APLICACIÓN DE LAS REDES BAYESIANAS PARA EL ANÁLISIS DE SENSIBILIDAD DE LOS PRECIOS DE OFERTA EN BOLSA DE LOS GENERADORES EN EL MERCADO MAYORISTA DE ENERGÍA ELÉCTRICA EN COLOMBIA

JOSÉ LUIS SANTIAGO LOZANO
PEDRO JHONANDER VARELA NUNCIRA

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICO-MECÁNICAS
ESCUELA DE INGENIERÍAS ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES
BUCARAMANGA
2007

APLICACIÓN DE LAS REDES BAYESIANAS PARA EL ANÁLISIS DE SENSIBILIDAD DE LOS PRECIOS DE OFERTA EN BOLSA DE LOS GENERADORES EN EL MERCADO MAYORISTA DE ENERGÍA ELÉCTRICA EN COLOMBIA

JOSÉ LUIS SANTIAGO LOZANO
PEDRO JHONANDER VARELA NUNCIRA

Proyecto de Grado en modalidad de investigación presentado como requisito para optar al título de Ingeniero Electricista.

Director: PhD. Rubén Darío Cruz Rodríguez

Codirector: Ing. Javier Augusto Hernández Romero

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICO-MECÁNICAS
ESCUELA DE INGENIERÍAS ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES
BUCARAMANGA
2007

Dedico este libro a Dios que gracias a él he realizado todos mis proyectos, a mis padres David y Hercilia que siempre han estado a mi lado apoyandome en todo, a mis hermanos David y Yuri que los quiero mucho y a mi novia Jeniffer Natalia que me motiva todos los días y me inspira en todo momento.

Pedro Varela

Al buen Dios.

A mis padres, José y Elida, por su fe en mi, su continuo apoyo, sus plegarias y su amor incondicional.

A mi hermano Carlos Andrés, quien con su vida me ha mostrado su tenacidad, entrega y madurez.

A José Duván, el pequeñín de la casa.

A toda mi familia…

A todas las personas que he amado profundamente…

A todos los amigos que han recorrido conmigo este camino de la vida…

José Luis

AGRADECIMIENTOS

Agradecemos la realización de este proyecto a:

Ph D. Ruben Dario Druz – por su dirección en la tesis.

Ing. Javier Hernandez – por su valiosa colaboración durante todo el desarrollo de la tesis como codirector del proyecto.

Cesar Martinez, Silvia Zarate, Cesar Cuta – estudiantes UIS por su amistad y colaboración.

Y a todos los que de alguna u otra manera contribuyeron al desarrollo de esta tesis de grado.

CONTENIDO

LISTA DE FIGURAS

LISTA DE ECUACIONES

LISTA DE TABLAS

LISTA DE ANEXOS

RESUMEN

TÍTULO: APLICACIÓN DE LAS REDES BAYESIANAS PARA EL ANÁLISIS DE SENSIBILIDAD DE LOS PRECIOS DE OFERTA EN BOLSA DE LOS GENERADORES EN EL MERCADO MAYORISTA DE ENERGÍA ELÉCTRICA EN COLOMBIA[1]

AUTORES
JOSÉ LUIS SANTIAGO LOZANO
PEDRO JHONANDER VARELA NUNCIRA[**]

PALABRAS CLAVES
Análisis de sensiblidad, árboles de clasificación, curva de demanda residual, generación, métodos de clasificación, precios de oferta, redes bayesianas, variables predictoras.

DESCRIPCIÓN

Con el desarrollo de esta tesis de grado se busca definir un modelo general para el análisis de sensibilidad de los precios de oferta de un conjunto de generadores aplicando métodos de clasificación basados en redes Bayesianas y árboles de clasificación. El análisis de sensibilidad de los precios de oferta en bolsa para cada central generadora es analizado desde dos puntos de vista, uno enfocado en el modelo descriptivo y el otro en el predictivo.

En cuanto al modelo descriptivo, el análisis se enfoca en la identificación de las variables más representativas para la fijación del precio de oferta y en encontrar las relaciones entre las diferentes variables predictoras que componen el modelo. Desde el punto de vista predictivo, el modelo se centra en el cálculo de eficiencias para los diferentes métodos de clasificación con el fin de establecer si el conjunto de variables predictoras son tenidas en cuenta por el generador para la fijación de su precio de oferta.

Este análisis de sensibilidad se plantea en una serie de etapas consecutivas, iniciando con la adquisición de los datos, seguidamente las etapas de adecuación de los datos (actualización y discretización de los datos), selección de variables, corrimientos de variables y eliminación de atípicos para finalmente aplicar las técnicas de clasificación para encontrar modelos gráficos y los análisis de resultados de cada generador.

[1] Proyecto en la modalidad de tesis.
[**] Facultad de ingenierías Físico-Mecánicas, Escuela de Ingenierías Eléctrica, Electrónica y Telecomunicaciones, Director: Ph.D Rubén Darío Cruz Rodríguez.

ABSTRACT

TITLE: APPLICATION OF BAYESIAN NETWORKS FOR SENSITIVITY ANALYSIS OF THE OFFER PRICES OF THE GENERATORS IN THE COLOMBIAN WHOLESALE MARKET OF ELECTRIC ENERGY[*].

AUTHORS
JOSE LUIS SANTIAGO LOZANO
PEDRO JHONANDER VARELA NUNCIRA[**]

KEYWORDS
Sensitivity Analysis, classification trees, residual demand curve, generation, classification methods, offer prices, Bayesian networks, predictor variables.

DESCRIPTION

The making of this Thesis is focused in order to define a general model for the sensitivity analysis of the offer prices of a set of generators applying classification methods based on Bayesian networks and classification trees. The sensitivity analysis of the offer prices for every generator central is analyzed from two different points of view, one focused into the descriptive model and the other one into the predictable model.

As far as the descriptive model is concerned, the analysis is approached into the identification of the most representative variables for the fixation of the offer price and into finding out the relationship between the different predictor variables that compound the model. From the predictive point of view, the model consists on the calculus of efficiencies for the different classification methods in order to establish whether the set of predictor variables are taken into account by the generator for the fixation of its offer price.

This sensitivity analysis is proposed as a series of consecutive stages, beginning with the acquisition of data, followed by data adaptation (update and discretization of data), selection of variables, offset of variables and elimination of abnormalities for applying finally classification techniques to find out graphic models and analysis of results of every generator.

[*] Proyect of grade
[**] Faculty Physical-mechanical. Electric, Electronic school of Engineers and Telecommunications. Manager. PhD. Rubén Darío Cruz Rodríguez.

1. INTRODUCCIÓN

Los agentes pertenecientes al Mercado Mayorista de Energía Eléctrica comercializan la energía eléctrica como un bien o servicio más; y estos usando el modelo de libre competencia del mercado, desean maximizar individualmente sus beneficios económicos. Como parte de esta característica propia del mercado, los agentes generadores varían continuamente sus precios de oferta con tal de mejorar sus ingresos; estos cambios en los precios de oferta de los generadores, son un fenómeno común en el día a día y característico en la formación del precio en bolsa.

Como parte de la investigación en mercados eléctricos y en particular del comportamiento de oferta de los generadores, el Grupo de Investigación en Sistemas de Energía Eléctrica – GISEL – de la Universidad Industrial de Santander, junto con el Centro de Productividad y Competitividad del Oriente, -CPC- han trabajado en conjunto en el proyecto matriz denominado: *"Modelo de Análisis de Mercados de Energía Eléctrica mediante la Aplicación de una Metodología que involucra Inteligencia Competitiva y Agentes Inteligentes"*. En el marco de este proyecto matriz, el estudio del comportamiento de las ofertas de los generadores, es un elemento importante en la monitorización del Mercado Eléctrico Colombiano. Así, con este trabajo de grado, se quiere describir el comportamiento de los precios de oferta en bolsa de un conjunto de generadores del Mercado de Energía Mayorista Colombiano.

En los precios de oferta de los agentes generadores, se han identificado ciertos elementos estratégicos que inciden en la formación de los mismos las cuales se denominan como variables estratégicas[2] y precios de la curva de demanda

[2] Unidad de Planeación Minero Energética – UPME. "Una Visión del Mercado Eléctrico Colombiano". 2004. Parte del ejercicio del Plan de Expansión.

residual. Encontrar la importancia y las relaciones de cada una de estas variables en la formación de los precios de oferta de los generadores, servirían de referencia y guía a los organismos de regulación y control del mercado eléctrico en la búsqueda y vigilancia de comportamientos nocivos o posiciones dominantes de algunos de estos agentes generadores, ya que son los precios de oferta de los generadores los que conforman el precio en bolsa y la componente de generación una de las mas significativas en el precio final del usuario.

Cada una de las variables estratégicas, así como el precio de oferta de cada uno de los generadores y algunas características de la curva de demanda residual, se modelaron con Redes Bayesianas y árboles de clasificación, herramientas matemáticas-probabilísticas dotadas con técnicas de clasificación, a partir de las cuales se encuentran relaciones descriptivas y predictivas entre cada una de las variables presentes en este estudio, logrando llevar a cabo en cada uno de los generadores, un análisis de sensibilidad[3] de los precios de oferta.

La información guía para la construcción de los modelos clasificatorios para el estudio de la sensibilidad de los precios se basan en la recopilación de datos para las variables estratégicas disponibles en los años 2003, 2004, 2005 y el primer semestre del 2006. Estos datos tomados del sistema de información Neón[4], perteneciente a XM S.A. E.S.P [5] se usaron para observar el comportamiento de 12 generadores (6 hidráulicos y 6 térmicos).

---

[3] El análisis de sensibilidad consiste en estudiar cómo la variación en los resultados de un modelo (matemático y computacional) puede depender cualitativa o cuantitativamente, de las variaciones sufridas por distintas fuentes en la entrada.

[4] http://www5.isa.com.co/neonweb/ Sitio web del sistema de información Neón.

[5] XM – Compañía de Expertos en Mercados S.A. E.S.P - es la empresa del Grupo ISA que presta servicios integrales de operación, administración y desarrollo de mercados mayoristas eléctricos.

2. OBJETIVOS

2.1 OBJETIVO GENERAL

Desarrollar un modelo gráfico mediante clasificadores basados en Redes Bayesianas que permita describir el comportamiento de los precios de oferta en bolsa de un conjunto de generadores del Mercado de Energía Mayorista Colombiano.

2.2 OBJETIVOS ESPECÍFICOS

Tabla 1. Objetivos específicos y resultados obtenidos.

Objetivo Específico	**Resultado Obtenido**
• Identificar el conjunto de generadores que formarán parte del caso de estudio de este proyecto.	Se conformó un conjunto de 12 generadores (6 hidráulicos y 6 térmicos) a partir del número de coincidencias del precio de oferta con el precio en bolsa. Ver capítulo 5.

• Adecuar las variables identificadas como estratégicas en la conformación de los precios de oferta y los datos de la curva de demanda residual de manera que permitan ser utilizados en la construcción de una Red Bayesiana.	Se adecuaron los datos en términos de actualización, discretización, selección de variables y corrimiento de variables y eliminación de atípicos. Ver sección 7.3
• Diseñar un modelo descriptivo basado en algoritmos de aprendizaje o métodos de búsqueda de Redes Bayesianas para el estudio de la sensibilidad de las ofertas en bolsa de los generadores seleccionados.	Se seleccionaron las técnicas de clasificación a partir de Redes Bayesianas y árboles de clasificación, con los cuales se determinarán las inferencias y los modelos gráficos. Ver capítulo 6, secciones 7.4 y 7.5 y capítulo 8.
• Aplicar el modelo obtenido para analizar el comportamiento de las ofertas para el periodo 01/01/2006 a 30/06/2006.	Se construyeron los métodos de clasificación discretos y continuos para los generadores seleccionados, con cada método de discretización, destacando resultados por generador, así como el desempeño y las eficiencias de los clasificadores de manera general. Ver capítulo 9

3. OFERTAS DE LOS GENERADORES Y FORMACIÓN DEL PRECIO EN BOLSA.

Bajo el modelo de compra y venta de energía en el cual se desarrolla actualmente el mercado eléctrico colombiano, las ofertas de los generadores son elementos dinámicos y presentes día a día en la formación del precio en bolsa.

