UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

ALGUNOS CLASIFICADORES BAYESIANOS

MONOGRAFÍA DE TRABAJO DE GRADO PARA OPTAR POR EL TÍTULO DE MATEMÁTICO
PROYECTO CURRICULAR DE MATEMÁTICAS

Monica Godoy Amado
Dirigido por: Luis Fernando Villarraga Poveda

Bogotá DC
Octubre de 2021

Resumen: *En la presente monografía se da lugar al estudio de las características que poseen las redes, las cuales permiten aprender sobre relaciones de dependencia, causalidad y d-separación, seguido a esto, se presentan las redes Bayesianas junto con la aplicación de la regla de la cadena y finalmente se da una introducción al estudio de algunos clasificadores bayesianos, que son tipos concretos de redes bayesianas, dando énfasis en el funcionamiento de los clasificadores de Naive Bayes, Naive Bayes aumentado a Arbol (TAN) y los K-dependientes (KDB), con el objetivo de ser finalmente aplicados a una base de datos de coronavirus en Colombia.*

Palabras clave: Claificadores, Naive-Bayes, TAN, k-dependientes

Agradecimientos: Agradezco primeramente a Dios por permitirme terminar esta etapa, a mi familia por su apoyo incondicional y en especial a mis amados hijos por creer en mi.

1. Introducción

La estadística bayesiana, es una rama de la matemática aplicada que tuvo auge en el siglo XIX, con los trabajos del matemático y físico francés Pierre Simón Laplace, quien realizo la formulación moderna, planteada inicialmente por Thomas Bayes; el avance de esta rama trajo consigo el planteamiento de las funciones "clasificadoras" que se aplican a determinados algoritmos, logrando la asociación de datos a una categoría, para el análisis y tratamiento de datos de forma ágil.

En el presente trabajo se dará una introducción al estudio de algunos clasificadores Bayesianos, que son una herramienta eficaz en el análisis de datos de gran dimensión, fundamentalmente en Machine learning y Big Data donde se hace una clasificación de sucesos con respecto a la información conocida (a priori) con el fin de encontrar, *el valor más probable* de la variable clase, dado el valor que toman sus variables características. La principal motivación para hacer este trabajo radico en ahí poca literatura en español que logre compilar todas la definiciones y herramientas de aplicación en una sola notación con su respectiva formalización matemática, de tal forma que pueda ser útil y entendible para la aplicación en cualquier rama de estudio sean ciencias de la salud, humanas, entre otras.

Esta monografía presenta los clasificadores de Naive Bayes, Naive Bayes aumentado a Arbol (TAN) y los K-dependientes (KDB), de una forma sencilla y clara, con el fin de ser consultados por cualquier persona, siendo aterrizados estos conceptos finalmente a través del software "Weka" que permite la aplicación de los modelos, dando análisis a la base de datos de coronavirus en Colombia.

2. Clasificadores Bayesianos

Para trabajar los clasificadores bayesianos es importante tener claros varios conceptos de probabilidad (eventos independientes, probabilidad condicionada, probabilidad conjunta, la ley multiplicativa de probabilidad, teorema de Bayes, entre otros) y teoría de grafos (nodo, arco, camino dirigido, un padre $pa(X)$, grafo conexo, acíclico y cíclico), dando así una introducción a las redes bayesianas que juegan un papel primordial en la comprensión de los clasificadores, por lo cual es necesario tener presente que las variables (eventos) las notaremos como características o clases que tienen un conjunto de estados contables o continuos, pero a lo largo de este trabajo se considera únicamente variables con un número finito de estados contables y cuando el estado de una variable es conocido, se dice que es *instanciada*,

Analizando el comportamiento de una red considere la situación en la figura 1, donde A tiene influencia sobre B, que a su vez tiene influencia en C, así la evidencia en A influirá en la certeza de B, que influye a su vez en la certeza de C, de esta forma la evidencia puede transmitirse a través de una *conexión en serie* [4] a menos que se conozca el estado de la variable en la conexión, es decir, si el estado de B se conoce, entonces el canal se bloquea, quedando A y C independientes,

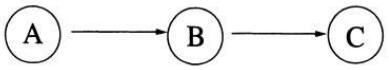


Figura 1: Ejemplo conexión en serie, cuando B es instanciado, bloquea la comunicación entre A y C

Ahora si la influencia pasa entre todos los hijos de A, como se muestra en la Figura 2, es una *conexión divergente* [4] a menos que A sea instanciado

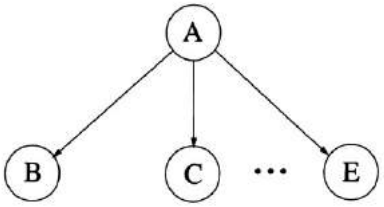


Figura 2: Ejemplo conexión divergente, si A esta instanciado, bloquea la comunicación entre sus hijos.

Si no se sabe nada sobre A, excepto lo que puede inferirse del conocimiento de sus padres $B, \cdots, E$, como se muestra en la Figura 3, es una *conexión convergente* [4], si los padres son independientes y la evidencia sobre uno de ellos no tiene influencia en la certeza de los demás donde el conocimiento de

una posible causa de un evento no dice nada sobre otras posibles causas, así la evidencia solo puede transmitirse si la variable en la conexión o uno de sus descendientes es instanciado.

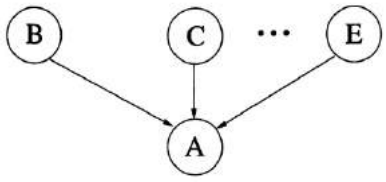


Figura 3: Ejemplo conexión convergente, si A cambia la certeza, se abre para la comunicación entre sus padres.

Definición 1. (d-separadas) [4] Dos variables distintas A y B en una red casual están *d-separadas* si, para todos los caminos entre A y B, hay un variable intermedia V (distinta de A y B) tal que

- La conexión es serial o divergente y V es instanciado ó
- La conexión es convergente, y ni V, ni ninguno de los descendientes de V han recibido evidencia.

