	INF.A 5
99.99	**0.000**	**9 . TipoEsacuela**	**737.52**	1	0
7.67	0.000	13 . InteligenciaGlobal	75.09	5	0
7.01	0.000	12 . TrabajoPadre	75.57	9	4
5.56	0.000	1 . FrecuenciaLecturaEscuela	36.22	2	0
4.63	0.000	11 . Sexo	22.78	1	0
4.33	0.000	7 . GustoTextosEscuela	23.57	2	0
3.37	0.000	5 . CuandoLee	15.77	2	0
2.55	0.005	4 . LibrosDadosMaestro	7.77	1	0
1.65	0.049	8 . GustoEscuela	6.01	2	1

CARACTERISATION PAR LES MODALITES DES CLASSES OU MODALITES DE TipoEsacuela

```
-------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                   DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------
---
 30.52  0.000  100.00  100.00  66.44  publica              TipoEsacuela             ubl
493
  5.63  0.000   80.69   38.13  31.40  mucho                FrecuenciaLecturaEscuela uch
233
  4.99  0.000   76.31   50.30  43.80  pr.industria         TrabajoPadre             r.i
325
  4.82  0.000   81.82   27.38  22.24  MB[1a13.5]           InteligenciaGlobal       B[1
165
  4.79  0.000   74.67   57.40  51.08  niño                 Sexo                     iño
379
  4.54  0.000   81.88   24.75  20.08  B[13.5a28.5]         InteligenciaGlobal       [13
149
  4.38  0.000   70.55   83.57  78.71  Gustan               GustoTextosEscuela       ust
584
  4.06  0.000   94.74    7.30   5.12  parado               TrabajoPadre             ara
38
  3.99  0.000   79.50   25.96  21.70  Trabajo              CuandoLee                rab
161
  2.74  0.003   68.50   88.24  85.58  Gustan               LibrosDadosMaestro       ust
635
  1.72  0.043   68.31   76.06  73.99  EnSilencio           CómoLee                  nSi
549
-------------------------------------------------------------------------------------
---
-------------------------------------------------------------------------------------
---
```

```
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                   DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------
---
 30.52  0.000  100.00  100.00  33.56  privada              TipoEsacuela             riv
249
  4.79  0.000   42.15   61.45  48.92  niña                 Sexo                     iña
363
  4.57  0.000   55.21   21.29  12.94  MA[83.5a98]          InteligenciaGlobal       A[8
96
  4.18  0.000   66.67   10.44   5.26  adm.banc.emp.        TrabajoPadre             dm.
39
  4.14  0.000   50.85   24.10  15.90  A[65a83.5]           InteligenciaGlobal       [65
118
  4.02  0.000   65.00   10.44   5.39  pr.liberales         TrabajoPadre             r.l
40
  3.34  0.000   51.25   16.47  10.78  poco                 FrecuenciaLecturaEscuela oco
80
  3.07  0.001   45.74   23.69  17.39  NoGustan             GustoTextosEscuela       oGu
129
  3.02  0.001   62.07    7.23   3.91  aVeces               GustoTextosEscuela       Vec
29
  2.93  0.002   38.00   65.46  57.82  bastante             FrecuenciaLecturaEscuela ast
429
  2.74  0.003   45.79   19.68  14.42  NoGustan             LibrosDadosMaestro       oGu
107
  2.46  0.007   50.00   10.84   7.28  comercio             TrabajoPadre             ome
54
  2.27  0.012   42.98   20.88  16.31  MeA[45a65]           InteligenciaGlobal       eA[
121
  1.93  0.027   34.95   89.96  86.39  si                   GustoEscuela             i
641
  1.92  0.028   36.91   56.63  51.48  Apetece              CuandoLee                pet
382
  1.75  0.040   40.29   22.49  18.73  oficios              TrabajoPadre             fic
139
-------------------------------------------------------------------------------------
---
```

B.2.11. Descripción de la Calificación Global

CARACTERISATION PAR LES QUESTIONS DE Calificación

V.TEST	PROBA	NUM . LIBELLE DE LA QUESTION	KHI-2	DEG.LIB	INF.A 5
53.60	**0.000**	**10 . Calificación**	**2968.00**	16	1
13.06	0.000	13 . InteligenciaGlobal	238.16	20	0
8.22	0.000	3 . DificultaLectura	93.25	8	2
4.96	0.000	16 . FrecuenciaLectura	45.11	8	0
3.72	0.000	5 . CuandoLee	31.87	8	0
2.82	0.002	11 . Sexo	16.54	4	0
2.54	0.006	6 . CómoLee	21.64	8	2
2.35	0.009	2 . LibrosEnCasa	20.26	8	2
1.62	0.053	8 . GustoEscuela	15.36	8	5
-1.93	0.973	1 . FrecuenciaLecturaEscuela	2.22	8	0