En este capítulo se quiere introducir al lector en los conceptos relacionados con generación, demanda, formación del precio en bolsa, así como otras variables asociadas con la dinámica del mercado eléctrico colombiano. Algunas de estos elementos asociados con la formación del precio en bolsa y presentes en el mercado eléctrico se establecieron como variables de estudio dentro de este trabajo de grado, con las cuales se quiere hacer un seguimiento al comportamiento de las ofertas de los agentes generadores en Colombia.

3.1 GENERACIÓN

La generación de energía eléctrica la componen todos aquellos lugares, equipos y componentes que producen la energía eléctrica a partir de otro tipo de energía. Dependiendo del tipo de energía usada para generar energía eléctrica, se destacan tres tipos de centrales de generación en Colombia, las cuales se detallan a continuación:

- Central Hidroeléctrica: Usan como medio de generación una caída de agua. Se pueden dividir básicamente en aquellas que tienen embalses, las filo de agua y las minicentrales.

- Central Térmica: Funcionan básicamente con carbón, gas, fuel oil, cogeneradores o combinaciones de las mismas, etc.

- Alternativas: Son aquellas que usan energías alternativas y cuya capacidad de generación aún no es suficientemente grande. Entre estas se destacan las Eólicas, Solares, Biogás, etc.

La generación de energía eléctrica en Colombia depende en gran parte, de las centrales hidroeléctricas; en donde aproximadamente el 64% de la capacidad instalada es de este tipo de generación[6]. Es así como el servicio de energía eléctrica en Colombia depende en buena parte del nivel de los embalses y de las condiciones meteorológicas e hidrológicas del territorio colombiano. Sin embargo, la legislación colombiana, ha creado herramientas para la inversión en sistemas de generación a través del Cargo por Confiabilidad, a través del cual a partir de una remuneración económica se garantiza la presencia y suministro de energía de estas plantas al sistema eléctrico a pesar de condiciones climáticas severas y bajo nivel de los embalses.

Cada generador hidráulico se caracteriza por tener un embalse propio, el cual representa una cantidad de energía acumulada y disponible para la generación de energía eléctrica. La sumatoria de la capacidad de reserva de energía de cada uno de los embalses del país se conoce como embalse agregado, el cual representa también la capacidad de almacenamiento de reservas energéticas en el sistema. A partir de la monitorización de esta variable y de los embalses individuales de cada uno de los generadores se toman medidas para garantizar el servicio de energía, inclusive en condiciones críticas hidrológicas.

[6] Dato a Febrero 15, 2007. Tomado de XM – S.A. E.S.P Características del Sector Eléctrico Colombiano. Gerencia Centro Nacional de Despacho. Seminario Introducción a la operación del SIN y a la Administración del Mercado.

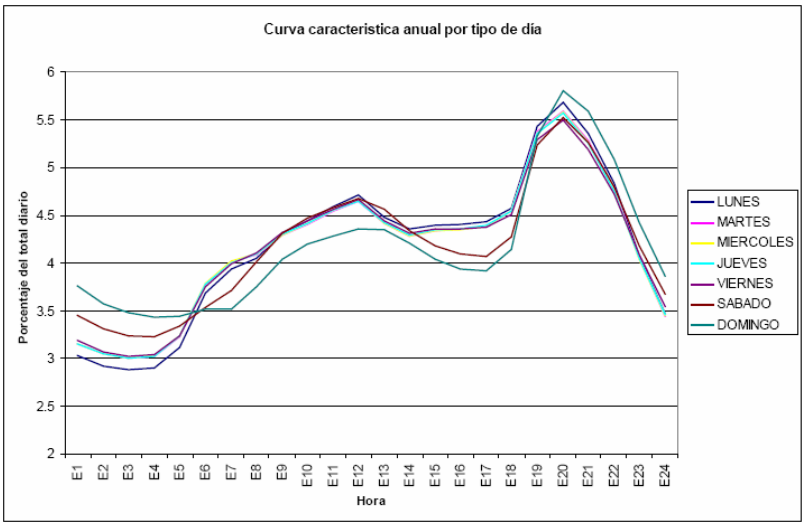
3.2 DEMANDA

Demanda es la cantidad de energía requerida por cargas o centros de demanda que continuamente exigen a un sistema eléctrico de potencia[7] el suministro de energía. La demanda de energía la componen el consumo de los usuario finales (residenciales, comerciales, industriales, etc), las pérdidas de energía, las exportaciones, el alumbrado público, el consumo propio, etc. La prestación adecuada de energía eléctrica, requiere de un equilibrio constante entre la demanda y la generación, ya que la generación ha de ajustarse continuamente de acuerdo a las condiciones que imponga la carga.

El organismo de operación y administración del mercado, XM – Compañía de Expertos en Mercados S.A. E.S.P – ha caracterizado por días y horas un comportamiento típico de la demanda de energía eléctrica en Colombia. Este comportamiento se puede ver en la figura 1.

[7] Un sistema eléctrico de potencia es el elemento básico para realizar la transformación y transmisión de energía a los centros de carga; su objetivo primordial es suministrar energía eléctrica a los diferentes usuarios en un área de servicio.

Figura 1. Curva característica de demanda por tipo de día[8]



En esta figura se puede destacar para cada uno de los días varios elementos característicos de cada una de las curvas. Algunos de estos son:

- Se presentan dos picos y dos valles en la demanda.

- El pico más significativo (el momento más alto de la demanda) se presenta alrededor de las horas 19, 20 y 21 (7 a 9:59 p.m.).

- El momento más bajo de la demanda se presenta en las horas de la madrugada, típicamente entre las horas 2, 3 y 4 (2 a 3:59 a.m.)

- Existen semejanzas entre los días normalmente hábiles (lunes a viernes), éstos se diferencian de los días sábados y Domingo.

[8] Tomado de Características del Sistema Eléctrico Colombiano. Gerencia Centro Nacional de Despacho. Seminario Introducción a la operación del SIN y a la administración del mercado.

La demanda se puede caracterizar para cada una de las 24 horas del día, en 24 franjas de energía o potencia correspondientes con cada una de las horas; con las cuales se aproxima el comportamiento de la demanda. Como parte de esta caracterización, las demandas horarias se pueden ordenar de manera ascendente o descendente para un día cualquiera, encontrando dentro de estas franjas algunos elementos claves para analizar el comportamiento de la demanda y del mercado de energía eléctrica. Entre estos se tienen:

- **Demanda máxima:** Franja de energía, la cual corresponde a la hora en la cual la carga consume más energía durante un día. Normalmente corresponde con la hora 20.

- **Demanda mediana** [9]: Franja de energía, la cual ocupa la posición central, colocando todos los valores de las demandas en orden creciente.

- **Demanda mínima:** Franja de energía, la cual corresponde con la hora en donde la carga consume menos energía durante un día.

La demanda de energía no solo cubre la demanda a los usuarios finales, sino también contempla otros elementos tales como las exportaciones (Transacciones Internacionales de Energía), consumo propio, el alumbrado público y las pérdidas. Todos estos elementos se cuantifican y se tienen en cuenta en la programación de la generación.

[9] La mediana es un número que supera a la mitad de los valores de la distribución y es superada por la otra mitad. Si el número de términos es impar, la mediana es el valor del elemento que ocupa el lugar central de la serie ordenada de elementos. Si el número de términos de la distribución es par, la mediana es el valor medio de los datos centrales.

3.3 BOLSA DE ENERGÍA, FORMACIÓN DEL PRECIO DE BOLSA Y CONTRATOS.

El suministro de energía eléctrica dejó de ser un asunto solamente técnico, para convertirse también en uno económico, en donde los distintos agentes pueden realizar acuerdos económicos para la entrega y consumo de energía eléctrica; así muchos de estos conceptos relacionados con figuras de contratación de energía, formación del precio en bolsa y bolsa de energía se han hecho cada vez más importantes en el conocimiento de la dinámica del mercado; estos son expuestos a continuación.

3.3.1 BOLSA DE ENERGÍA[10]

Es el sistema de información, manejado por el ASIC[11], sometido a las reglas del mercado mayorista de energía eléctrica, en donde los generadores y comercializadores del mercado mayorista ejecutan actos de intercambios de ofertas y demandas de energía, hora a hora. Luego el ASIC se encarga de liquidar y entregar los valores monetarios correspondientes a cada una de las partes.

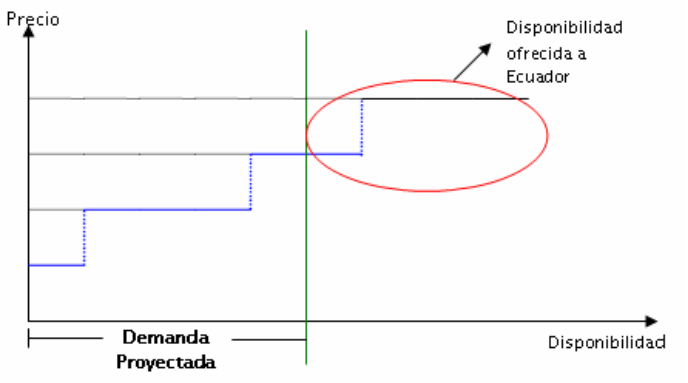
[10] Tomado del sitio web www.xm.com.co

[11] ASIC: Administrador del Sistema de Intercambios Comerciales. Dependencia del Centro Nacional de Despacho encargada de registro de fronteras comerciales, de los contratos de energía a largo plazo; de la liquidación, facturación, cobro y pago del valor de los actos, contratos, transacciones y en general de todos las transacciones que resulten del intercambio de energía en la bolsa entre generadores y comercializadores.

3.3.2 FORMACIÓN DEL PRECIO EN BOLSA

Antes de las 8 AM del día anterior al despacho, se realiza el predespacho ideal en el cuál los generadores envían sus ofertas para atender la demanda del siguiente día. Estas presentan dos componentes: *precio y disponibilidad;* las ofertas hechas por los generadores son planas y diarias y se ordenan según el precio de menor a mayor hasta cubrir la demanda proyectada[12] como se observa en la figura 2, la disponibilidad por encima de la demanda proyectada es ofrecida a Ecuador para las Transacciones Internacionales de Electricidad de corto plazo – TIE.

Figura 2. Predespacho ideal

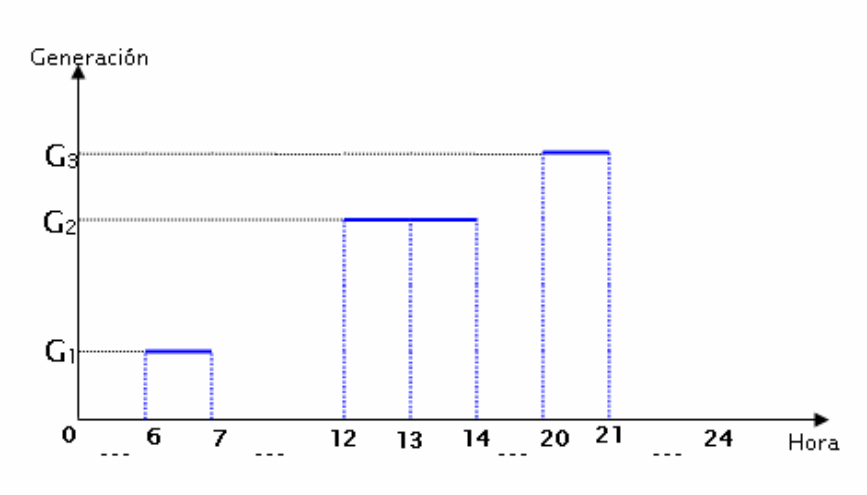


En el mismo día anterior al despacho, después de la elaboración del predespacho ideal, teniendo en cuenta las restricciones del sistema y las TIEs, se lleva a cabo el despacho programado, que es la programación horaria de la generación para cada uno de los agentes generadores. En la figura 3, se puede observar como ejemplo la programación hora a hora para un generador X. El generador X esta

[12] En la demanda proyectada se tienen en cuenta las perdidas del sistema.

programado para generar de 6-7 una cantidad G_1, de 12-14 una cantidad G_2 y de 20-21 una cantidad G_3.