Si A y B no están d-separadas, son *d-conectadas*.

Ejemplo 1. En la figura 1 diremos que A y C son d-separados dado B, en la figura 2 decimos que $B, C, \cdots, E$ están d-separados dado A

Definición 2. Una *red bayesiana* [4] consiste en:

- Un conjunto de variables y arcos dirigidos
- Cada variable tiene un conjunto finito de estados mutuamente excluyentes.
- Un grafo dirigido acíclico
- Para cada variable A con los padres $pa(A) = \{B_1, \ldots, B_n\}$ tiene una distribución de probabilidad conjunta $P(A|B_1, \ldots, B_n) = P(A|pa(A))$

En redes bayesianas las conexiones representan dependencia estadística entre las variables, si la parte grafica de una red causal se mantiene para redes bayesianas entonces la d-separación corresponde a independencia, por consiguiente, si en una red bayesiana dos variables X_1 y X_2 están d-separadas dado un conjunto C, se dice que son independientes dado C, es decir $P(X_1|X_2, C) = P(X_1|C)$, para hacer más clara esta propiedad considérese $P(X)$ una distribución de probabilidad sobre un conjunto

de variables $X = \{X_1, \ldots, X_n\}$, aplicando sobre la distribución la ley multiplicativa de probabilidad repetidamente se obtiene un producto de distribuciones de probabilidad condicional

$$\begin{aligned} P(X) &= P(X_1, X_2, \ldots, X_n) \\ &= P(X_1 \cap (X_2 \cap \cdots \cap X_n)) \\ &= P(X_1 | X_2 \cap \cdots \cap X_n) P(X_2 \cap \cdots \cap X_{n-1} \cap X_n) \\ &\vdots \\ &= P(X_1 | X_2 \cap \cdots \cap X_n) P(X_2 | X_3 \cap \cdots \cap X_n) \cdots P(X_{n-1} | X_n) \end{aligned}$$

la anterior propiedad es conocida como la *regla de la cadena*[4]. En el caso de que los nodos o variables de una red bayesiana sean $X_1, X_2, \ldots, X_n$, las relaciones de dependencia e independencia condicionada en la red permiten demostrar que la probabilidad conjunta viene dada como se muestra en el siguiente propiedad.

Regla de la cadena para redes Bayesianas 1. *Sea una red Bayesiana (BN) sobre* $X = \{X_1, X_2, \ldots, X_n\}$, *entonces* BN *tiene específicamente una única distribución de probabilidad conjunta* $P(X)$, *dada por:*

$$P(X) = \prod_{i=1}^{n} P(X_i | pa(X_i))$$

donde $pa(X_i)$ *son los padres de* X_i *en* BN.

Clasificadores

La clasificación de sucesos con respecto a alguna información conocida (a priori), se conoce como problema de clasificación, el cual consiste en asignar un objeto descrito por un conjunto características $X_j = (x_{j1}, \cdots, x_{jm})$, a una de las clases posibles, $c_1, c_2, \cdots, c_k$ de C, para $j \in \{1, 2, \cdots, m\}$ y $k \leq j$, buscando así que la probabilidad de pertenecer a la clase, dados los atributos, sea la máxima, es decir:

$$f(X_j) = c_i \quad \text{si} \quad (\max_{c}(P(C|X_1, X_2 \cdots X_m)) = \quad (\max_{c_i \in C}(P(c_i|X_1, X_2, \cdots, X_m))$$

Los clasificadores junto con las redes bayesianas permiten comprender las relaciones entre la causalidad y dependencia de las variables. El objetivo de los clasificadores Bayesianos es predecir el valor de la variable clase dada una determinada configuración de variables características, aportando la simplificación de premisas con el fin de que puedan ser aplicados en la resolución de problemas reales.

3. Naive Bayes

Las redes bayesianas fueron diseñadas para capturar las propiedades de independencia en los dominios que se modelan. Sin embargo, los primeros sistemas de diagnóstico bayesianos se construyeron sobre la base de modelos más simples, estos son los llamados modelos Naive Bayes (NB) también conocidos como modelo ingenuo (simple) de Bayes, son redes bayesianas donde se supone que las variables de información son independientes dada la variable de hipótesis.

Planteamiento: Sea C la variable clase, que toma K valores posibles, donde las clases c_i son excluyentes, es decir, que dos clases no pueden darse al unísono, con $X_1, \ldots, X_m$ las m variables características y sea c^* la búsqueda del diagnóstico más probable a posteriori, una vez conocidas las caracteristicas observadas, se plantea como la búsqueda del estado de la variable C con mayor probabilidad a posteriori, es decir

$$c^* = arg\left(\max_c P(C = c|X_1 = x_1, \cdots, X_m = x_m)\right) \tag{1}$$

donde $P(C = c|X_1 = x_1, \cdots, X_m = x_m)$ se calcula utilizando el teorema de Bayes,

$$P(C = c|X_1 = x_1, \ldots, X_m = x_m) = \frac{P(C = c)P(X_1 = x_1, \ldots, X_m = x_m|C = c)}{P(X_1 = x_1, \ldots, X_m = x_m)} \tag{2}$$

dado que el objetivo es calcular la clase c^* de C con mayor probabilidad a posteriori, el denominador de la ecuación 2 puede ser ignorado ya que es invariante entre las clases, por consiguiente la ecuación estará dada por

$$P(C = c|X_1 = x_1, \ldots, X_m = x_m) \propto P(C = c)P(X_1 = x_1, \ldots, X_m = x_m|C = c) \tag{3}$$

Paradigma de Naive Bayes

El clasificador Naive Bayes se basa en la suposición de que cada variable característica X_i es condicionalmente independiente de las demás características dada la clase:

$$P(X_i|X_j, C) = P(X_i|C)$$

para cualquier j, con $j \neq i$. Este planteamiento nos lleva, de manera natural a la definición de red bayesiana con la estructura que se muestra en la Figura 4