CARACTERISATION PAR LES MODALITES DES CLASSES OU MODALITES DE Calificación

```
-------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                   DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------
---
 29.77  0.000  100.00  100.00  29.65  notable              Calificación             ota
220
  4.17  0.000   35.55   70.45  58.76  Facilidad            DificultaLectura         aci
436
  3.12  0.001   42.37   22.73  15.90  A[65a83.5]           InteligenciaGlobal       [65
118
  3.04  0.001   43.75   19.09  12.94  MA[83.5a98]          InteligenciaGlobal       A[8
96
  2.60  0.005   37.19   33.64  26.82  Ambas                CuandoLee                mba
199
  2.49  0.006   39.67   21.82  16.31  MeA[45a65]           InteligenciaGlobal       eA[
121
  2.28  0.011   51.85    6.36   3.64  pocos                LibrosEnCasa             oco
27
```

```
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                    DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
 22.83  0.000  100.00  100.00  11.73  sobresaliente        Calificación             obr
87
  7.13  0.000   37.50   41.38  12.94  MA[83.5a98]          InteligenciaGlobal       A[8
96
  6.88  0.000   18.12   90.80  58.76  Facilidad            DificultaLectura         aci
436
  3.86  0.000   19.70   45.98  27.36  mucho                FrecuenciaLectura        uch
203
  2.76  0.003   13.66   86.21  73.99  EnSilencio           CómoLee                  nSi
549
  2.29  0.011   18.64   25.29  15.90  A[65a83.5]           InteligenciaGlobal       [65
118
-------------------------------------------------------------------------------------------
---
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                    DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
 29.66  0.000  100.00  100.00  29.11  bien                 Calificación             ien
216
  2.88  0.002   34.16   57.41  48.92  niña                 Sexo                     iña
363
  2.61  0.005   38.26   26.39  20.08  B[13.5a28.5]         InteligenciaGlobal       [13
149
  2.17  0.015   31.63   74.54  68.60  muchos               LibrosEnCasa             uch
509
  1.87  0.031   35.40   26.39  21.70  EnVozAlta            CómoLee                  nVo
161
  1.78  0.037   36.44   19.91  15.90  poco                 FrecuenciaLectura        oco
118
-------------------------------------------------------------------------------------------
---
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                    DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
 99.99  0.000  100.00  100.00   6.60  suspenso             Calificación             usp
49
  5.68  0.000   17.58   59.18  22.24  MB[1a13.5]           InteligenciaGlobal       B[1
165
  2.68  0.004   11.80   38.78  21.70  EnVozAlta            CómoLee                  nVo
161
  2.68  0.004   11.80   38.78  21.70  Trabajo              CuandoLee                rab
161
  2.53  0.006   12.71   30.61  15.90  poco                 FrecuenciaLectura        oco
118
  2.40  0.008   13.64   24.49  11.86  no                   GustoEscuela             o
88
  2.35  0.009    9.54   55.10  38.14  AlgunaDificultad     DificultaLectura         lgu
283
  2.20  0.014   21.74   10.20   3.10  MuchaDificultad      DificultaLectura         uch
23
  1.64  0.051    8.26   61.22  48.92  niña                 Sexo                     iña
363
-------------------------------------------------------------------------------------------
---
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                    DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
```

```
27.98  0.000  100.00  100.00  22.91  suficiente          Calificación              ufi
170
  6.95  0.000   44.24   42.94  22.24  MB[1a13.5]          InteligenciaGlobal        B[1
165
  5.45  0.000   33.92   56.47  38.14  AlgunaDificultad    DificultaLectura          lgu
283
  2.75  0.003   27.18   60.59  51.08  niño                Sexo                      iño
379
  2.62  0.004   31.06   29.41  21.70  Trabajo             CuandoLee                 rab
161
  1.80  0.036   27.67   33.53  27.76  bastantes           LibrosEnCasa              ast
206
  1.76  0.040   29.66   20.59  15.90  poco                FrecuenciaLectura         oco
118
-------------------------------------------------------------------------------------------
---
```