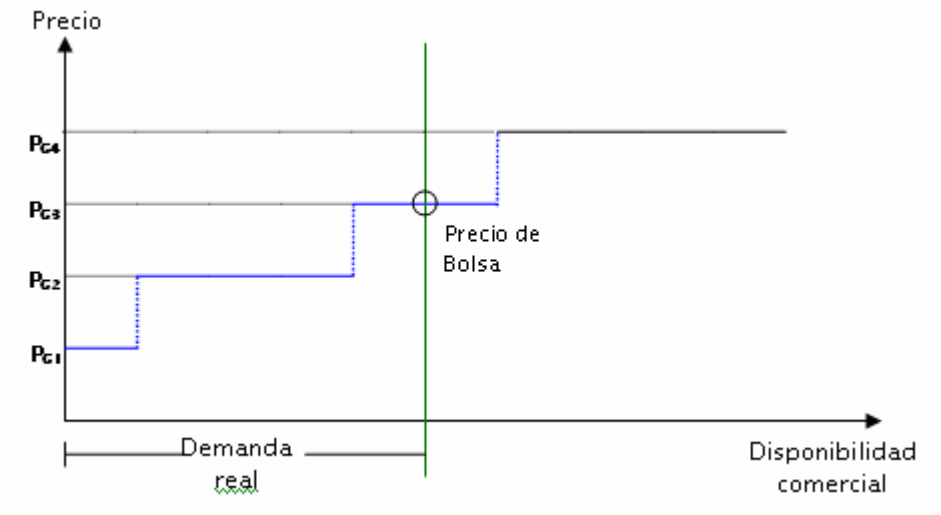
Figura 3. Despacho programado



Durante el día del despacho mediante una operación llamada redespacho, puede alterarse el despacho programado, es decir, se puede alterar la cantidad de energía a generar en una o más franjas horarias para cualquiera de los agentes generadores. Este cambio se debe notificar al agente generador mínimo 90 minutos antes de la hora de inicio de la respectiva franja horaria modificada.

El día siguiente al despacho se realiza el despacho ideal, el cual tiene en cuenta la demanda real y la disponibilidad comercial de las plantas de generación. En este procedimiento no se tienen en cuenta las restricciones del sistema y los precios de oferta de los generadores se ordenan de menor a mayor hasta completar la demanda real como se observa en la figura 4.

Figura 4. Formación del precio en bolsa para una franja horaria.



3.3.3 CONTRATOS

Son acuerdos comerciales entre dos partes para la compra y venta de energía entre Generadores y Comercializadores para atender parcial o totalmente los compromisos comerciales del comprador que participa en el Mercado de Energía Mayorista – MEM[13]. La forma, contenido, vigencia, mercado que atiende y condiciones se acuerdan libremente entre las partes. Entre algunos tipos o figuras de contratos están:

- **Pague lo contratado:** El agente se compromete a pagar toda la energía contratada, así esta sea consumida o no. En la resolución CREG 024-95 Anexo A-3 se dan las condiciones de pago cuando el consumo es mayor o menor de la energía contratada.

[13] Tomado de Contratos de Energía a Largo Plazo. GERENCIA MERCADO DE ENERGÍA MAYORISTA. TRANSACCIONES EN BOLSA. Junio de 2006.

- **Pague lo demandado:** El agente comprador paga solo (a precio de contrato) lo que consume. Otras condiciones de este tipo de contrato se detallan en la resolución CREG 024-95 Anexo A-3.

3.4 RESTRICCIONES DEL SISTEMA ELÉCTRICO Y RECONCILIACIONES.

El Sistema Interconectado Nacional posee límites operativos y físicos en la transferencia de potencia por las líneas de transmisión [14], interconexiones internacionales, etc. Cada límite se denomina restricción y con cada uno se asegura un suministro seguro y confiable del servicio de electricidad. Sin embargo, las restricciones encarecen el costo de la operación. Los costos asociados a generaciones obligatorias se ven reflejados en una figura aplicada a los generadores denominadas reconciliaciones, las cuales se explicarán mas adelante.

[14] Las líneas de transmisión son los elementos que transportan y distribuyen la energía eléctrica a niveles de tensión de 220 kV y 500 kV.

3.4.1 RESTRICCIONES.

Para garantizar tensiones adecuadas[15], estabilidad del sistema de potencia e intercambios seguros a nivel del Sistema de Transmisión Nacional y a nivel regional, se requiere una cantidad de generación adecuada para llevar a cabo estos propósitos, estas generaciones se denominan restricciones.

3.4.2 RECONCILIACIONES.

Reconciliación es la diferencia que se presenta entre la generación real y la generación ideal de una planta o recurso de generación. Este tipo de generación se produce para cubrir las restricciones del sistema.

De acuerdo al tipo de diferencia que exista, positiva o negativa, el agente vende reconciliación (recibe) ó compra reconciliación (paga). La reconciliación positiva se asocia con generaciones fuera de mérito, es decir aquellas ofertas que tienen que despachar independientemente de su precio de oferta, ya que las condiciones de operación técnicas así lo requieren o por generaciones de seguridad. Los costos de reconciliación negativa están asociados con generaciones desplazadas en el despacho real por generaciones de seguridad fuera de mérito.

Las reconciliaciones se pagan a un precio, el cual está debidamente reglamentado en resoluciones CREG (CREG 034- 2001, CREG 034-2001, CREG 084-2005, CREG 034-2001 y CREG 084-2005). Cuando la reconciliación es positiva, el

[15] En Colombia existen distintos niveles de tensión dependiendo de la cantidad de energía que se transporte por la línea de transmisión. Básicamente se distinguen tres tipos de redes dependiendo los niveles de tensión. Estos son el Sistema de Transmisión Nacional (STN) el cual lo componen el conjunto de líneas cuyos niveles de tensión son mayores o iguales a 220 kV, los sistemas de transmisión regionales (STR) cuyos niveles de tensión son menores a 220 kV y a su vez no pertenezcan a un sistema de distribución local y los sistemas de distribución local (SDL), encargados de distribuir la energía eléctrica a nivel municipal, distrital o local. Otra distinción en cuatro niveles de tensión en la distribución de la energía eléctrica se pueden encontrar en la resolución CREG 082 de 2002.

precio reconocido al generador por reconciliación es el mínimo entre el precio oferta y el precio de referencia (definido por resoluciones CREG); la reconciliación negativa, se cancela al promedio entre el precio de oferta y el precio de bolsa nacional. La formula para calcular los precios de reconciliación negativa son iguales para generadores térmicos e hidráulicos, mientras la formula para calcular los precios de referencia de la reconciliación positiva dependen de si el generador es térmico o hidráulico.

Durante el desarrollo de este capítulo se abordaron algunas de las variables más significativas en la formación del precio en bolsa y que a su vez pueden constituir las ofertas de los generadores. Entre las mencionadas dentro de este capítulo están el precio en bolsa, el embalse agregado y embalse propio, los contratos, las reconciliaciones y la disponibilidad de los generadores. En el siguiente capítulo se abordarán otras variables relacionadas con el estudio así como las mencionadas anteriormente, para finalmente caracterizar y conformar el conjunto de variables con el cual se conformará el estudio de este trabajo de grado.

4. VARIABLES ESTRATÉGICAS EN LA CONSTRUCCIÓN DEL PRECIO DE OFERTA DE LOS GENERADORES.

Algunas de las variables nombradas en el capítulo anterior así como otras tomadas de la curva de demanda residual que se revisan en este capítulo conforman el conjunto de variables denominadas como estratégicas y precios de la curva de demanda residual, respectivamente, que finalmente forman parte del estudio en este trabajo de grado.

Antes se introducirán las variables tomadas de la Curva de Demanda Residual, para luego mostrar el conjunto total de variables que forman parte del estudio de este trabajo de grado.

4.1 CURVAS DE DEMANDA RESIDUAL (CDR).

La curva de demanda residual se construye para una franja horaria y se realiza para un generador en particular. Para la construcción de esta curva es necesario conocer la disponibilidad y precio de las empresas generadoras, y también se requiere conocer la demanda en la franja horaria en la que se requiere el cálculo de la curva de demanda residual.

A continuación se describirá mediante un ejemplo como es este proceso, desde la construcción de la oferta agregada hasta conocer la curva de demanda residual de una empresa generadora.

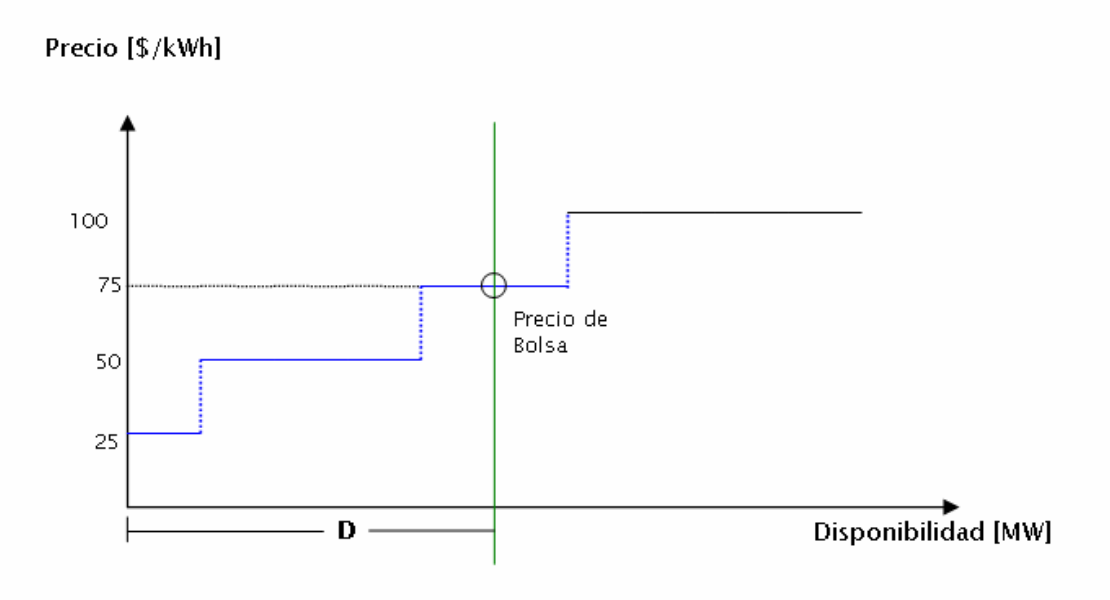
Por ejemplo, para una franja horaria se tiene la oferta de los generadores X, Y, Z y W, para atender una demanda D de 50 [MW] y se desea conocer la curva de demanda residual para el generador Z.

Tabla 2. Datos de los generadores para ejemplo de demanda residual

Generador	Precio [$/MWh]	Disponibilidad [MW]
X	25	10
Y	50	30
Z	75	20
W	100	40

Con base en la información de precio y disponibilidad enviada por las empresas generadoras, el operador del mercado ordenando el precio de los generadores de menor a mayor como se observa en la Tabla 2, construye la curva de oferta agregada S(p) que se observa en la figura 5.

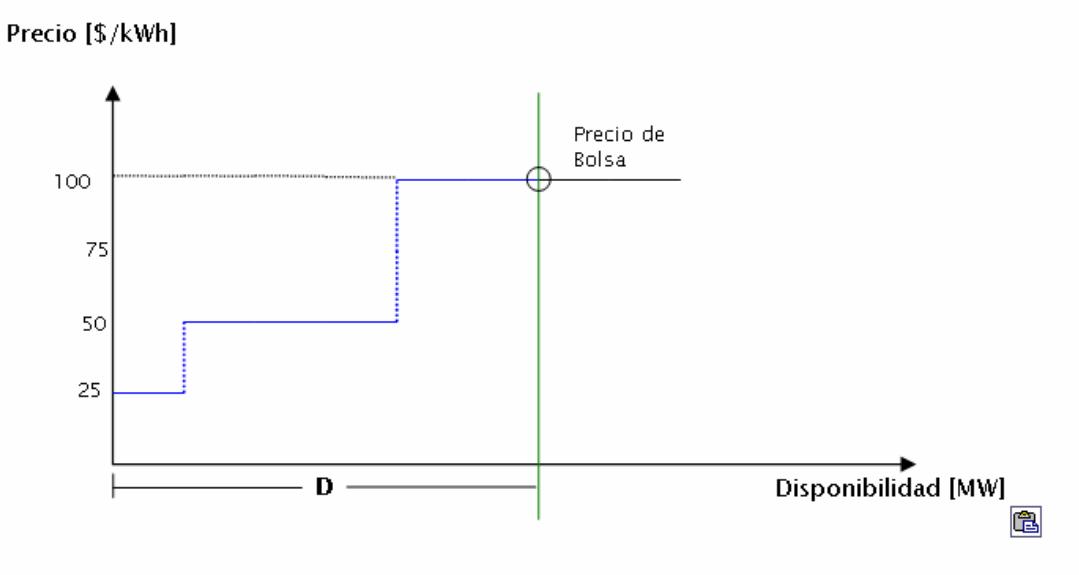
Figura 5. Curva de oferta agregada S(p)



De la figura 5, se observa que la demanda D se satisface con los generadores X, Y y Z; y el precio de bolsa es el mismo del generador Z, que fue el último generador empleado para completar la demanda total.