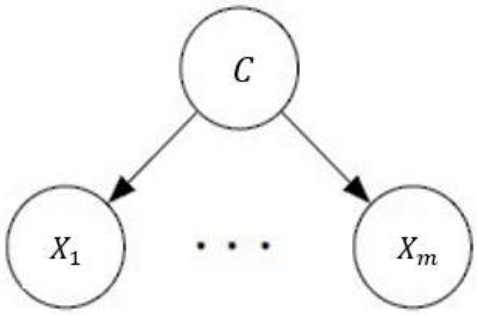


Figura 4: Un modelo de Naive Bayes

El paradigma Naive Bayes se basa en dos premisas establecidas sobre las variables predictoras (caracteristicas) y la variable a predecir (clases), dichas premisas son:

- Las clases son excluyentes, es decir, la variable C a predecir toma uno de sus k posibles valores: $c_1, \ldots, c_k$.
- Las variables caracteristicas son condicionalmente independientes dada la clase, si se conoce el valor de la variable clase, el conocimiento del valor de cualquiera de las características es irrelevante para el resto de los hallazgos.

bajo estas condiciones, aplicando la regla de la cadena para redes bayesianas, donde los padres de X_i estan dados por $pa(X_i) = C$ y que los $X_1, X_2, \ldots, X_m$ son condicionalmente independientes entre sí dada la clase C

$$P(X_1 = x_1, \ldots, X_m = x_m | C = c) = \prod_{i=1}^{m} P(X_i = x_i | C = c)$$

entonces la ecuación 3 se convierte en

$$P(C = c | X_1 = x_1, \ldots, X_m = x_m) = P(C = c) \prod_{i=1}^{m} P(X_i = x_i | C = c) \tag{4}$$

y la ecuación 1 junto con la ecuacion 4 permite calcular la clase más probable c^*, una vez conocidas las características $(x_1, \ldots, x_m)$ de un determinado evento

$$c^* = arg\left(\max_c P(C = c) \prod_{i=1}^{m} P(X_i = x_i | C = c) \right) \tag{5}$$

En general el modelo Naive Bayes proporciona un rendimiento optimo, incluso cuando se suprime el supuesto de independencia. Esto se debe en parte al hecho de que para muchos problemas de

diagnóstico es de mayor interés identificar la clase más probable, entonces el modelo Naive Bayes se puede utilizar sin afectar el rendimiento del sistema.

4. Naive Bayes aumentado a árbol (TAN)

El clasificador naive Bayes tiene una alta precisión en la clasificación de datos, pero su rendimiento decrece debido a que las caracteristicas no siempre son condicionalmente independientes como se asume. A continuación, se presentara un enfoque para resolver esta limitación

Plantiamento: Un clasificador naive Bayes aumentado a árbol (TAN), es una red bayesiana donde la variable clase no tiene padres

$$pa(C) = \emptyset$$

y cada característica $X_1, \ldots, X_m$ tiene la variable clase como padre, a excepción del nodo de clase, cada nodo de la variable característica X_i tiene como máximo otro nodo de variable característica, como su nodo principal, es decir, $pa(X_i) = \{C, X_j\}$ con $i \neq j$, donde $i, j = 1, \ldots, m$ y

$$|P_a(X_i)| \leq 2$$

por lo tanto, cada atributo puede tener un arista ascendente, si la red cumple con estas condiciones es un **árbol**. La red de la Figura 5 es un ejemplo del modelo TAN. Este clasificador busca el estado de la variable C con mayor probabilidad a posteriori, como se hizo con naive Bayes, en la ecuación 1 [10]

$$c_{TAN} = arg\left(\max_c P(C = c|X_1 = x_1, \cdots, X_m = x_m)\right) \tag{6}$$

con

$$P(C = c|X_1 = x_1, \ldots, X_m = x_m) = \mu P(C = c)P(X_1, \ldots, X_m|C = c)$$

donde $\mu = 1/P(X_1 = x_1 \ldots, X_m = x_m)$ es una constante de normalización y es invariante entre las clases. Dada la propiedad de la regla de la cadena para redes Bayesianas, se tiene que

$$P(C = c|X_1 = x_1, \ldots, X_m = x_m) = P(C = c)\prod_{i=1}^{m} P(X_i = x_i|pa(X_i)) \tag{7}$$

Por lo tanto, el clasificador TAN se puede describir a partir de

$$c_{TAN} = arg\left(\max_c P(C = c)\prod_{i=1}^{m} P(X_i = x_i|pa(X_i))\right) \tag{8}$$

donde $pa(X_i)$ tiene dos formas:

1. $pa(X_i) = c$, en el que X_i no tiene como padre una característica, se reduce a un naive Bayes.
2. $pa(X_i) = \{C, X_j\}$ donde X_i tiene como padre una característica X_j y su respectiva clase.

El paso clave, consiste en encontrar un padre diferente de la clase C, de modo que el clasificador TAN requiere un algoritmo de aprendizaje, donde se busca la estructura de árbol que más se aproxime a la distribución real de las variables, esto se basa en una medida de divergencia de información, definida de la siguiente manera

Definición 3. Sea (P) la distribución real y $(P*)$ la aproximada :

$$DI(P(X)||P(X)*) = \sum_X P(X) \log\left(\frac{P(X)}{P*(X)}\right) \tag{9}$$

esto es conocido como, la *divergencia de Kullback-Leibler*

El objetivo es minimizar DI, para esto se presenta la divergencia entre pares de variables características, a través de la siguiente función de *información mutua*[1]

$$I(X_i, X_j) = DI(P(X_i, X_j)||P(X_i)P(X_j)) \tag{10}$$

con $i \neq j$, de esta manera se puede ver la divergencia de información en función de la suma de las informaciones mutuas de todos los pares de variables características que constituyen el árbol.