B.2.12. Descripción del sexo

CARACTERISATION PAR LES QUESTIONS DE Sexo

V.TEST	PROBA	NUM . LIBELLE DE LA QUESTION	KHI-2	DEG.LIB	INF.A 5
99.99	**0.000**	**11 . Sexo**	**738.00**	1	0
7.41	0.000	6 . CómoLee	60.83	2	0
4.85	0.000	4 . LibrosDadosMaestro	24.85	1	0
4.63	0.000	9 . TipoEsacuela	22.78	1	0
3.20	0.001	7 . GustoTextosEscuela	14.58	2	0
3.14	0.001	8 . GustoEscuela	14.18	2	0
2.93	0.002	5 . CuandoLee	12.77	2	0
2.82	0.002	10 . Calificación	16.54	4	0

CARACTERISATION PAR LES MODALITES DES CLASSES OU MODALITES DE Sexo

```
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                  DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
 31.82  0.000  100.00  100.00  51.08  niño                Sexo                      iño
379
  7.82  0.000   59.56   86.28  73.99  EnSilencio          CómoLee                   nSi
549
  5.07  0.000   73.83   20.84  14.42  NoGustan            LibrosDadosMaestro        oGu
107
  4.79  0.000   57.40   74.67  66.44  publica             TipoEsacuela              ubl
493
  3.57  0.000   69.32   16.09  11.86  no                  GustoEscuela              o
88
  3.43  0.000   65.12   22.16  17.39  NoGustan            GustoTextosEscuela        oGu
129
  3.27  0.001   62.73   26.65  21.70  Trabajo             CuandoLee                 rab
161
  2.75  0.003   60.59   27.18  22.91  suficiente          Calificación              ufi
170
  2.15  0.016   55.69   47.76  43.80  pr.industria        TrabajoPadre              r.i
325
  2.06  0.020   60.17   18.73  15.90  poco                FrecuenciaLectura         oco
118
-------------------------------------------------------------------------------------------
---
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                  DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
```

```
31.82  0.000  100.00  100.00  48.92  niña                 Sexo                      iña
363
  7.00  0.000   73.29   32.51  21.70  EnVozAlta            CómoLee                   nVo
161
  5.07  0.000   52.76   92.29  85.58  Gustan               LibrosDadosMaestro        ust
635
  4.79  0.000   61.45   42.15  33.56  privada              TipoEsacuela              riv
249
  3.66  0.000   51.64   91.18  86.39  si                   GustoEscuela              i
641
  2.88  0.002   57.41   34.16  29.11  bien                 Calificación              ien
216
  2.50  0.006   71.88    6.34   4.31  ambos                CómoLee                   mbo
32
  2.30  0.011   51.20   82.37  78.71  Gustan               GustoTextosEscuela        ust
584
  2.18  0.015   55.78   30.58  26.82  Ambas                CuandoLee                 mba
199
  1.64  0.051   65.52    5.23   3.91  aVeces               GustoTextosEscuela        Vec
29
  1.64  0.051   61.22    8.26   6.60  suspenso             Calificación              usp
49
-------------------------------------------------------------------------------------------
---
```