Para encontrar la curva de demanda residual del generador Z, se procede a construir la curva de oferta agregada para el resto de los generadores $S_{-1}(p)$, que es equivalente a restar de la curva de oferta agregada S(p) la oferta del generador Z.

Figura 6. Curva de oferta agregada $S_{-1}(p)$



En la figura 6 se puede observar la curva de oferta agregada para el resto de los generadores $S_{-1}(p)$. Para cumplir con la demanda total, la salida del generador Z obliga a la entrada del siguiente generador que en este caso es el generador W. También se puede notar de la figura 6, que el precio de bolsa ha aumentado al valor del generador W (100 $/MWh) y que para cumplir con la demanda total se emplea la disponibilidad de los generadores X e Y, además de una parte de la disponibilidad del generador W.

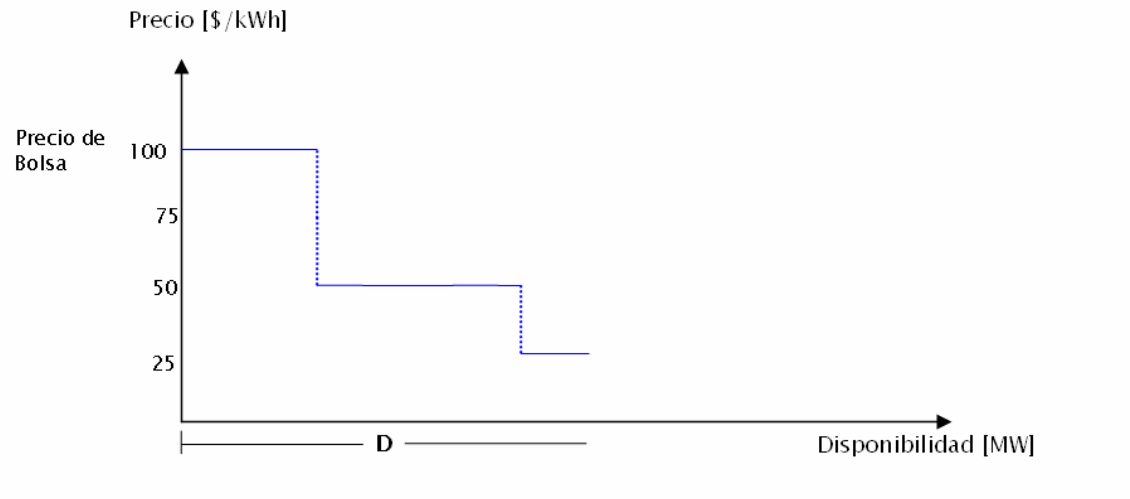
Finalmente, para obtener la curva de demanda residual para el generador Z, $R_Z(p)$, a la demanda total D se le resta la curva de oferta agregada de las demás firmas, como se define en la ecuación 1:

$$R_Z(p) = D - S_{-1}(p)$$

Ecuación 1. Definición de la demanda residual

Al realizar esta operación se obtiene la curva de la figura 7.

Figura 7. Curva de demanda residual para el generador Z, $R_Z(p)$



4.1.1 CONSTRUCCIÓN DE LA CURVA DE DEMANDA RESIDUAL

La construcción de la curva de demanda residual para cada generador se realiza teniendo en cuenta el proceso anteriormente descrito pero con las siguientes salvedades:

- El primer escalón de la curva de demanda residual es la suma de las inflexibilidades [16] del día anterior [17] de los demás generadores y la

[16] Mayor información acerca de declaración de inflexibilidades se puede consultar en la resolución CREG 0991668 de 1999

[17] Se usa la del día anterior debido a que en un escenario real los datos actuales serian desconocidos.

inflexibilidad del día actual[18] del generador al que se le esta calculando la curva de demanda residual.

- Seguidamente se empiezan a ubicar los generadores por orden ascendente de precio, teniendo en cuenta que la disponibilidad de cada generador es la disponibilidad del día anterior[19] menos la inflexibilidad del día anterior.

- La curva de demanda de residual se calcula para tres diferentes valores de demanda, la demanda máxima, la demanda mediana y la demanda mínima.

4.1.2 ELEMENTOS DE LA CURVA DE DEMANDA RESIDUAL QUE SE TOMAN COMO VARIABLES DEL PROYECTO

Como parte de la caracterización del comportamiento de las ofertas de los agentes generadores en el mercado mayorista de energía eléctrica, se han tomado ciertos elementos claves asociados con la demanda, así como con la oferta de los generadores en la construcción de la curva de demanda residual para cada uno de los días en estudio.

La construcción de una curva de demanda residual tiene asociada una franja horaria, de esta manera, en un día se pueden encontrar 24 franjas horarias de demanda asociadas con cada una de las horas del día. La construcción y análisis de cada una de las curvas de demanda residual, para el comportamiento de las ofertas sería un trabajo dispendioso. Es por esto, que en este trabajo de grado, se

[18] Se usa la del día actual debido a que es un escenario real, el generador que esta realizando el análisis de la curva de demanda residual conoce su inflexibilidad para ese día.

[19] Se usa la del día anterior debido a que en un escenario real los datos actuales serian desconocidos.

tomaron en cuenta tres demandas características para cada uno de los días de estudio, la demanda máxima, demanda mediana y demanda mínima[20].

Para cada una de las curvas de demanda residual caracterizadas por la demanda máxima, mediana y mínima se obtienen las variables P1D y P2D como se observan en la Figuras 8, Figura 9 y Figura 10. El precio P1D corresponde al mayor precio que se obtiene en la construcción de la curva de demanda residual de un generador X y el precio P2D es el mayor precio que puede ofrecer un generador X (al que se le esta calculando la curva de demanda residual) para garantizar que toda su disponibilidad será despachada.

Dependiendo del tipo de demanda, se le asocia una letra de subíndice a las variables P1D y P2D, obteniendo seis (6) precios característicos de la curva de demanda residual que se indican a continuación:

- Demanda máxima: Precio1Dx y Precio2Dx.
- Demanda mediana: Precio1Dd y Precio2Dd.
- Demanda mínima: Precio1Dn y Precio2Dn.

A continuación se muestran las figuras de los precios correspondientes con cada una de las demandas.

[20] Las definiciones de cada una de estas demandas se pueden detallar en el númeral 3.2

Figura 8. Demanda residual para demanda máxima

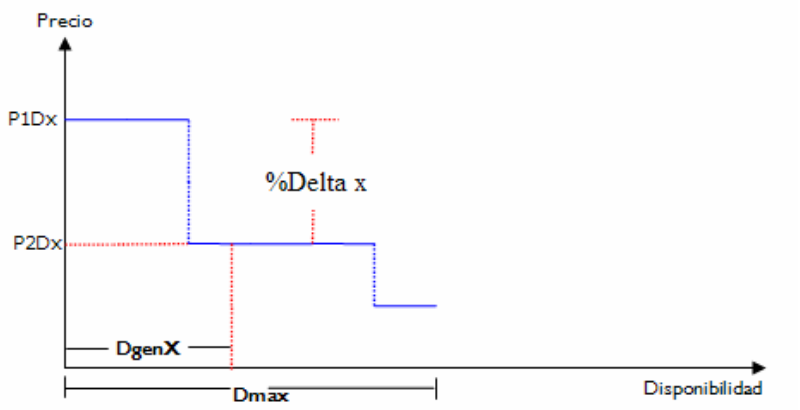


Figura 9. Demanda residual para demanda mediana

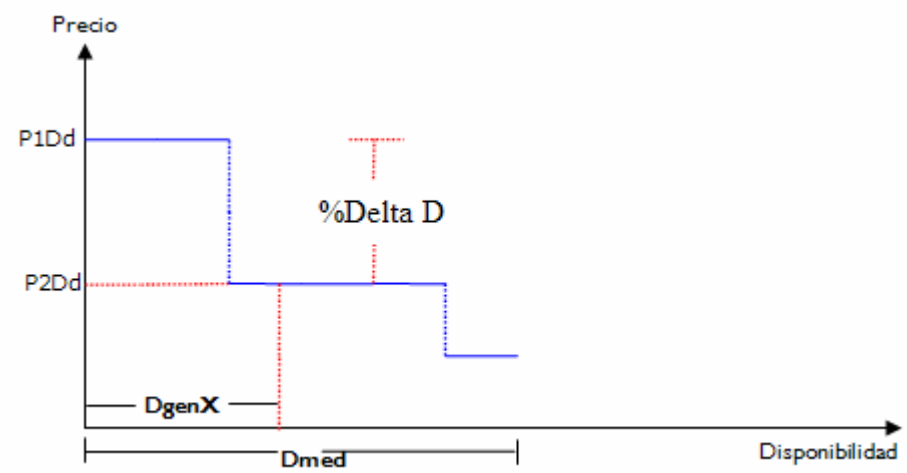
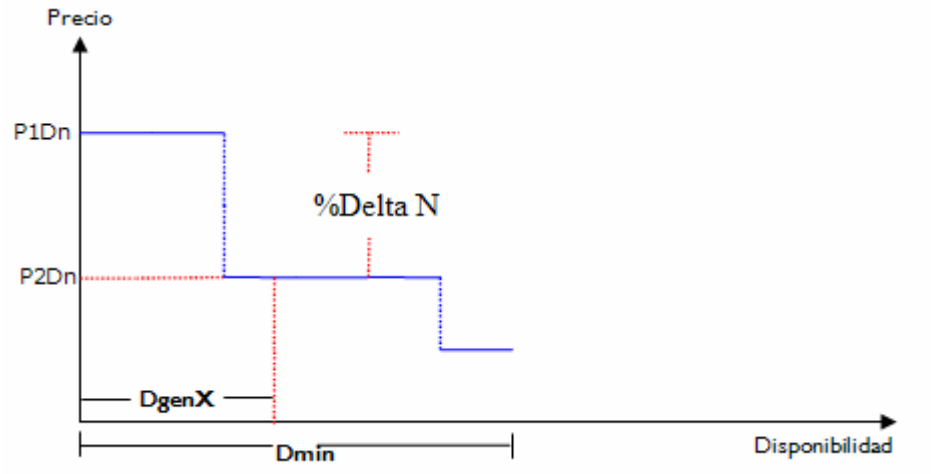


Figura 10. Demanda residual para demanda mínima



4.2 CONJUNTO DE VARIABLES ESTRATÉGICAS SELECCIONADAS EN EL ESTUDIO DE LAS OFERTAS DE LOS GENERADORES.

La elección de las variables estratégicas en este proyecto de grado se basa en la determinación de aquellas que más impacto han tenido en la construcción de los precios de oferta de los generadores. Como parte de esta selección de variables, se tomó en cuenta el documento "Una Visión del Mercado Eléctrico Colombiano. UPME" [UPME, 2004], el cual investiga y determina algunos elementos estratégicos en la formación de los precios de oferta; en este documento, se condensan algunos resultados de estudios al Mercado de Energía en Colombia. Entre estos elementos estratégicos se destacan:

- Precios de los contratos.
- Generación del despacho ideal.
- Precio de bolsa.
- Ventas en bolsa.
- Ingresos por reconciliaciones.
- Las inflexibilidades[21].
- Para las plantas térmicas, el embalse agregado del sistema.
- Para las plantas térmicas, la generación térmica del sistema.
- Para las plantas térmicas, los precios de oferta de otras plantas de la misma empresa.
- Los cambios de regulación.

Algunas de estas variables junto con otras son tomadas como estratégicas para cada uno de los generadores en su formación del precio de oferta. A continuación

[21] Las inflexibilidades fueron tratadas de un modo especial durante el desarrollo de este trabajo de grado. No se constituyó directamente dentro del conjunto de variables estratégicas, sin embargo, a la disponibilidad del generador se sustrajo la inflexibilidad declarada por cada uno de ellos. Este tratamiento se realizó durante los años 2003, 2004, 2005 y los seis primeros meses del 2006.

se muestran las variables tomadas en cuenta para la construcción de los modelos clasificatorios (para el estudio de la sensibilidad de los precios) y consideradas como estratégicas en este trabajo de grado, destacando de paso su importancia e impacto para cada una de ellas, sin dejar de tener en cuenta las recomendaciones hechas por estudios y análisis del Mercado Eléctrico Colombiano.