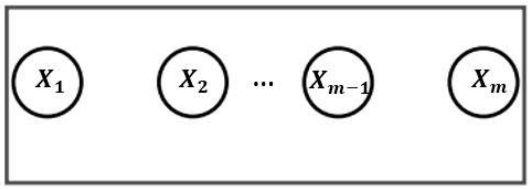
$$I(X_i, X_j) = \sum_{X_i} \sum_{X_j} P(X_i, X_j) \log\left(\frac{P(X_i, X_j)}{P(X_i)P(X_j)}\right)$$

dado que las variables características X_i y X_j estan condicionadas a la variable clase C, la informacion mutua condicionada (CMI) [8] entre ellas se define como

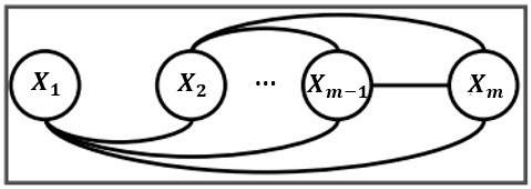
$$I(X_i, X_j|C) = \sum_C \sum_{X_i} \sum_{X_j} P(C, X_i, X_j) \log\left(\frac{P(X_i, X_j|C)}{P(X_i|C)P(X_j|C)}\right) \tag{11}$$

A partir de esta ecuación Friedman y Col (1997) establece la construcción del algoritmo denominado Tree Augmented Network (TAN), con el fin de encontrar la relación entre la cantidad de información mutua condicionada, de la siguiente forma

1. Calcule $I(X_i, X_j|C)$ entre cada par de variables con $i \neq j$

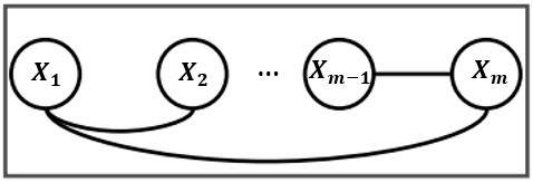


2. Construya un grafo completo no dirigido con m nodos, en el que los vértices sean las variables características $X_1, X_2, \ldots X_m$ donde la probabilidad de la arista que conecta X_i con X_j está dada por $I(X_i, X_j|C)$



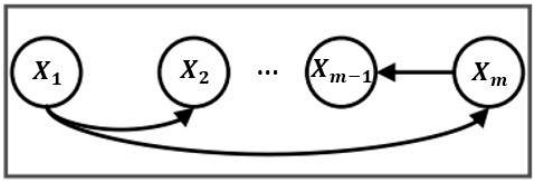
3. Asignar las dos aristas de mayor peso (probabilidad) al árbol a construir

4. Examinar la siguiente arista de mayor peso, y añadirla al árbol a no ser que forme un ciclo, en cuyo caso se descarta y se examina la siguiente arista de mayor peso.



5. Repetir el paso anterior hasta que se hayan seleccionado $m-1$ aristas.

6. Transforme el árbol no dirigido resultante en uno dirigido eligiendo una variable raiz y estableciendo la dirección de todos los vertices para que estén hacia afuera.



7. Construya un modelo TAN agregando un vértice etiquetado con C y agregando un arco desde C a cada X_i

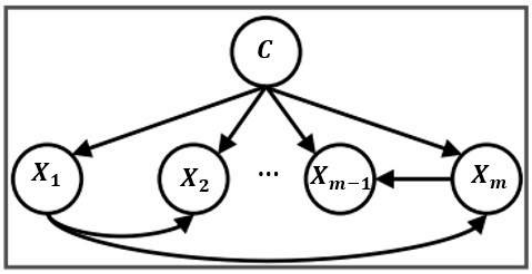


Figura 5: Un modelo de naive Bayes aumentado a árbol TAN

Ya que este modelo es una extensión natural del modelo naive Bayes, donde el espacio entre los arcos a seleccionar, tiene una relación de dependencia más fuerte entre dos atributos (características). De esta manera, los arcos seleccionados y los nodos de atributos construyen un árbol.

Usualmente los clasificadores TAN generados por este algoritmo son estables y es difícil mejorar su rendimiento de clasificación por la técnica de empaquetado (Bootstrap) que consiste en la iteración del proceso, en los pasos 3, 4 y 5 se realiza este proceso conocido como el algoritmo de Kruskall que parte de las $n(n-1)/2$ probabilidades obtenidas en el paso anterior para construir el árbol de expansión con el máximo peso[l].

5. Clasificador Bayesiano *k-dependiente*

La noción de clasificadores $k-dependientes$ fue introducida por Sahami (1996), donde la probabilidad de cada variable característica está condicionado por la clase y como máximo a otras k características, lo que da como resultante un modelo naive Bayes de dependencia k, notado por (kDB).

Plantiamiento: Un clasificador *k-dependiente* es una red bayesiana que contiene la estructura de un clasificador naive Bayes y permite que cada característica X_i tenga máximo k nodos de características como padres, es decir $P_a(X_i) = \{C, X_{d_i}\}$ donde X_{d_i} es un conjunto de como máximo k nodos de características y $pa(C) = \emptyset$.

Para dar inicio a la construcción de estos modelos, se establece un valor k, que representa , el número máximo de nodos padres que pueda tener cualquier característica, estableciendo la complejidad de los clasificadores deseados, estas van desde una estructura naive Bayes simple donde $k = 0$, hasta estructuras que abarcan otros enfoques como TAN, donde los estimadores tienen dependencia, con un padre distinto a la clase, en este caso $k = 1$, también se da en redes bayesias completamente generalizadas con $k = n$ [l]

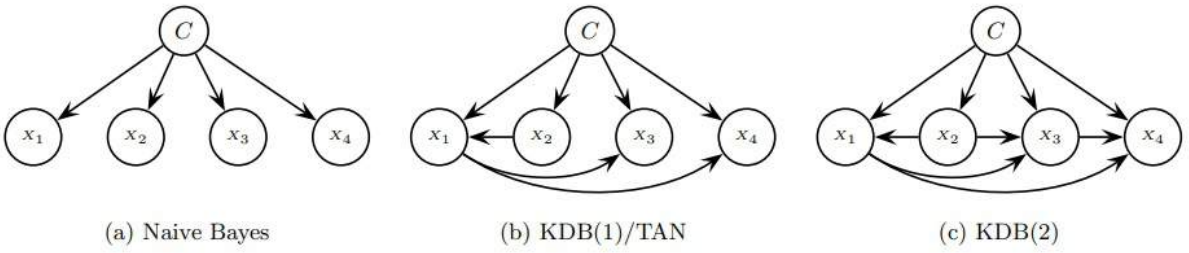


Figura 6: Ejemplo de clasificadores kDB

En la imagen 6 vemos un ejemplo del modelo Naive Bayes (a) y TAN (b) que son casos particulares de un modelo más general.