B.2.13. Descripción del trabajo del padre

```
CARACTERISATION PAR LES QUESTIONS  DE TrabajoPadre
+--------+--------+---------------------------------+---------+---------+---------+
| V.TEST |  PROBA | NUM . LIBELLE DE LA QUESTION    |  KHI-2  | DEG.LIB | INF.A 5 |
+--------+--------+---------------------------------+---------+---------+---------+
|  79.02 |  0.000 |  12 . TrabajoPadre              | 6678.00 |    81   |   72    |
|   7.01 |  0.000 |   9 . TipoEsacuela              |   75.57 |     9   |    4    |
|   2.72 |  0.003 |  13 . InteligenciaGlobal        |   75.02 |    45   |   17    |
+--------+--------+---------------------------------+---------+---------+---------+

CARACTERISATION PAR LES MODALITES DES CLASSES OU MODALITES   DE TrabajoPadre
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                  DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
 31.65  0.000  100.00  100.00  43.80  pr.industria         TrabajoPadre              r.i
325
  4.99  0.000   50.30   76.31  66.44  publica              TipoEsacuela              ubl
493
  2.81  0.003   54.36   24.92  20.08  B[13.5a28.5]         InteligenciaGlobal        [13
149
  2.15  0.016   47.76   55.69  51.08  niño                 Sexo                      iño
379
-------------------------------------------------------------------------------------------
---
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                  DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
                                0.94  jubilado                                  ubi
7
  1.69  0.046    3.23   42.86  12.53  MeB[28.5a45]         InteligenciaGlobal        eB[
93
-------------------------------------------------------------------------------------------
---
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                  DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
```

```
26.46  0.000  100.00  100.00  18.73  oficios              TrabajoPadre              fic
139
  2.02  0.022   25.58   23.74  17.39  NoGustan             GustoTextosEscuela        oGu
129
  1.75  0.040   22.49   40.29  33.56  privada              TipoEsacuela              riv
249
  1.66  0.049   34.78    5.76   3.10  MuchaDificultad      DificultaLectura          uch
23
  1.65  0.049   23.60   27.34  21.70  EnVozAlta            CómoLee                   nVo
161
  1.64  0.051   21.91   44.60  38.14  AlgunaDificultad     DificultaLectura          lgu
283
-------------------------------------------------------------------------------------------
---
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                    DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
 99.99  0.000  100.00  100.00   4.58  funcionario          TrabajoPadre              unc
34
  1.96  0.025    8.05   35.29  20.08  B[13.5a28.5]         InteligenciaGlobal        [13
149
  1.64  0.051    5.73   73.53  58.76  Facilidad            DificultaLectura          aci
436
  1.62  0.053    6.90   41.18  27.36  mucho                FrecuenciaLectura         uch
203
  1.62  0.053   12.50   11.76   4.31  ambos                CómoLee                   mbo
32
-------------------------------------------------------------------------------------------
---
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                    DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
 99.99  0.000  100.00  100.00   5.26  adm.banc.emp.        TrabajoPadre              dm.
39
  4.18  0.000   10.44   66.67  33.56  privada              TipoEsacuela              riv
249
  2.04  0.021   10.42   25.64  12.94  MA[83.5a98]          InteligenciaGlobal        A[8
96
-------------------------------------------------------------------------------------------
---
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                    DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
 99.99  0.000  100.00  100.00   5.39  pr.liberales         TrabajoPadre              r.l
40
  4.02  0.000   10.44   65.00  33.56  privada              TipoEsacuela              riv
249
  1.97  0.025   10.42   25.00  12.94  MA[83.5a98]          InteligenciaGlobal        A[8
96
  1.76  0.039    9.32   27.50  15.90  A[65a83.5]           InteligenciaGlobal        [65
118
-------------------------------------------------------------------------------------------
---
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                    DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
 99.99  0.000  100.00  100.00   5.12  parado               TrabajoPadre              ara
38
  4.06  0.000    7.30   94.74  66.44  publica              TipoEsacuela              ubl
493
  2.97  0.002    9.01   55.26  31.40  mucho                FrecuenciaLecturaEscuela  uch
233
  2.20  0.014    8.37   44.74  27.36  mucho                FrecuenciaLectura         uch
203
-------------------------------------------------------------------------------------------
```

```
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                    DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
 19.29  0.000  100.00  100.00   7.28  comercio             TrabajoPadre              ome
54
  2.46  0.007   10.84   50.00  33.56  privada              TipoEsacuela              riv
249
  1.81  0.035   12.50   22.22  12.94  MA [83.5a98]          InteligenciaGlobal        A[8
96
-------------------------------------------------------------------------------------------
---
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                    DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
  2.18  0.014    3.39   57.14  15.90  A[65a83.5]           InteligenciaGlobal        [65
118
  1.60  0.055    1.65   85.71  48.92  niña                 Sexo                      iña
363
-------------------------------------------------------------------------------------------
---
```