- Precio de oferta: esta variable es quizás la más importante en la construcción de la generación horaria por despacho de méritos para los generadores del mercado, así como en la competencia que existe entre cada uno de los generadores.

- Embalse agregado: la vigilancia de esta variable, puede generar especulaciones en los precios de oferta sobre todo en tiempos de sequía, tanto por generadores térmicos como por hidráulicos.

- Embalse propio**:** esta variable es importante para el generador propio en cuanto a la toma de decisiones en su precio de oferta diario y al control del embalse bajo condiciones hidrológicas críticas.

- Contratos en [$], en [MWh] y en [$/MWh]: con esta variable se quiere determinar el impacto de las figuras de contratación en la formación y decisión del precio en bolsa, ya que estos elementos cada vez, se han vuelto más significativos en el dinero que reciben los agentes generadores.

- Reconciliación Positiva y Negativa: con estas variables se quiere determinar la importancia de las restricciones del sistema en las ofertas de los generadores térmicos e hidráulicos para la formación del precio en bolsa.

- Disponibilidad: esta variable permite ver el impacto de la cantidad de energía disponible para cada uno de los días en la fijación del precio de oferta de cada uno de los generadores.

- Variables características de la curva de demanda residual: con la curva de demanda residual se quiere cuantificar el impacto de la presencia o no, de un agente ante los demás, y las respuestas de los demás agentes ante las decisiones adoptadas por el agente evaluado. Estas variables dan una idea de la posible influencia en la formación del precio en bolsa que puede ejercer un generador.

4.2.1 USO DE UNIDADES PARA LA CARACTERIZACIÓN DE LAS VARIABLES.

La importancia de la caracterización y cuantificación de las unidades de las variables para el establecimiento de las relaciones establecidas entre ellas y la variable clase (Precio), es un paso importante en la determinación del modelo. A continuación se mostrarán las unidades establecidas para las variables tomadas en el modelo:

- Precios: generalmente tienen asociada un valor por unidad de energía, es decir representan una unidad de dinero (peso, dólar, etc.) por unidad de energía, generalmente asociado con el consumo de energía eléctrica (kWh, MWh, etc.), de esta manera el precio se representa como el cociente entre estas dos variables ($/kWh, $/MWh, etc.). La unidad usada para representar los precios de oferta de los generadores y los precios de la curva de demanda residual fue de $/MWh (pesos por MegaWatt hora).

- Embalse agregado y embalse propio: El embalse representa una cantidad de energía potencialmente útil para la generación de energía, por lo tanto su cantidad se asocia con energía, casi siempre haciendo referencia a unidades de energía asociadas con la generación de energía eléctrica, tales como MWh (Megawatthora) o kWh (kilowatthora). La unidad asociada para la cuantificación de esta variable en este trabajo es MWh.

- Disponibilidad: Con esta variable se quiere mostrar la cantidad de energía que hora a hora el generador está en capacidad de entregar al sistema. Típicamente cada generador tiene asociada una capacidad máxima entregable al sistema, limitada obviamente por la cantidad de máquinas que trabaje y las condiciones operativas de las mismas. Esta unidad, generalmente se da en MW (MegaWatt); esta fue la unidad elegida para el tratamiento de esta variable durante el proyecto.

- Reconciliación positiva y reconciliación negativa: se pueden representar de distintas formas; en ingresos por reconciliaciones (unidades de dinero) o en cantidad de energía entregada de esta manera (MWh). Las unidades escogidas para la representación de esta variable fueron los MWh.

- Contratos: se representan tres variables diferentes, los contratos en unidades de dinero (pesos), en cantidades de energía (MWh) y el cociente entre estas dos variables ($/MWh). Las variables y las unidades elegidas son las que se describen anteriormente.

5. SELECCIÓN DE LOS GENERADORES

Del total de los generadores que aportan energía al Sistema Interconectado Nacional, las plantas centralizadas que participan dentro del despacho entre hidráulicas y térmicas representan el 96.2% total de la generación mientras que las plantas menores y cogeneradores representan el 3.8% restante[22]. Involucrar todos los generadores dentro de este estudio, generaría un trabajo muy dispendioso que incluye una cantidad de tiempo que esta fuera del alcance de esta tesis de grado. Es así, como se escogió un conjunto de doce agentes generadores, seis hidráulicos y seis térmicos, a partir de los cuales se estudiará el comportamiento de sus precios de oferta.

Para la selección de este conjunto de generadores se define el criterio que permite realizar dicha tarea. Este es conocido como 'número de coincidencias del precio de oferta con el precio de bolsa' y se muestra su aplicación a los generadores del mercado eléctrico colombiano.

5.1 NÚMERO DE COINCIDENCIAS DEL PRECIO DE OFERTA CON EL PRECIO DE BOLSA.

La selección del conjunto de generadores para el desarrollo de este proyecto se basa en el análisis de uno de los indicadores del comportamiento del MEM, suministrado por la Superintendencia de Servicios Públicos Domiciliarios, conocido como, número de coincidencias del precio de oferta con el precio de bolsa. Este indicador establece el número de coincidencias de los precios de oferta de un

[22] Tomado de www.xm.com.co a Diciembre 31 de 2006.

agente con respecto al precio de bolsa. El análisis se hace horario para todos los agentes. Para un agente X significa que en una hora determinada el precio de bolsa fue el mismo que ofertó el agente por alguna de sus plantas, y se calcula como el un número de coincidencias del agente sobre el número de coincidencias en el mes para esa hora[23].

La importancia de este indicador radica en conocer con que frecuencia los precios de oferta de los generadores coinciden con el precio de bolsa, así mismo asegurar como este grupo de generadores son líderes del mercado en la formación del precio de bolsa. Estos datos están disponibles en: http://www.superservicios.gov.co/energiagas/energia_ind_comp_mem.htm y son suministrados por la Superintendencia de Servicios Públicos Domiciliarios.

5.2 APLICACIÓN DEL INDICADOR "NÚMERO DE COINCIDENCIAS DEL PRECIO DE OFERTA CON EL PRECIO DE BOLSA" PARA LA SELECCIÓN DE LOS GENERADORES.

El análisis de este indicador se realizó con datos de Julio de 2004 a Enero de 2006. Una vez analizado este indicador en el periodo de tiempo establecido se obtuvieron los resultados que se observan en la Tabla 3.

Como se puede observar de esta Tabla 3, la columna 1 indica la empresa generadora, la columna 2 indica la planta generadora en particular de la empresa (en azul se encuentran las plantas hidroeléctricas y en rojo las plantas térmicas), la columna 3 indica el número de coincidencias de la planta generadora, la columna 4 indica el porcentaje de participación en horas del agente generador,

[23] Ver http://www.superservicios.gov.co/energiagas/energia_ind_comp_mem.htm

tomando como referencia que durante el periodo de estudio hubo 13920 franjas horarias y el cálculo de esta participación se da como el cociente entre el número de coincidencias y el número de franjas horarias establecidas anteriormente. Finalmente en la columna 5 se indica la capacidad de la planta generadora en Megavatios.

Tabla 3. Coincidencias del precio de oferta con el precio de bolsa de los generadores

Empresa	**Planta Generadora**	**Coincidencias**	**Porcentaje de participación**	**Capacidad (MW)**
ISAGEN	San Carlos	4362	31.34	1240
EMGESA	Guavio	3036	21.81	1150
CHIVOR	Chivor	1493	10.73	1000
EEPPM	Guatrón	1247	8.96	512
EEPPM	Porce II	854	6.14	405
EMGESA	Paraíso-Guaca	709	5.09	276
EPSA	Alto y bajo anchicayá	703	5.05	439
CEN. HID. BET	Betania	644	4.63	540
EEPPM	La tasajera	635	4.56	306
CORELCA	Tebsa-total	313	2.25	890
EEPPM	Guatapé	308	2.21	560
EEPPM	Playas	256	1.84	204
EPSA	Calima	247	1.77	132
EGETSA	Prado 4	195	1.4	
EGETSA	Prado	195	1.4	

GEST. ENERG.	Paipa 4	168	1.21	150
EEPPM	Riogrande I	145	1.04	
ISAGEN	Jaguas	136	0.98	170
EEPPM	Miel I	127	0.91	396
TERMOTASAJERO	Tasajero I	103	0.74	163
CHEC	Termodorada	93	0.67	52
CORELCA	Termoflores	74	0.53	150
URRA	Urrá	58	0.42	340
TERMOFLORES	Termoflores 3	29	0.21	
GEST. ENERG.	Paipa 2	28	0.2	68
TERMOFLORES	Termoflores 2	24	0.17	
TERMOYOPAL	Termoyopal 2	23	0.17	30
GEST. ENERG.	Paipa 3	23	0.17	68
EMGESA	Zipa ISA 4	14	0.1	
EEPPM	La vuelta (planta menor)	13	0.09	19,8
GEST. ENERG.	Paipa 1	13	0.09	28
EPSA	Salvajina	11	0.08	285
EMGESA	Zipa ISA 5	8	0.06	
CORELCA	Termoguajira 1	4	0.03	151
ISAGEN	Termocentro 1	4	0.03	300
CORELCA	Termobarranquilla 4	3	0.02	
ISAGEN	Venezuela	3	0.02	
PROELÉCTRICA	Proeléctrica 1	2	0.01	
CORELCA	Termoguajira 2	1	0.01	151

Para el desarrollo de este trabajo de grado se escogen 6 generadores hidráulicos y 6 generadores térmicos. Con base en el mayor número de coincidencias del precio de oferta con el precio en bolsa de los generadores (ver Tabla 3) se determina que estos generadores deben ser:

- Hidráulicos: San Carlos, Guavio, Chivor, Guatrón, Porce II y Paraíso-Guaca.

- Térmicos: Tebsa, Paipa 4, Tasajero I, Termodorada, Termoflores y Termoflores 3.

6. TÉCNICAS DE CLASIFICACIÓN USADAS EN EL ANÁLISIS DE SENSIBILIDAD DE LOS PRECIOS DE OFERTA EN BOLSA.

Las técnicas de clasificación son un tema de vital importancia para el desarrollo de este trabajo pues son estas herramientas, las que arrojan información acerca de la sensibilidad de los precios de oferta. En este capitulo se detallan las dos técnicas de clasificación empleadas que son las Redes Bayesianas y los Árboles de clasificación. Para la técnica de Redes Bayesianas se profundiza en los métodos de clasificación Naives Bayes y TAN.

Para entender de una mejor manera los conceptos acerca de Redes Bayesianas y Árboles de clasificación, este capitulo inicia presentando los principales conceptos relacionados con probabilidad y métodos de clasificación supervisada para posteriormente profundizar en los modelos de clasificación Naives Bayes, TAN y Árboles de clasificación.

6.1 REDES BAYESIANAS

6.1.1 TÉRMINOS RELACIONADOS CON MÉTODOS DE CLASIFICACIÓN.

La teoría de la probabilidad usa el Teorema de Bayes[24] para realizar inferencias, este permite mejorar la certeza que se tiene de un suceso o conjunto de sucesos a la luz de nuevos datos u observaciones. Básicamente este teorema, permite pasar

[24] Thomas Bayes, Matemático británico y ministro presbiteriano (1702-1761), conocido por haber formulado el Teorema de Bayes, el cuál fue publicado póstumamente en el trabajo titulado "*An Essay towards solving a Problem in the Doctrine of Chances"*. El teorema que lleva su nombre se refiere a la probabilidad de un suceso condicionado por la ocurrencia de otro suceso.

de la probabilidad *a priori* P(suceso), a la probabilidad *a posteriori* P(suceso/observaciones).

Para la comprensión de la terminología relacionada con Redes Bayesianas, Teorema de Bayes y Árboles de clasificación, es necesario conocer previamente ciertos conceptos relacionados con estos temas. A continuación se presentan algunos de estos conceptos:

Probabilidad a priori: También conocida como probabilidad clásica. Se obtiene del conocimiento de un evento, de una experimentación (un suceso) o de un razonamiento lógico, en el cual todos los eventos posibles han de ser igual de probables. Su formulación matemática es sencilla.