El algoritmo kDB adopta una estrategia para identificar la estructura gráfica del clasificador resultante; basandose en los conceptos de información mutua e información mutua condicional (CMI). En el proceso de construcción de la estructura kDB, todos los nodos de atributos deben clasificarse previamente en orden descendente de acuerdo con la información mutua (IM) entre cada atributo predictivo X_i y el nodo de clase C

$$I(X_i, C) = \sum_{X_i, C} P(X_i, C) log \frac{P(X_i, C)}{P(X_i)P(C)} \tag{12}$$

se inicializa un conjunto vacío S_i, para realizar un seguimiento de los nodos ya considerados, una vez que el atributo X_i ingresa al modelo, las k variables con el CMI más alto $I(X_i, X_j|C)$ serán seleccionadas como su nodo padre de $S_{i-1} = \{X_1, \ldots, X_{i-1}\}$ las variables que han entrado en el modelo.

Por lo tanto, para cualquiera de los primeros $k+1$ atributos en la secuencia, seleccionarán indiscriminadamente todos los atributos que existen frente a ellos en el modelo como sus padres, para los demás atributos, elegirán k atributos padre, correspondientes al valor máximo de $I(X_i, X_j|C)$, donde X_j se clasifica antes que X_i, entonces, la ecuación 7 se transformará:

$$P(C|X_1 = x_1, \ldots, X_m = x_m) \propto P(C) \prod_{i=1}^{m} P(X_i|C, X_{i1}, \ldots, X_{ip}) \tag{13}$$

donde $X = \{X_{i1}, \ldots, X_{ip}\}$ es el conjunto de variables principales del atributo X_i y $p = \text{mín}(i-1, k)$, obsérvese que las primeras k variables agregadas al modelo tendrán menos de k padres.

Supongamos que el orden de los atributos es $\{X_1, X_2, \ldots, X_m\}$; entonces X_i tendrá $i-1$ padres, con $i \leq k$ donde las $m-k$ variables restantes tendrán exactamente k padres y el clasificador kDB esta

dado por

$$c_k^* = arg\,\text{máx}\, P(C|X_1 = x_1, \ldots, X_m = x_m) \propto arg\,\text{máx}\left(P(C = c) \prod_{i=1}^{m} P(X_i|C, X_{i1}, \ldots, X_{ip}) \right) \qquad (14)$$

La idea básica del algoritmo consiste en generalizar el algoritmo propuesto por Fridman y Col (1997) permitiendo que cada variable tenga un número de padres, sin contar la variable clase C, acotado por k (valor arbitrario), que es el valor máximo permitido de dependencia de características, flexibilizando la determinación de k por medio de la obtención de un umbral de cantidad de información mutua, el cual debería de ser sobrepasado, para que el correspondiente arco fuese incluido, captura asi en gran parte la eficiencia computacional del modelo Naive Bayes

1. Para cada característica X_i, calcule la información mutua, $I(X_i, C)$, donde C es la clase.
2. Calcule la información mutua condicional $I(X_i, X_j|C)$, para cada par de características X_i y X_j, donde $i \neq j$.
3. Inicialice la lista de variables, $S = \{\emptyset\}$
4. Inicializar la red Bayesiana a construir, BN, con un único nodo, el correspondiente a la variable C
5. Repita hasta que S incluya todas las variables características del dominio
   a) Seleccione la característica $X_{\text{máx}}$, que no está en S y tiene el valor más grande $I(X_{\text{máx}}, C)$
   b) Agregue un nodo a BN que representa $X_{\text{máx}}$
   c) Agregue un arco de C a $X_{\text{máx}}$ en BN
   d) Añada $m = \text{mín}(|S|, k)$ arcos, de las m variables características distintas X_j en S con el valor más alto para $I(X_{\text{máx}} X_j|C)$.
   e) Agregue $X_{\text{máx}}$ a S.
6. Calcule las tablas de probabilidad condicional de acuerdo con la estructura de red Bayesiana y devuélvalo al clasificador de kDB

Nota: $X_{\text{máx}}$ es la variable característica X con mayor cantidad de información mutua respecto a C es decir $I(X_{\text{máx}}, C) = \text{máx}_{X \notin S} I(X, C)$. En esta descripción del algoritmo, el paso 5.4 requiere que agregar m padres a cada nueva característica agregada al modelo. Para hacer el algoritmo más robusto, considere una variante en el paso 5.4 considerando m características distintas X_j en S con el valor más alto para $I(X_{\text{máx}}, X_j|C)$, y solo agregar arcos de X_j a $X_{\text{máx}}$ si $I(X_{\text{máx}}; X_j|C) > \theta$, donde θ es un umbral de información mutua, esto permite una mayor flexibilidad al no forzar la inclusión de dependencias que no parecen existir cuando el valor de k se establece demasiado alto.

6. Aplicación

Para la aplicación de los clasificadores Naive, TAN y k-dependientes, se toma de la página del instituto nacional de salud de Colombia una base en la sección coronavirus a cohorte de 31 de agosto de 2021, que corresponde a la ultima base completa con todos los registros de casos. El archivo plano que se descargo contiene 4.909.086 registros con 23 variables.