B.2.14. Descripción del nivel de inteligencia global

```
CARACTERISATION PAR LES QUESTIONS  DE Inteligencia Global
+--------+--------+---------------------------------+---------+---------+---------+
| V.TEST |  PROBA | NUM . LIBELLE DE LA QUESTION    |  KHI-2  | DEG.LIB | INF.A 5 |
+--------+--------+---------------------------------+---------+---------+---------+
|  59.70 |  0.000 |  13 . InteligenciaGlobal        | 3710.00 |    25   |    0    |
|  13.06 |  0.000 |  10 . Calificación              |  238.16 |    20   |    0    |
|   7.67 |  0.000 |   9 . TipoEsacuela              |   75.09 |     5   |    0    |
|   5.45 |  0.000 |   3 . DificultaLectura          |   55.57 |    10   |    5    |
|   4.89 |  0.000 |   5 . CuandoLee                 |   48.44 |    10   |    0    |
|   4.22 |  0.000 |  16 . FrecuenciaLectura         |   40.84 |    10   |    0    |
|   2.72 |  0.003 |  12 . TrabajoPadre              |   75.02 |    45   |   17    |
+--------+--------+---------------------------------+---------+---------+---------+
CARACTERISATION PAR LES MODALITES DES CLASSES OU MODALITES   DE Inteligencia Global
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                    DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
 23.34  0.000  100.00  100.00  12.53  MeB [28.5a45]         InteligenciaGlobal        eB[
93
  2.55  0.005   14.39   84.95  73.99  EnSilencio           CómoLee                   nSi
549
  1.73  0.041   25.93    7.53   3.64  pocos                LibrosEnCasa              oco
27
-------------------------------------------------------------------------------------------
---
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                    DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
```

```
27.76  0.000  100.00  100.00  22.24  MB [1a13.5]           InteligenciaGlobal        B[1
165
  6.95  0.000   42.94   44.24  22.91  suficiente           Calificación              ufi
170
  5.68  0.000   59.18   17.58   6.60  suspenso             Calificación              usp
49
  5.68  0.000   39.75   38.79  21.70  Trabajo              CuandoLee                 rab
161
  5.67  0.000   33.57   57.58  38.14  AlgunaDificultad     DificultaLectura          lgu
283
  4.82  0.000   27.38   81.82  66.44  publica              TipoEsacuela              ubl
493
  4.65  0.000   39.83   28.48  15.90  poco                 FrecuenciaLectura         oco
118
  2.11  0.017   43.48    6.06   3.10  MuchaDificultad      DificultaLectura          uch
23
  2.05  0.020   28.57   27.88  21.70  EnVozAlta            CómoLee                   nVo
161
  1.85  0.032   30.68   16.36  11.86  no                   GustoEscuela              o
88
  1.60  0.055   37.04    6.06   3.64  pocos                LibrosEnCasa              oco
27
-------------------------------------------------------------------------------------------
---
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                    DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
 25.39  0.000  100.00  100.00  16.31  MeA[45a65]           InteligenciaGlobal        eA[
121
  2.49  0.006   21.82   39.67  29.65  notable              Calificación              ota
220
  2.27  0.012   20.88   42.98  33.56  privada              TipoEsacuela              riv
249
  1.71  0.044   18.35   66.12  58.76  Facilidad            DificultaLectura          aci
436
-------------------------------------------------------------------------------------------
---
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                    DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
 26.99  0.000  100.00  100.00  20.08  B[13.5a28.5]         InteligenciaGlobal        [13
149
  4.54  0.000   24.75   81.88  66.44  publica              TipoEsacuela              ubl
493
  2.81  0.003   24.92   54.36  43.80  pr.industria         TrabajoPadre              r.i
325
  2.61  0.005   26.39   38.26  29.11  bien                 Calificación              ien
216
  1.96  0.025   35.29    8.05   4.58  funcionario          TrabajoPadre              unc
34
  1.59  0.055   21.10   89.93  85.58  Gustan               LibrosDadosMaestro        ust
635
-------------------------------------------------------------------------------------------
---
-------------------------------------------------------------------------------------------
---
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                    DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
 25.19  0.000  100.00  100.00  15.90  A[65a83.5]           InteligenciaGlobal        [65
118
  4.14  0.000   24.10   50.85  33.56  privada              TipoEsacuela              riv
249
  3.12  0.001   22.73   42.37  29.65  notable              Calificación              ota
220
  2.29  0.011   25.29   18.64  11.73  sobresaliente        Calificación              obr
87
  1.76  0.039   27.50    9.32   5.39  pr.liberales         TrabajoPadre              r.l
40
-------------------------------------------------------------------------------------------
---
```

```
V.TEST  PROBA ---- POURCENTAGES ----  MODALITES                                    DEN
POIDS
              CLA/MOD MOD/CLA GLOBAL  CARACTERISTIQUES     DES VARIABLES
-------------------------------------------------------------------------------------------
---
 23.59  0.000  100.00  100.00  12.94  MA[83.5a98]          InteligenciaGlobal        A[8
96
  7.13  0.000   41.38   37.50  11.73  sobresaliente        Calificación              obr
87
  4.57  0.000   21.29   55.21  33.56  privada              TipoEsacuela              riv
249
  3.61  0.000   20.69   43.75  27.36  mucho                FrecuenciaLectura         uch
203
  3.43  0.000   16.51   75.00  58.76  Facilidad            DificultaLectura          aci
436
  3.04  0.001   19.09   43.75  29.65  notable              Calificación              ota
220
  2.70  0.004   15.85   70.83  57.82  bastante             FrecuenciaLecturaEscuela  ast
429
  2.12  0.017   17.59   36.46  26.82  Ambas                CuandoLee                 mba
199
  2.08  0.019   14.73   78.13  68.60  muchos               LibrosEnCasa              uch
509
  2.04  0.021   25.64   10.42   5.26  adm.banc.emp.        TrabajoPadre              dm.
39
  1.97  0.025   25.00   10.42   5.39  pr.liberales         TrabajoPadre              r.l
40
  1.91  0.028   14.39   82.29  73.99  EnSilencio           CómoLee                   nSi
549
  1.81  0.035   22.22   12.50   7.28  comercio             TrabajoPadre              ome
54
-------------------------------------------------------------------------------------------
---
```