P(A)=número de resultados en que se presenta el evento/Número total de resultados

Ecuación 2. Probabilidad del evento A

Probabilidad a posteriori bajo condiciones de dependencia estadística[25]:. Probabilidad de que se presente un evento, dado que otro evento ya se ha presentado. Su notación matemática es:

$$P(B/A) = \frac{P(BA)}{P(A)}$$

Ecuación 3. Probabilidad condicional de que se de el evento B dado que se halla dado el A.

[25] La dependencia estadística existe cuando la probabilidad de que se presente algún suceso depende o se ve afectada por la presentación de otro evento

Probabilidad conjunta bajo condiciones de dependencia estadística: Probabilidad de que se presenten dos o más eventos simultáneamente o en sucesión. Su notación matemática es:

$$P(BA) = P(B/A) \cdot P(A)$$

Ecuación 4. Probabilidad conjunta para que los eventos A y B se den al mismo tiempo.

Probabilidad marginal: Probabilidad de que se presente un solo evento.

Nodo: Representación gráfica de una variable en un modelo de clasificación.

Arco: Línea o flecha que une dos nodos y que representa una dependencia probabilística entre estas variables.

Grafo o red: Conjunto finito de nodos (variables) y arcos. Los arcos de un grafo pueden ser dirigidos o no dirigidos, dependiendo de si se considera o no el orden de los nodos.

Arco dirigido: Es aquel cuya dirección está especificada entre dos nodos. Representa una dependencia probabilística entre estas variables en el sentido establecido.

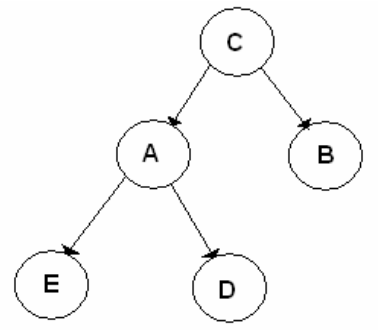
Grafo dirigido: Un grafo cuyos arcos son dirigidos. Para este es importante el orden del par de nodos que definen cada arco.

Padre: Nodo a partir del cual parte un arco dirigido. Dependiendo del modelo clasificatorio puede tener uno o varios hijos. (Ver figura 11).

Hijo: Nodo al que llega el sentido de la flecha de un arco dirigido. Dependiendo del modelo clasificatorio puede tener uno o varios padres. (Ver figura 11).

Nodo raíz: Nodo a partir del cual se establecen las primeras relaciones de dependencia y arcos dirigidos de un modelo bayesiano. Dicho nodo no posee padre y a partir de este se genera la primera descendencia (hijos). (Ver figura 11).

Figura 11. Grafo o red. La variable C es el nodo raíz. La variable A es padre de E y D. Mientras la variable C es padre de A y B.



Camino cerrado: Es aquel en el cual el nodo final e inicial son los mismos.

Ciclo: Es un camino cerrado en un ciclo dirigido.

Variable clase: también conocida como variable clasificadora. Variable que se le da un tratamiento especial y de la cuál se desea obtener predicciones de su comportamiento. A partir de esta variable surgen las relaciones por arcos y las dependencias probabilísticas con las demás variables en estudio. Para este trabajo de grado la variable clase es el precio de oferta de los generadores.

Variable discreta: Variable a la cual se le ha dado un tratamiento a partir un método de discretización. Cada elemento de una variable discreta tiene asociado una etiqueta o un estado en particular.
Variable continua: Variable que refleja el comportamiento de un determinado elemento con los datos originales de este.

Nodo Terminal: Es aquel nodo que no tiene hijos.

Árbol: Caso particular de poli árbol en el cuál cada nodo tiene un solo padre a excepción del nodo raíz, que no tiene padres.

Datos para Aprendizaje: Conjunto de datos que reflejan el comportamiento de las variables estratégicas y con los cuales se construyen y entrenan los métodos clasificatorios. Para este trabajo estos datos son los correspondientes a los años 2003, 2004 y 2005.

Datos para Inferencia: Conjunto de datos que reflejan el comportamiento de las variables estratégicas y con los cuales se evalúan los métodos clasificatorios. Para este trabajo estos datos son los correspondientes al primer semestre del 2006.

Clasificación supervisada: Consiste en clasificar nuevos objetos basándose en la información de una muestra ya clasificada. En este tipo de clasificación se da un tratamiento especial a la variable clase y las demás variables son conocidas como variables predictoras.

Información Mutua: Método que calcula la relación que existe entre una variable y otra dada. De esta manera se mide que tanta relación, información o conocimiento puede tener una variable con respecto al estado de la segunda. Esta medida se encuentra acotada en el intervalo [0,1]; valores cercanos a 1 dan una

alta correlación, mientras que valores cercanos a 0 indican independencia entre las variables analizadas.

La expresión matemática de este algoritmo se expresa a continuación, en donde: X es la variable a evaluar, C es la variable clase, r_x es el número de estados que X puede tomar, y r_c es el número de estados que la variable clase puede tomar.

$$IM(X;C) = \sum_{i=1}^{r_x} \sum_{j=1}^{r_c} P(x_i, c_j) \log\left(\frac{P(x_i, c_j)}{P(x_i)P(c_j)} \right)$$

Ecuación 5. Información mutua entre la variable clase (C) y la variable a evaluar (X).

$P(x_i)$ y $P(c_j)$ son probabilidades marginales y se calculan mediante la suma de las probabilidades de todos los eventos conjuntos en los que se presenta el evento sencillo, esto se realiza por presentarse en condiciones de dependencia estadística. $P(x_i,c_j)$ es la probabilidad conjunta, es decir, la probabilidad de que los dos eventos x_i y c_j se presenten juntos. Se calcula de la siguiente manera:

$$P(x_i, c_j) = \frac{N(x_{i,}, c_j)}{N}$$

Ecuación 6. Probabilidad conjunta para los eventos $x_{i,}$ c_j

En donde $N(x_i,c_j)$ es el número de observaciones totales en las cuales los eventos x_i y c_j se den a la vez y N es el conjunto de observaciones totales.

A continuación se ilustra un ejemplo de este método, donde se tienen tres variables definidas así:

- La variable var03 es la variable clase, puede tomar los estados [0 1 2 3 4].
- La variable var02 puede tomar los estados [s0 s1 s2 s3 s4].
- La variable var01 puede tomar los estados [low médium high].

Se requiere encontrar el ranking de variables, para las variables var01 y var02 dado que la var03 es la variable clase.
La base de datos para trabajar este ejercicio es la Tabla 4

Tabla 4. Estados que pueden tomar las variables var02 y var03

var01	**var02**	**var03**
High	S0	2
High	S2	2
Médium	S2	2
Low	S2	0

Se comienza con la variable var01: como se puede observar de la Tabla 5 el valor de la probabilidad conjunta (color amarillo) se encuentra observando la frecuencia de los datos en la Tabla 5, por ejemplo P(médium,2)=2/4=0.5.

Tabla 5. Probabilidades conjuntas y marginales de los estados de las variables var01 y var02

ESTADOS	**0**	**1**	**2**	**3**	**4**	
high	0	0	**0,5**	0	0	0,5
medium	0	0	**0,25**	0	0	0,25
low	**0,25**	0	0	0	0	0,25
	0,25	0	0,75	0	0	

Una vez se tienen todas las probabilidades conjuntas, se suman las filas y las columnas para obtener las probabilidades marginales (color verde). Con las probabilidades conjuntas y marginales se aplica la fórmula de la ecuación 5, que para este caso seria:

$$I_p = 0.5 * \log\left(\frac{0.5}{0.75 * 0.5}\right) + 0.25 * \log\left(\frac{0.25}{0.75 * 0.25}\right) + 0.25 * \log\left(\frac{0.25}{0.25 * 0.25}\right)$$

De la anterior expresión se obtiene que Ip es igual a 0.24421905 para var01.

Para la variable var02 se realiza el mismo procedimiento, en la Tabla 6 se observan las probabilidades conjuntas y marginales. Repitiendo el proceso anteriormente descrito para var01 se obtiene que Ip es igual a 0.036893107 para la var02.

Tabla 6. Probabilidades conjuntas y marginales de los estados de las variables var02 y var03.

ESTADOS	**0**	**1**	**2**	**3**	**4**	
S0	0	0	**0,25**	0	0	0,25
S1	0	0	0	0	0	0
S2	**0,25**	0	**0,5**	0	0	0,75
S3	0	0	0	0	0	0
S4	0	0	0	0	0	0
	0,25	0	0,75	0	0	

El resultado del ranking seria:

Var01: 0.24421905

Var02: 0.036893107

Los resultados usando la aplicación Elvira[26] se muestran en la figura 12 para el mismo ejercicio.

[26] Aplicación para Redes Bayesianas.

Figura 12. Resultados del software Elvira usando Información Mutua para tres variables, incluyendo la variable clase.

Tomado del software Elvira

Al comparar los resultados se observa que son los mismos. De estos resultados se puede concluir que la correlación entre la variable clase y la var02 es muy baja.

Este algoritmo no solo es útil en el uso de clasificadores bayesianos, sino también en el ranking de variables y en algunas modificaciones al algoritmo, llevándolo a expresar ahora la *Información Mutua Condicionada.*

Teorema De Bayes: Básicamente el teorema de Bayes es el resultado de probabilidades condicionales. Sean A y B dos eventos con P(A). Así se tiene:

$$P(A/X) = \frac{P(X/A) \cdot P(A)}{P(X)} = \frac{P(A,X)}{P(X)}$$

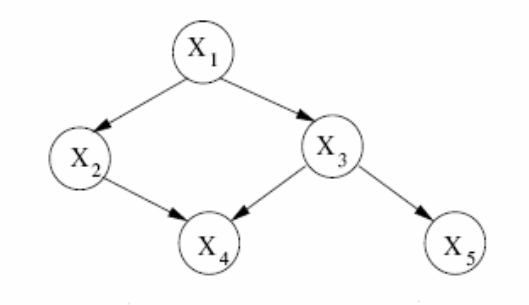
Ecuación 7. Teorema de Bayes

El uso principal de este teorema en probabilidad, es conocer el estado de unos eventos bajo ciertas condiciones y como a partir de un conocimiento básico y certero, se puede inferir o conocer la probabilidad de un cierto estado.

6.1.2 DEFINICIÓN DE RED BAYESIANA

Una Red Bayesiana es un modelo gráfico acíclico direccional que muestra las relaciones de dependencia e independencia entre variables, descrito por una distribución de probabilidad, en el cual cada nodo es una variable y cada arco una dependencia probabilística; este modelo usa el teorema de Bayes para el cálculo de dichas dependencias.

Figura 13. Estructura de una Red bayesiana



Tomado de [ARMAÑANZAS, 2004].

Las Redes Bayesianas permiten un doble uso: descriptivo y predictivo. En la parte descriptiva, los algoritmos de aprendizaje de Redes Bayesianas buscan descubrir las relaciones de independencia y/o relevancia entre sus variables, revelando

muchas relaciones de interés. Estas relaciones van desde una dependencia completa hasta una dependencia funcional entre las variables. En su uso predictivo, las redes bayesianas se usan como clasificadores. Así mismo, las Redes Bayesianas, expresan de forma numérica la “fuerza” de las relaciones entre las variables; esta parte cuantitativa del modelo se especifica mediante distribuciones de probabilidad [Hernández, Ramírez, 2004]. Algunas de las características más importantes de una Red Bayesiana son:

- Permiten aprender a partir de relaciones de dependencia y causalidad.
- Combinan conocimiento con datos.
- Evitan el sobre-ajuste de los datos y pueden manejar bases de datos incompletas.

La construcción de una Red Bayesiana a partir de datos, se divide en dos etapas: el aprendizaje estructural y el aprendizaje paramétrico. El aprendizaje estructural consiste en la obtención de las relaciones de dependencia e independencia entre las variables relacionadas; con el aprendizaje paramétrico se obtienen las probabilidades a priori y condicionales necesarias para la construcción del modelo dado. Los métodos Bayesianos son útiles para:

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

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

6.1.3 APRENDIZAJE DE LAS REDES BAYESIANAS.

El problema del aprendizaje de Redes Bayesianas consiste básicamente en dado un conjunto de datos, hallar el grafo acíclico dirigido que represente de un mejor modo las relaciones de dependencia e independencia presentes en los datos. La construcción de este modelo puede ser atacado de dos maneras diferentes; una, mediante la ayuda de expertos en este campo, con un considerable costo de tiempo y posibles errores en las aproximaciones, o dos, mediante el aprendizaje automático tanto de la estructura a partir de una base de datos.