Posteriormente se realizó el tratamiento de los datos, por medio del software estadístico Stata 14, el cual consistió en quitar los pacientes fallecidos y recuperados dado que sobre ellos no hay probabilidad de cambio de estado, se hizo una minería de datos sobre las variables de ubicación (ubicación del paciente, hospitalizado, en casa o en hospitalización uci), sexo; se crearon grupos etarios como, primera infancia, Infancia, Juventud, Adolescentes, Adulto y Ancianos tomando como referencia los grupos etarios que maneja el sistema de salud colombiano por medio de la UPC basados en la clasificación de la OMS.

De esta manera la base de datos queda con un total de 45.256 registros, por lo cual se decide trabajar sobre un tamaño muestral, dada la capacidad de hardware con la que se contaba en el momento y a la gran cantidad de registros contenidos en la base de datos, por lo cual, el cálculo del tamaño muestral se realizó con un nivel de confianza de 95 % y un margen de error del 5 %, lo cual corresponde a 381 registros, con ayuda del comando “sample” de Stata, se extrajo muestras aleatorias de los datos en la memoria sin remplazo, obteniendo así una nueva base.

Modelación bayesiana

La modelación se realizó por medio del software estadístico especializado en aprendizaje automático “Weka” el cual cuenta con licenciamiento libre y una sección nombrada como “Bayesian Network Classifiers” con comandos para obtener los modelos de interés, con sus respectivas funciones de aprendizaje y criterios de selección que tiene como interfaz el paquete bnclassify del software R.

Dada la fuente de información la modelación se define de la siguiente manera. La variable clase se define como la ubicación del paciente (hospitalización, hospitalización uci y casa), inicialmente se tomaron las variables características: edad, genero (Masculino y Femenino), ciclo vital (infancia, primera infancia, adolescentes, juventud, adulto y ancianos), departamento (Bogotá, Antioquia, Valle, Boyacá, Santander y Cundinamarca), fuente tipo de contagio (En estudio, Familiar, Comunitario) estado (Leve, moderado, grave).

Se sube en "Weka"la base de datos de validación y el software crea una base de datos de entrenamiento, con los cuales se corren los respectivos comandos para, Naive (weka.classifiers.bayes.NaiveBayes), TAN (weka.classifiers.bayes. net.search.local.TAN -S BAYES) y para el caso de los k-dependientes se toma k=2 el cual notaremos como el modelo 2DB (weka.classifiers.bayes.net. search.local.K2 -P3-S BAYES), en este modelo hay un máximo de tres padres contando con la clase, los cuales arrojan como resultado la siguiente tabla:

	Naive	TAN	2-DB
Instancias correctamente clasificadas	99,73 %	98,68 %	100 %
Kappa	0,99	0,97	1
RECM	12,95 %	19,50 %	5,40 %

A partir de estos criterios de decisión se puede observar, que el clasificador bayesiano 2DB es el más óptimo, con un porcentaje mayor de instancias correctamente clasificadas, siguiendo de Naive bayes y finalmente TAN, lo cual es soportado por (RECM) la raíz del error cuadrático medio que muestra que el modelo TAN tiene el mayor porcentaje (19,50 %) de diferencia entre los valores predichos por el modelo y los valores observados, sin embargo el coeficiente kappa de Cohen muestra que los tres clasificadores tienen un valor mayor de $0,81$ por consiguiente un nivel de concordancia casi perfecto en la modelación. Los clasificadores Bayesianos también permiten cotejar los datos con el fin de comprender el comportamiento de estos con respecto a la clase:

	Casa	Hospital	Hospital UCI
Departamento			
Bogotá	101	7	4
Antioquia	81	43	5
Valle	30	21	5
Boyaca	6	11	4
Santander	8	17	1
Cundinamarca	6	29	2
Total	232	128	21
	61 %	33,5 %	5,5 %
Ciclo vital			
Primera infancia	7	2	1
Infacia	9	1	1
Adolecentes	20	3	1
Juventud	52	9	3
Adulto	131	98	11
Ancianos	13	15	4
Edad			
Promedio	37,8	53,2	54,7
σ	19,2	19,4	17,3
Genero			
Hombre	100	62	11
Mujer	132	66	10

La clasificación de los datos permite analizar que el 61% de las personas que contraen coronavirus se quedan en la casa, mientras que el 33,5% requiere de atención hospitalaria y 5,5% entran en la unidad de cuidados intensivos (UCI), los ciclos vitales en que menos se presentan contagios es primera infancia e infancia, los adultos ocupan el mayor porcentaje de personas contagiadas que necesitan atención hospitalaria seguido de los ancianos, la edad promedio de contagio que se encuentra en casa es de 37,8 años, mientras que el promedio de edad para ser hospitalizado es de 53,2 años y entrar en UCI 54,7 años finalmente con respecto al género, las mujeres tienen un porcentaje mayor de contagio, aunque no requieren atención hospitalaria, el comportamiento entre hospitalización y UCI es muy similar entre hombres y mujeres. La clasificación de los datos permite observar la dependencia que ahí entre las variables características y su comportamiento, arrojando los diagramas correspondientes para cada modelo:

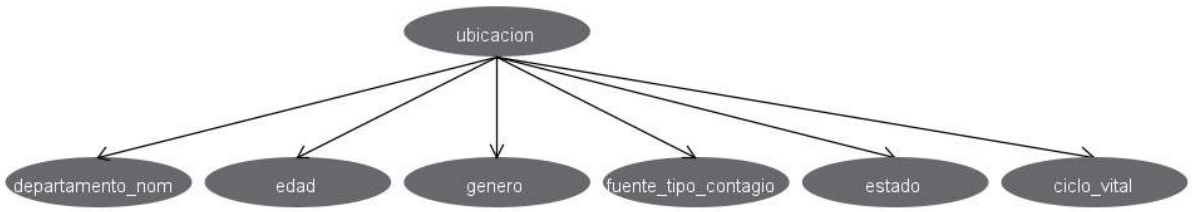


Figura 7: Modelo Naive Bayes correspondiente a la clasificación

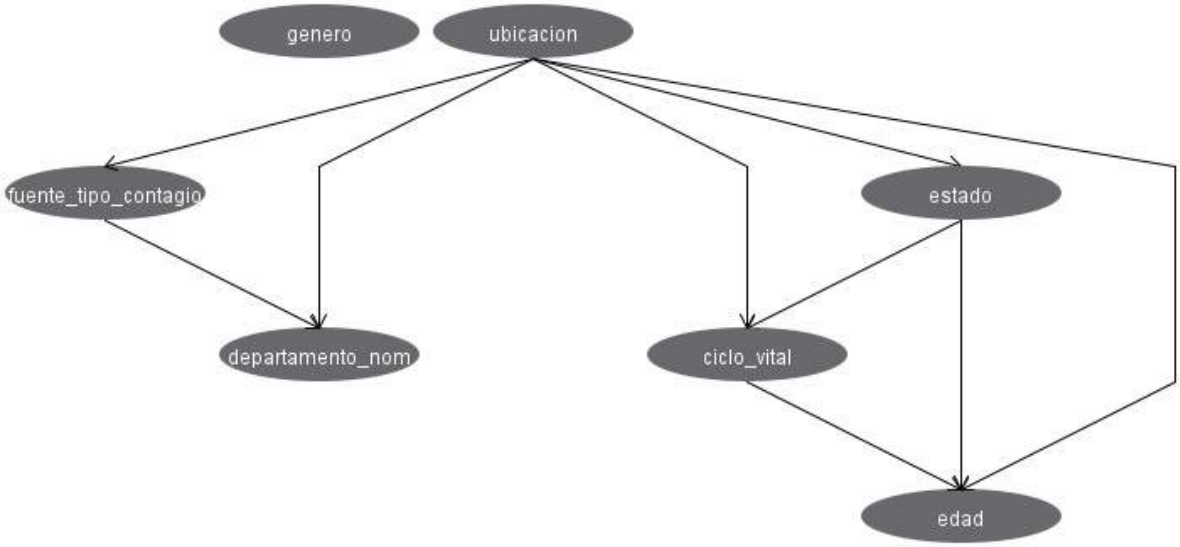


Figura 8: Modelo 2DB correspondiente a la clasificación

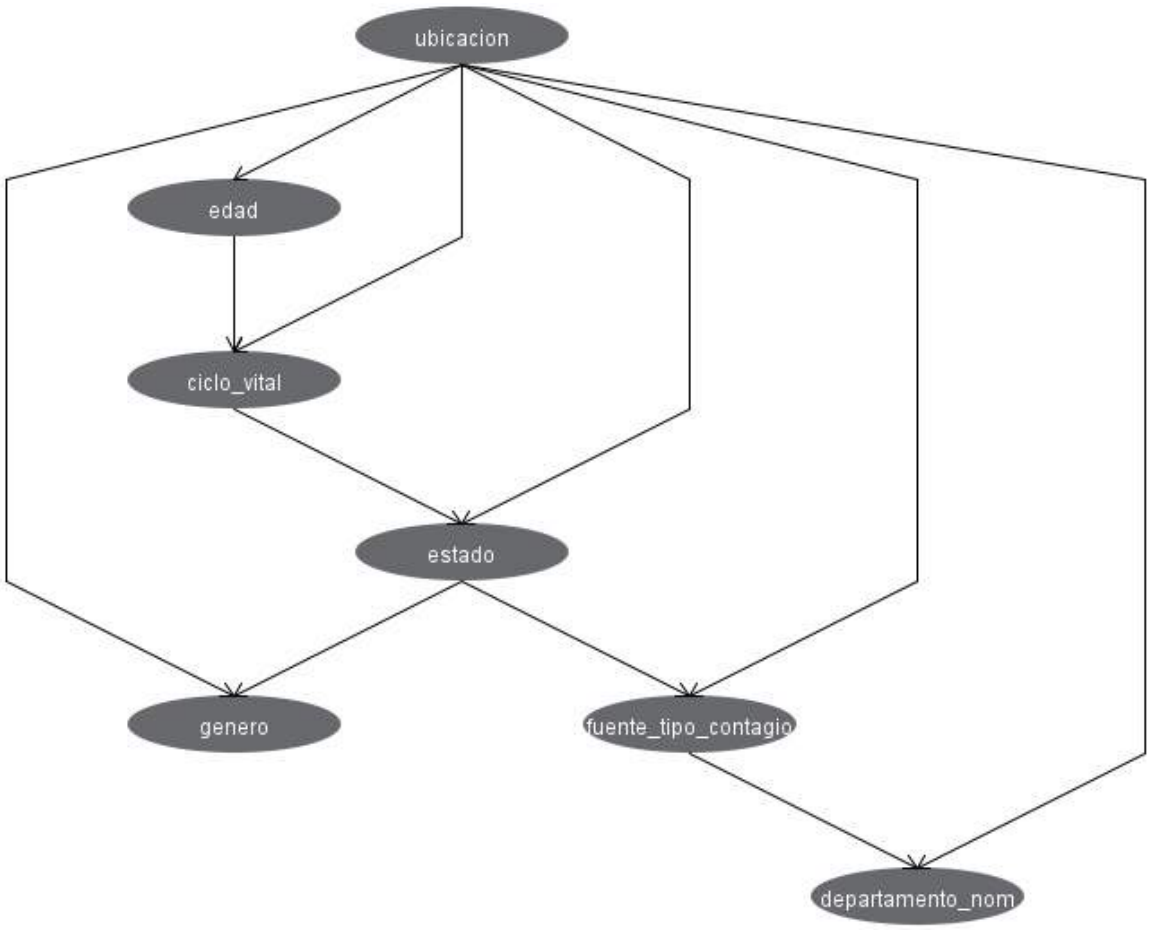


Figura 9: Modelo TAN correspondiente a la clasificación

Al observar las imágenes generadas por "Wekaçorrespondientes a la modelación de cada clasificador, donde cada grafo es una conexión divergente y la clase ubicación no es instanciada, por lo tanto las variables son d-separadas y el modelo de Naive Bayes solo tiene como padre a la variable clase, el modelo TAN creo arcos entre las variables características con mayor probabilidad de dependencia, un ejemplo claro de esto es que la edad es dependiente del ciclo vital, lo cual es evidente y se dejó en la modelación estas variables con el objetivo de constatar que el algoritmo efectivamente encuentra las variables con mayor dependencia; en el clasificador bayesiano 2DB el número máximo de padres que puede tener una variable es dos, donde la variable edad es la única que tiene más de dos padres distintos de la clase y note que la variable género es independiente de las demás variables, tomando el modelo esta variable como no significativa, aunque en el clasificador TAN la variable genero si tiene dependencia entre la variable clase (ubicación) y la variable estado, en los modelos TAN y 2DB se ve la relación entre la fuente tipo de contagio y el departamento, igualmente que entre ciclo vital y el

estado de un paciente.

7. Conclusiones

Los clasificadores bayesianos son una herramienta útil en el análisis de datos ya que no requieren una alta complejidad en la modelación, sin embargo es fundamental comprender la estructura matemática de cada clasificador para realizar un análisis adecuado de los datos, considerando que cada uno de los clasificadores son redes bayesianas, d-separadas y que la dependencia que se establece en cada modelo se basa en encontrar la máxima probabilidad entre cada una de las características, arrojando como resultado la red más optima.