C. Estudio del "El color en la comunicación socia" [César R., Correa M. T., 1999].

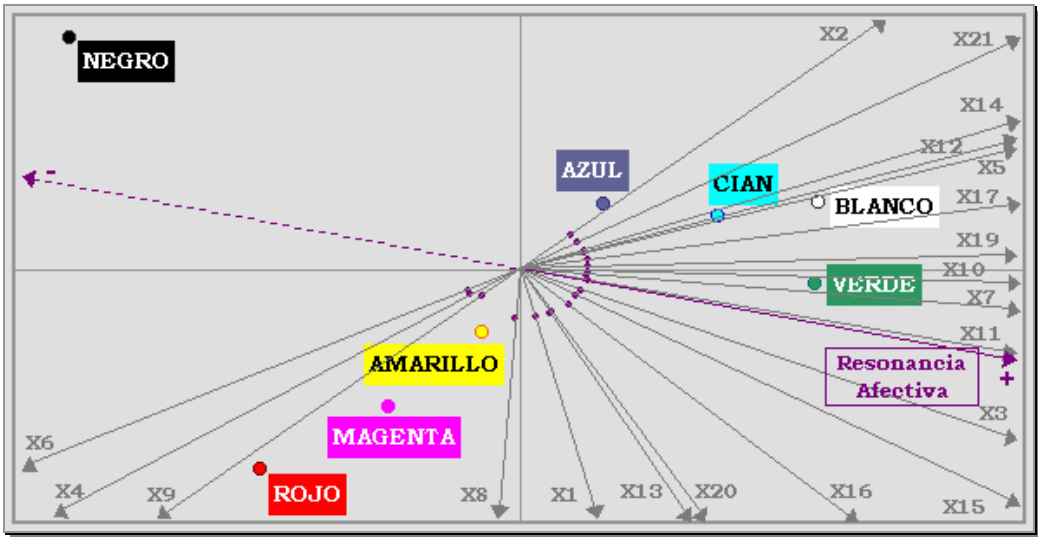
C.1. Análisis de la resonancia afectiva (RA) de cada color. Cartografiado

Con datos obtenidos del Diferencial Semántico, se efectuó el análisis de la resonancia afectiva (RA) de cada color estudiado, para todos los sujetos de la muestra. Para ello se construyó la Tabla de Promedios por cada color, de las calificaciones obtenidas de todas las variables de escala (los adjetivos). De acuerdo a lo establecido en la técnica del Diferencial Semántico (DS), que establece una escala entre el rango 1 a 7, los valores promedios de puntajes de todas las escalas, para cada color establece la Resonancia Afectiva (RA) según los siguientes criterios:

1. (Promedio entre 1 y 4) = Resonancia Afectiva Negativa RA (-) negativa.
2. (Promedio igual a 4) = Resonancia Afectiva RA indiferente.
3. (Promedio entre 4 y 7) = Resonancia Afectiva RA (+) positiva.

El gráfico factorial de las componentes principales (ACP), extraídos con los promedios de escala, por cada color y para todos los sujetos de la muestra, visualiza las relaciones con la (RA) global de cada color y las variables de escala asociadas

COMPONENTES PRINCIPALES (ACP)



Cuando dos *variables continuas* (X1 a X21, promedios por cada color de todos los individuos) *están perfectamente correlacionadas*, el ángulo entre ellas, en el plano, será cercano a cero y los vectores se superpondrán. Cuando las *dos variables son directamente opuestas* el ángulo entre ellas es de 180º. Las variables sin relación alguna entre sí en la variación, se visualizan por medio de *dos vectores en ángulo recto* en el plano (ortogonales). La dirección de los vectores indica la dimensión con la cual cada variable está asociada. La proximidad entre un punto – individuo (los colores) y la dirección de la variable, significa en "promedio" que esta variable tiene un valor "ALTO" o "BAJO" para este individuo y la asocia fuertemente.