El problema de clasificación con Redes Bayesianas, se puede abarcar de distintas maneras y técnicas existentes. Lo propuesto en este trabajo de grado es integrar un método de aprendizaje híbrido que combine técnicas de los árboles de decisión con redes bayesianas.

Una vez construida la Red Bayesiana, esta se convierte en un buen medio para realizar tareas de inferencia probabilística. Dada una evidencia sobre el valor de algunas variables de la red, se pueden calcular la distribución de probabilidad de otras variables de interés.

6.1.4 CLASIFICADOR NAIVES BAYES.

Este método de clasificación supervisado es el más sencillo dentro de los clasificadores basados en Redes Bayesianas y se ha demostrado que sus resultados pueden llegar a ser tan competitivos con técnicas como Redes

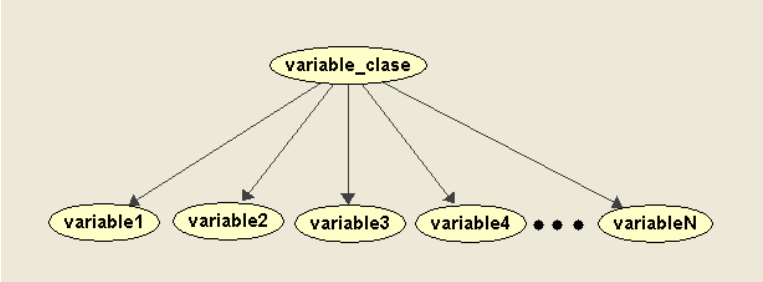
Neuronales, Árboles de decisión, entre otros, e incluso en algunos casos logran superarlos [Hernández, Ramírez, 2004].

La sencillez de este método radica en una suposición hecha en la construcción del modelo, ya que se asume una relación de independencia entre las variables predictoras conocida el valor de la variable clase.

6.1.4.1 CONSTRUCCIÓN DEL MODELO NAIVES BAYES

Teniendo como base la suposición nombrada anteriormente, la simplificación introducida en la construcción del modelo es demasiado notoria y permite que la estructura o modelo descriptivo se reduzca a un nodo raíz (variable clase), y un conjunto de nodos hijos (variables predictoras) con un único padre (variable clase). En la figura 14 se observa la estructura general de un modelo de Naives Bayes.

Figura 14. Modelo general de Naives Bayes.



Elaborado con el software Elvira.

6.1.4.2 INFERENCIA SOBRE EL MODELO NAIVES BAYES

Para inferir el modelo de clasificación construido, se parte de la regla más básica desde el punto de vista probabilístico para la inferencia, que es el teorema de Bayes. Como se menciono en la terminología, el teorema de Bayes se define de la siguiente manera:

$$P(A/X) = \frac{P(X/A)\cdot P(A)}{P(X)} = \frac{P(A,X)}{P(X)}$$

Ecuación 8. Teorema de Bayes

Al cambiar la sintaxis al clasificador Naives Bayes, el cual consta de una variable clase llamada C y un conjunto de variables predictoras llamadas A_1, A_2,… y A_n, el teorema de Bayes se define así:

$$P(C/A_1,A_2,...,A_n) = \frac{P(A_1,A_2,...,A_n/C)\cdot P(C)}{P(A_1,A_2,...,A_n)}$$

Ecuación 9. Teorema de Bayes para múltiples variables.

Como se observa de la ecuación anterior, aun no se contempla la posibilidad de que la variable clase tiene diferentes valores, es decir, que es una variable discreta. Al tener en cuenta este aspecto de que $c_k = \{c_1, c_2,\ldots, c_n\}$, el problema de la inferencia se transforma en identificar el estado de la variable clase con mayor probabilidad. Esta hipótesis con la probabilidad a posteriori más alta dada las variables predictoras se conoce como hipótesis MAP y aplicada a este problema se expresa así:

$$c_{MAP} = \arg máx P(c_k / A_1, A_2, ..., A_n) = \arg máx \frac{P(A_1, A_2, ..., A_n / c_k) \cdot P(c_k)}{P(A_1, A_2, ..., A_n)}$$

Ecuación 10. Hipótesis CMAP

La ecuación anterior se puede simplificar debido a que la probabilidad del denominador es la misma para cada posible estado de la variable clase (c_k), por lo tanto, la ecuación se transforma en:

$$c_{MAP} = \arg máx P(A_1, A_2, ..., A_n / c_k) \cdot P(c_k)$$

Ecuación 11. Hipótesis CMAP simplificada

Como el modelo Naives Bayes asume independencia entre las variables predictoras, la tabla de probabilidad $P(A_1, A_2, ..., A_n / c_k)$ se puede factorizar como el producto de n tablas que solo involucran dos variables para finalmente obtener la ecuación del c_{MAP} ajustada al modelo de clasificación de Naives Bayes así:

$$c_{MAP} = \arg máx \left(p(c_k) \bullet \prod_{i=1}^{n} p(A_i \mid c_k) \right)$$

Ecuación 12. Representación matemática de la hipótesis CMAP

La ecuación anterior se emplea para la inferencia del modelo de clasificación de Naives Bayes en general, pero la forma en que se calcula la probabilidad condicional $p(A_i|c_k)$ depende de la naturaleza de las variables predictoras, es decir,

de si éstas son variables discretas o continuas. Con la variable clase no se tiene este inconveniente porque queda claro que ésta es de naturaleza discreta.

6.1.4.2.1 INFERENCIA SOBRE EL MODELO CON VARIABLES PREDICTORAS DISCRETAS

Cuando las variables son discretas la probabilidad condicional se calcula teniendo en cuenta las frecuencias de aparición de los casos favorables y los casos totales que se obtienen de la base de datos empleada para el aprendizaje del modelo así:

$$P(A_i \mid c_k) = \frac{n(A_i, c_k)}{n(c_k)}$$

Ecuación 13. Probabilidad condicional

Esta ecuación se conoce como la estimación por máxima verosimilitud pero presenta una dificultad cuando el número de casos favorables es cero ya que el resultado de la probabilidad condicional también lo es. Para suavizar este resultado se usa el estimador basado en la ley de la sucesión de Laplace que consiste en agregar un caso más a los casos favorables y adicionar el número de estados de la variable clase al número de casos totales como se muestra en la siguiente ecuación:

$$P(A_i \mid c_k) = \frac{n(A_i, c_k) + 1}{n(c_k) + \Omega_c}$$

Ecuación 14. Estimador de Laplace.

Con esta última ecuación no se presentan resultados nulos para la probabilidad condicional como si ocurren en la estimación por máxima verosimilitud.

6.1.4.2.2 INFERENCIA SOBRE EL MODELO CON VARIABLES PREDICTORAS CONTINÚAS

Cuando las variables predictoras son continuas se asume que estas siguen distribuciones de probabilidad normal y las probabilidades condicionales son calculadas de la siguiente manera:

$$P(A_i \mid c_k) = N(\mu,\sigma) = \frac{1}{\sqrt{2\cdot\pi}\cdot\sigma}\exp\left(-\frac{(X-\mu)^2}{2\cdot\sigma^2}\right)$$

Ecuación 15. Probabilidades condicionales para variables continúas

Cada variable predictora tiene su propia media y desviación estándar. El valor X representa el valor de la evidencia de la variable predictora con el cual se desea inferir la variable clase.

6.1.5 CLASIFICADOR TREE AUGMENTED NAIVE BAYES (TAN)

TAN (del inglés *Tree Augmented Naives Bayes*) es una versión modificada del clasificador Naives Bayes. Este clasificador a diferencia del Naives Bayes, establece ciertas relaciones de dependencia entre las variables, sin dejar de tener en cuenta la relación existente entre cada una de las variables y la variable clase (para el caso de este trabajo de grado es el precio), intentando de esta manera mejorar el nivel de acierto del clasificador. Esta estructura sigue siendo una Red Bayesiana con forma de árbol, es acíclica y además cada una de las variables predictoras tiene máximo dos padres.

El establecimiento de las relaciones de dependencia entre cada una de las variables se da a partir de una pequeña modificación del algoritmo base de

Información Mutua, llamado *Información Mutua Condicionada*; el cual mide la cantidad de información que una variable Y proporciona sobre otra variable X supuesto que el valor de la variable clase (C) es conocido. La formulación matemática es:

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

Ecuación 16. Información Mutua Condicionada.

6.1.5.1 CONSTRUCCIÓN DEL MODELO TAN

Para la construcción del modelo del clasificador TAN, a continuación se ilustra el funcionamiento del algoritmo paso a paso:

- Calcular la Información Mutua Condicionada, *I(X,Y/C),* para cada una de las posibles relaciones entre todas las variables exceptuando la variable clase, asignando a cada una de las relaciones su valor correspondiente.

- Ordenar las Informaciones Mutuas Condicionadas *I(X,Y/C)* de mayor a menor.

- Partir de un árbol, en el cual no existan arcos ni relaciones entre las variables.

- Unir las dos variables que arrojaron el mayor peso entre las Informaciones Mutuas Condicionadas.

- Asignar la siguiente unión de mayor peso y añadirla al árbol sin que se forme un ciclo, si se llega a formar se descarta y se examina el siguiente arco con mayor peso.

- Repetir el paso anterior hasta recorrer todas las informaciones mutuas condicionadas calculadas.

- Escoger una variable como nodo raíz, para así empezar a direccional cada uno de los arcos.

- Añadir la variable clase (C) y las direcciones de los arcos a cada una de las variables predictoras.

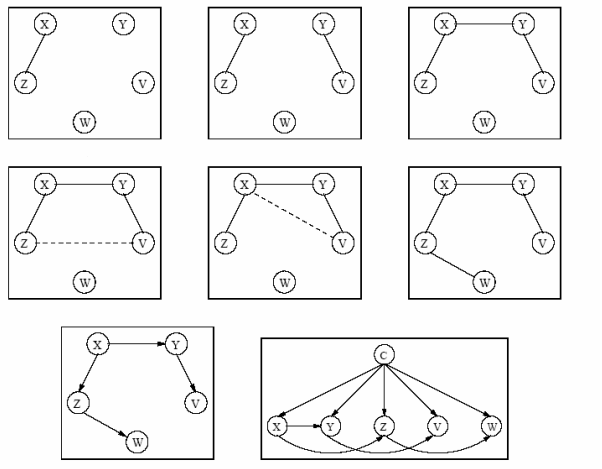
Para una total compresión del algoritmo TAN, la figura 15 da una idea de la construcción del árbol. Previamente se tienen ordenadas hipotéticamente, una serie de variables, que se muestran en la Tabla 7.

Tabla 7. Informaciones Mutuas Condicionadas de un conjunto de variables ordenadas de mayor a menor [Larrañaga & Inza, 2000].

Orden	*IMC*
1	X,Z
2	Y,V
3	X,Y
4	Z,V
5	X,V
6	Z,W
7	X,W
8	Y,Z
9	Y,W
10	V,W

Figura 15. Construcción del árbol de TAN teniendo en cuenta el orden establecido de la IMC(X,Y/C). La construcción se da a partir del algoritmo establecido para el mismo. En el último recuadro se puede ver el árbol TAN obtenido finalmente.

Figura tomada de [Larrañaga, Inza, 2000]



6.1.5.2 INFERENCIA EN EL CLASIFICADOR TAN

Debido a que la variable clase tiene *k* posibles valores, al igual que las variables predictoras, lo que interesa es identificar el estado más probable de todos los posibles estados que puede tomar la variable clase. Dentro de los clasificadores Bayesianos, esta probabilidad es aquella que presenta la máxima probabilidad a posteriori dado los atributos (variables), y es conocida como la hipótesis máxima a posteriori o hipótesis MAP (*maximum a posteriori*).

La estimación del estado más probable para el clasificador TAN parte del producto de todas las posibles probabilidades condicionales que se pueden dar (uniones presentes en el árbol que arrojó la construcción del algoritmo TAN). En este producto están presentes la probabilidad a priori de la variable clase, las probabilidades a posteriori de cada una de las variables predictoras dada la variable clase (C) y las probabilidades a posteriori de las variables predictoras dado algunas de las mismas variables predictoras (dado por las informaciones mutuas condicionadas).