Naive Bayes y TAN son capaces de describir la relación de dependencia entre atributos en una estructura en forma de árbol para cualquier conjunto de datos, diferente sucede con la aplicación del clasificador k-dependiente puesto que, si no selecciona el conjunto de variables considerables y de arcos que tienen relaciones de dependencia significativas, es decir el valor de k para construir el modelo, sino que seleccionamos aleatoriamente un subconjunto de relaciones de dependencia en el desarrollo del algoritmo, es probable que obtengamos un clasificador naive Bayes o TAN cuyo desempeño de clasificación no sea el mejor, esto se observa en la práctica, cuando se disminuyen el número de características a analizar de la base de datos de coronavirus a 4 variables característica y se toma un k=2, el modelo 2DB arroja un modelo naives Bayes (figura10) con un 72,7% de instancias correctamente clasificadas, disminuyendo su optimización el modelo.

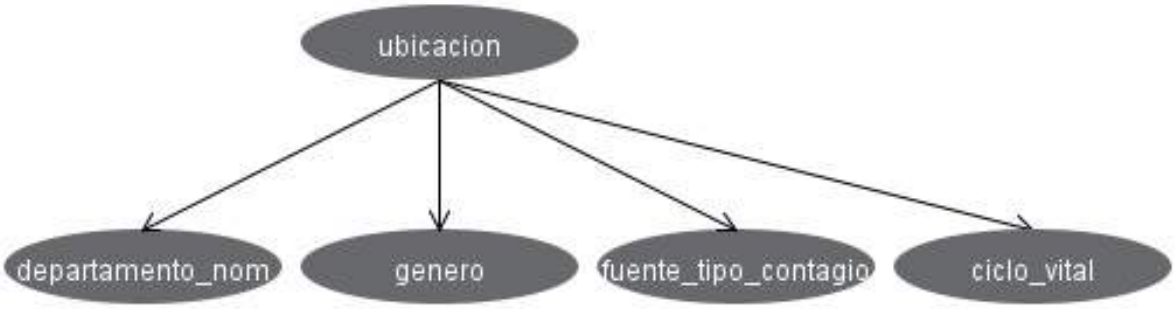


Figura 10: Modelo 2DB con 4 variables y k=2

Referencias

[1] Arcadio Rubio y José Antonio Gámez, *Flexible learning of k-dependence Bayesian network classifiers.*Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation, Jul 2011, Dublin, Ireland. pp.1219, ff10.1145/2001576.2001741ff.

[2] B. Chandra, Manish Gupta y M. P. Gupta, Thomas D. Nielsen*Robust Approach for Estimating Probabilities in Naive-Bayes Classifier* Indian Institute of Technology, New Delhi, India, 2015

[3] Chow, C. y Liu, C, *Approximating discrete probability distributions with dependence trees*, IEEE Trans. on Info. Theory, 14:462–467, 1968

[4] Finn V. Jensen y Thomas D. Nielsen, *Bayesian Networks and Decision Graphs*, segunda edicion, Editorial Springe, New York, USA, págs. 26-37, 2007

[5] Friedman, N, Geiger, D. y Goldszmidt, M.,*Bayesian network classifiers. Machine learning*, págs 131–163.

[6] Jose A. Gamez,Serafin Moral y Antonio SalmeronPhilippe Weber, *Advances in Bayesian Networks*, volumen 2, Great Britain and the United States by ISTE Ltd and John Wiley Sons, Inc, 2016

[7] Philippe Weber y Christophe Simon, *Benefits of Bayesian Network Models*, Editorial Springer, Verlag, Berlin Heidelberg, págs. 220-230, 2004

[8] Ren, H.J., Wang, X.C., Guo, Q.L. y Zhang, R., *Spatial prediction of oil and gas distribution using Tree Augmented Bayesian network, Computers and Geosciences* 2020,doi:https://doi.org/10.1016/j.cageo.2020.104518.

[9] Ren, H.J. y Wang, X.C. *Scalable Structure Learning of K-Dependence Bayesian Network Classifier.* IEEE Access,2020 doi:10.1109/access.2020.3035175

[10] Shang-Cai Ma, y Hong-Bo Shi. *Tree-augmented naive Bayes ensembles.* International Conference on Machine Learning and Cybernetics (IEEE Cat. No.04EX826) 2004. doi:10.1109/icmlc.2004.1382010