De esta manera se obtienen como "positivos" en la Resonancia Afectiva, los colores Verde, Blanco y Cian; "indiferentes" el Rojo, Azul y Amarillo y los de Resonancia afectiva "Negativa" el Negro.

C.2. Visualización de los estados de ánimo licitados por los colores. Cartografiado

Para obtener la información completa (relaciones lineales y no lineales), se hace necesario extraer la información con mayor "Contraste". Para ello, los valores cuantitativos de las escalas de adjetivos (los promedios), se "Discretizan" en rangos o clases, transformando a esos valores, en variables cualitativas (nominales), en cinco modalidades cada una. El criterio de Discretización es el siguiente:

REFERENCIAS PARA LA INTERPRETACIÓN de LOS RESULTADOS

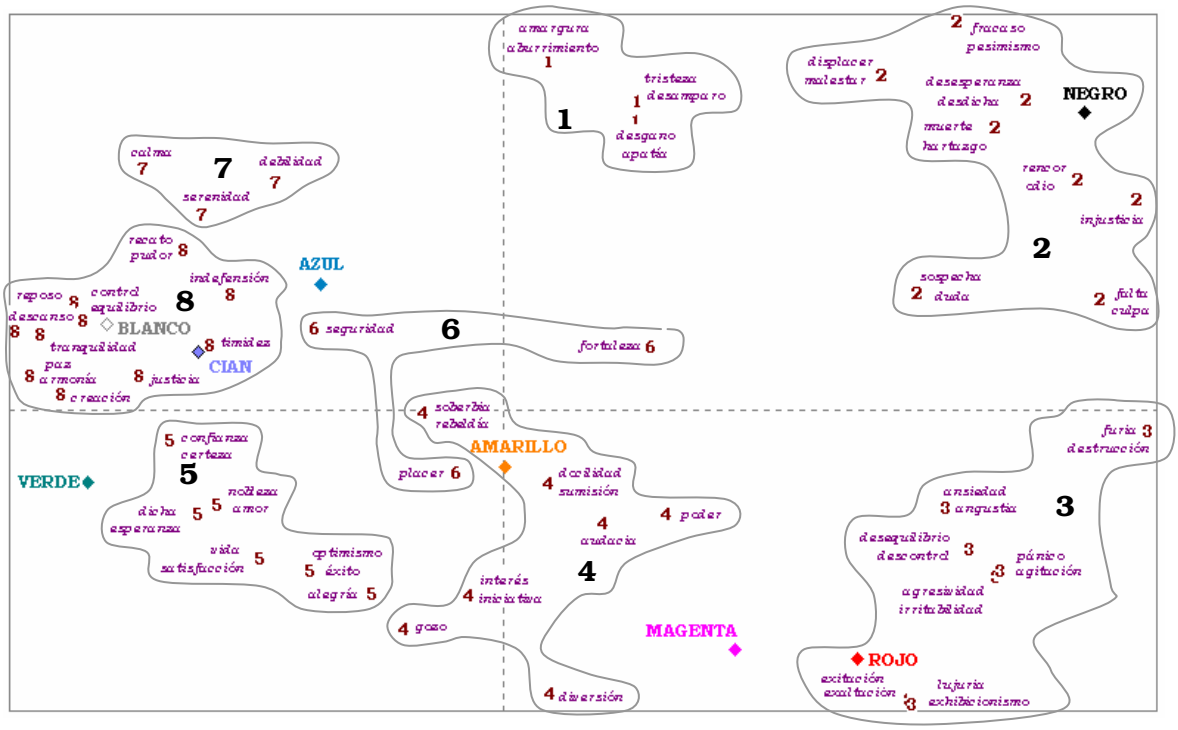
	ALTO+	ALTO-	MEDIO+	nsnc	MEDIO-	BAJO-	BAJO+
	7	6	5	4	3	2	1
	Alta connotación de	Media connotación de	Baja connotación de		Alta connotación de	Media connotación de	Baja connotación de
X1	Iniciativa / Interés			INDIFERENTE	Apatía / Desgano		
X2	Pudor / Recato				Lujuria / Exhibicionismo		
X3	Nobleza / Amor				Rencor / Odio		
X4	Fortaleza / Audacia				Timidez / Debilidad		
X5	Calma / Descanso				Agresividad / Irritabilidad		
X6	Fortaleza / Poder				Debilidad / Indefensión		
X7	Justicia				Injusticia		
X8	Diversión / Placer				Aburrimiento / Amargura		
X9	Sumisión / Docilidad				Rebeldía / Soberbia		
X10	En Paz / En Armonía				Culpa / Falta		
X11	Confianza / Certeza				Sospecha / Duda		
X12	Control / Equilibrio				Descontrol / Desequilibrio		
X13	Alegría / Seguridad				Tristeza / Desamparo		
X14	Calma / Seguridad				Agitación / Pánico		
X15	Vida / Satisfacción				Muerte / Hartazgo		
X16	Éxito / Optimismo				Pesimismo / Fracaso		
X17	Tranquilidad / Serenidad				Ansiedad / Angustia		
X18	Dicha / Esperanza				Desdicha / Desesperanza		
X19	Creación / Serenidad				Destrucción / Furia		
X20	Gozo / Placer				Malestar / Displacer		
X21	Reposo / Serenidad				Excitación / Exaltación		

Az Medio+ X14 → *significa una Baja connotación de calma / seguridad en el azul (casi indiferente)*

Se obtiene así el CARTOGRAFIADO DE LOS DATOS, que visualiza con mayor contraste de información las relaciones entre colores y connotaciones. [César, 2005]

En el *gráfico factorial* de los adjetivos se proyecta la distribución de los colores en función de los adjetivos que connotan. Se obtiene una visualización de la naturaleza de los estados de ánimo licitados por los colores.

Las distancias cercanas de puntos en el plano, indican “asociación” entre el color con el adjetivo. Las distancias lejanas, indican rechazo y puede decirse que connotan lo opuesto.



Así por ejemplo el COLOR VERDE puede analizarse de la siguiente forma: los factores asociados son de alto+ valores connotativos en sentimientos de *calma*, *descanso*, *justicia*, *paz*, *armonía*, *control* y *equilibrio*; *seguridad*, *serenidad*, *esperanza*, *dicha* y *nobleza*.

1. El estado de ánimo **"apatía"**, es fundamentalmente transmitido por el *negro* y también en menor medida por el *azul*. *Aburrimiento, amargura, apatía, desamparo, desgano, tristeza*
2. El estado de ánimo **"deterioro"**, es transmitido por el *negro*, sobre todo. *culpa, desdicha, desesperanza, displacer, duda, falta, fracaso, hartazgo, injusticia, malestar, muerte, odio, pesimismo, rencor, sospecha*
3. El estado de ánimo **"depresión"**, es fundamentalmente transmitido por el *rojo*, además de por el *magenta*, *negro*; y en menor medida por el *amarillo*. *Agitación, agresividad, angustia, ansiedad, descontrol, desequilibrio, destrucción, exaltación, exhibicionismo, excitación, furia, irritabilidad, lujuria, pánico*
4. El estado de ánimo **"poder"**, es fundamentalmente transmitido por el *magenta* y *rojo* y en menor medida por el *verde*. *Audacia, diversión, docilidad, gozo, iniciativa, interés, poder, rebeldía, soberbia, sumisión*
5. El estado de ánimo **"alegría"**, es fundamentalmente transmitido por el *Verde*, y en mucha menor medida por *blanco*, *cian* y *magenta*. *Alegría, amor, certeza, confianza, dicha, esperanza, éxito, nobleza, optimismo, satisfacción, vida*
6. El estado de ánimo **"seguridad"**, es fundamentalmente transmitido por el *Rojo*, además del *cian*, *amarillo*, *azul* y también *blanco*. *Fortaleza, placer, seguridad*
7. El estado de ánimo **"calma"**, es fundamentalmente transmitido por el *blanco* y el *cian*, además del *azul* y en menos medida por el *amarillo*. *calma , debilidad, serenidad*
8. El estado de ánimo **"armonía"**, es fundamentalmente transmitido por el *verde* y *blanco*, además del *cian* y en menor medida el *azul*. *Armonía, control, creación, descanso, equilibrio, indefensión, justicia, paz, pudor, recato, reposo, timidez, tranquilidad*