Otra manera sencilla de entender la fórmula para el cálculo de la inferencia del clasificador TAN es recorriendo todas las variables que conforman la estructura del árbol, y en cada una de ellas observar la cantidad de arcos que le llegan. A partir de los enlaces que lleguen a estas, estas serán las probabilidades condicionales establecidas conjuntamente entre la variable estudiada y aquellas de donde vienen los enlaces, a excepción de la variable clase para la cual solo se evalúa su probabilidad marginal. Tras haber recorrido todas las variables se hace el producto entre todas las probabilidades condicionales y la probabilidad marginal de la variable clase.

Para el modelo clasificatorio TAN obtenido en la figura 15, se obtiene la expresión para la inferencia mostrada en la ecuación 17:

$$P(c \mid x, y, z, v, w) = P(c) \cdot P(x \mid c) \cdot P(y \mid x, c) \cdot P(z \mid x, c) \cdot P(v \mid y, c) \cdot P(w \mid z, c)$$

Ecuación 17. Inferencia para el modelo clasificatorio TAN para el ejemplo de la figura 15.

Sin embargo, algunas de estas probabilidades condicionales necesitan corregirse, ya que algunas de las frecuencias de aparición que se obtienen de la base de datos aparecen como cero y esto llevaría a dar una probabilidad de cero para un posible estado de la variable clase en el modelo. Esta corrección es similar a la hecha para el clasificador Naives Bayes. De esta manera con esta corrección, se esta garantizando una mínima probabilidad para todas las configuraciones y estados posibles. El estimador que se usa para hacer dicha corrección se denomina *estimador basado en la ley de la sucesión de Laplace*, el cual se detalla a continuación:

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

Ecuación 18. Estimador basado en la ley de la sucesión de Laplace.

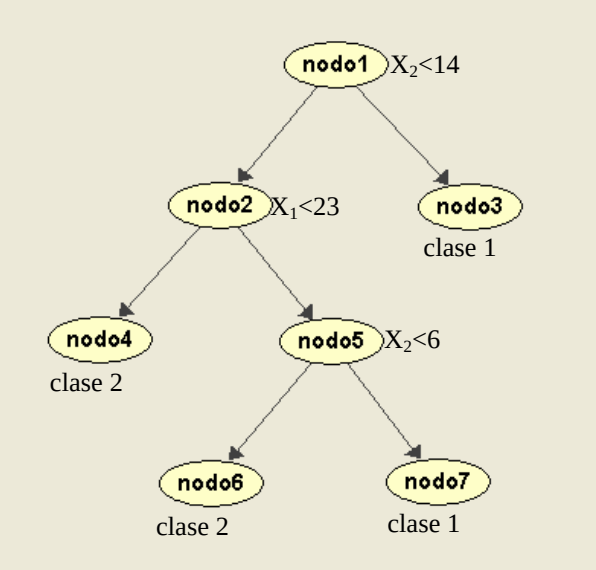
6.2 ÁRBOLES DE CLASIFICACIÓN.

Para describir la sensibilidad de los precios de oferta en bolsa de los generadores del mercado de energía mayorista colombiano basado en modelos gráficos se explora como una alternativa la técnica conocida como Árboles de clasificación, cuya motivación nace de la similitud de esta técnica con las Redes Bayesianas permitiendo un doble uso, descriptivo (para desarrollar un modelo gráfico) y predictivo.

Los árboles de clasificación son una técnica empleada para el reconocimiento de patrones[27]. Esta técnica, como otros métodos se basa en unas reglas definidas para la clasificación pero a diferencia de los mismos permite conocer cuales variables predictoras son las más importantes para el modelo.

Los árboles de clasificación representan un proceso de decisión multi-estado donde en cada estado se debe tomar una decisión binaria y están conformados de nodos internos y nodos terminales.

Figura 16. Esquema de un árbol de clasificación[28]



[27] Esta técnica es desarrollada en la Toolbox 'Computational Statistics Handbook with MATLAB, Statistical Pattern Recognition, Classification Trees'.

[28] Elaborado con el software Elvira.

En la figura 16 se pueden observar unos nodos que tienen asociadas una condición (nodo 1, nodo 2 y nodo 5), éstos nodos son llamados nodos internos y entre sus características más importantes están que tienen una condición asociada, poseen solo dos nodos hijos (nodo izquierdo y nodo derecho) y un solo nodo padre, exceptuando al nodo raíz el cual no tiene padre. Además de estos nodos internos existen otros nodos conocidos como nodos terminales (nodo 3, nodo 4, nodo 6 y nodo 7), los cuales se caracterizan por tener una clase asociada explicita[29]. Esta clase asociada a cada nodo (interno o terminal) de un árbol de clasificación corresponde con uno de los posibles estados de la variable clase y es calculada como la máxima probabilidad condicional entre todos los estados de la variable clase.

Cuando es construido un árbol de clasificación, se puede clasificar un nuevo caso básicamente evaluando éste en el árbol construido, teniendo en cuenta que cuando se evalúa la condición de un nodo interno y se cumple, el siguiente paso continua en el nodo hijo izquierdo mientras que sino se cumple, el siguiente paso es el nodo hijo derecho. También se debe tener en cuenta que el proceso de evaluar el árbol termina al evaluar un nodo terminal porque al ocurrir esto automáticamente se le asigna una clase al nuevo caso y la clasificación esta terminada.

A continuación se ilustran un par de ejemplos donde se asume que el árbol de clasificación mostrado en la figura 16 fue construido con base en dos variables predictoras (X_1, X_2) y una variable clase con dos posibles estados (clase 1, clase 2). Como primer ejemplo, X_1 y X_2 toman los valores de 10 y 25 respectivamente. Al evaluar este nuevo caso en el nodo raíz donde la condición es $X_2<14$, ésta no se cumple debido a que X_2 es 25 entonces como resultado se continua en el nodo

[29] Un nodo interno también tiene asociada una clase la cual permanece implicita, pero que con el proceso de podado (Ver 5.2.2) podría pasar a una clase asociada explicita si el nodo interno es modificado a nodo terminal durante el proceso.

hijo de la derecha (nodo 3), el cual es un nodo terminal por lo tanto queda definida la clasificación a la clase 1 sin importar el valor de la variable X_1.

Como segundo ejemplo X_1 y X_2 toman los valores de 50 y 2 respectivamente. Al evaluar este nuevo caso en el nodo raíz, donde la condición es $X_2<14$, esta se cumple debido a que X_2 es 2 entonces como resultado se continua en el nodo hijo de la izquierda (nodo 2) el cual no es un nodo terminal por lo tanto no queda definida la clasificación. Al contrario por ser el nodo 2 interno, aparece una nueva condición $X_1<23$ la cual no se cumple porque X_1 es 50 por lo tanto el proceso continua en el nodo hijo derecho (nodo 5) que es un nodo interno con la condición $X_2<6$. La anterior condición se cumple porque X_2 es 2 y el siguiente paso continúa en el nodo hijo izquierdo (nodo 6) el cual es un nodo terminal y el nuevo caso es clasificado en la clase 2.

Conocido de manera general el funcionamiento de un árbol de clasificación se procede a explicar los aspectos más importantes desde la construcción del árbol hasta la inferencia en el mismo. La técnica de los árboles de clasificación consta básicamente de tres etapas principales que son el crecimiento del árbol, el podado del árbol para obtener subárboles y la elección del mejor subárbol.

6.2.1 CRECIMIENTO DEL ÁRBOL[30]

Para encontrar el modelo gráfico de esta técnica de clasificación, el primer paso es realizar el proceso de crecimiento del árbol de clasificación con el set de datos de aprendizaje, el cual consiste en cortar el set de datos inicial en particiones cada vez más pequeñas de tal manera que las particiones creadas lleguen a ser puras en terminos del grupo de la clase.

[30] Para ver detalladamente el programa ver Anexo G.

Para realizar el crecimiento de un árbol es importante definir las dos condiciones del proceso que son:

- Condición de parada: indica donde el proceso de crecimiento del árbol termina. En general, la condición de parada se cumple cuando todos los nodos terminales presentes en el árbol cumplan con el número máximo de datos por nodo que es la variable conocida como 'maxn' y además la condición de impureza que permite detener el proceso si todos los datos pertenecen a una misma clase. La medida de impureza utilizada es 'Gini diversity index' y se define como i(t) donde $p(w_j|t)$ corresponde a las probabilidades condicionales de los posibles estados de la variable clase respecto al nodo del árbol que se esta analizando.

$$i(t) = 1 - \sum_{j=1}^{J} p^2(w_j \mid t)$$

Ecuación 19. Medida de impureza para Árboles de clasificación

- Condición de Corte: es usada para que a partir de un nodo del árbol, se construyan sus dos nodos hijos (nodo izquierdo y nodo derecho) siempre y cuando se tenga presente que el objetivo es elegir el corte que obtenga el mayor decrecimiento de impureza del corte s en el nodo t, que se denota $\Delta i(s,t)$. Donde p_R y p_L son la proporción de datos enviados a la derecha y a la izquierda respectivamente.

$$\Delta i(s,t) = i(t) - p_R i(t_R) - p_L i(t_L)$$

Ecuación 20. Decrecimiento de impureza.

A continuación se realiza un ejemplo con el generador Guatrón para revisar con más detalle el proceso de crecimiento de un árbol de clasificación. En este ejemplo se establece un valor máximo de datos por nodo de 300 (maxn) y la base de datos para la construcción del árbol esta definida con 1089 datos de aprendizaje.

Antes del crecimiento del árbol de clasificación es importante el cálculo del número y proporción de datos para cada estado de la variable precio. Esta proporción de datos son las mismas probabilidades a priori.

En este momento el árbol de clasificación cuenta con un solo nodo, el cual por defecto es establecido como nodo terminal y el número de datos pertenecientes al nodo son los 1089 datos de aprendizaje. La medida de impureza de este nodo es calculada con la ecuación 19 utilizando las probabilidades a priori (proporción de datos) y para este ejemplo práctico el valor de impureza del primer nodo es 0.55126. Este nodo no tiene padre por ser el nodo raíz y ahora la tarea es encontrar los dos nodos hijos, para lo cual debemos hallar la variable de corte y el nodo de corte de este nodo.

A partir de este momento empieza el crecimiento del árbol de clasificación y para esto se deben inspeccionar las condiciones de parada. El proceso de crecimiento debe continuar en un nodo si el nodo es terminal, si la impureza es mayor que cero y si el número de datos del nodo es mayor que el número máximo de datos permitidos que en este caso es 300. En este ejemplo las tres condiciones se cumplen porque el nodo raíz es terminal, la impureza es 0.55125 la cual es mayor que cero y el número de nodos es 1089 que es mayor a 300.

Para establecer cual será la variable de corte y el valor de corte se debe recorrer cada una de las variables predictoras y para cada una éstas se debe recorrer cada uno de lo posibles cortes. El corte de la variable que arroje el mayor decrecimiento

de impureza será elegido como el valor de corte de nodo y su respectiva variable como la variable de corte.

Siguiendo con este ejemplo el recorrido comienza con la variable predictora inicial. El primer paso para establecer los valores de corte es ordenar los datos de la variable predictora y eliminar los valores repetidos. Los primeros tres valores para la primer variable predictora del generador Guatrón son 1356, 4307 y 4327 y los valores de corte se forman así: los primeros valores de corte se obtienen operando los datos así:
Primer valor de corte: 1356 + (4307-1356)/2 = 2831.5
Segundo valor de corte: 4307 + (4327-4307)/2 = 4317
El tercer valor de corte se obtiene de operar el tercero y cuarto dato y así sucesivamente.

Siguiendo con el procedimiento, una vez tenemos el valor de corte, en este caso 2831.5, se encuentra el numero de datos de la variable que se esta analizando (en este caso la variable 1) que están por debajo de ese valor de corte los cuales irán al nodo izquierdo hijo y el número de datos que están por encima de ese valor de corte los cuales irán al nodo derecho hijo. En este ejemplo, para la primer variable predictora el número de datos por debajo del corte son 43 y el número de datos por encima del corte son 1088.

A continuación se calculan las probabilidades de la ecuación 20, p_L y $p_{R,}$ como el porcentaje de datos por debajo y por encima del valor de corte respectivamente, los cuales en este ejemplo son 0.0009 y 9.9991. Posteriormente se obtiene para los dos posibles nodos hijos el número de datos por estado de la variable precio que junto con las probabilidades a priori y el número de datos por estado del precio del nodo padre (en este caso el nodo raíz) son necesarios para calcular las probabilidades condicionales para cada uno de los dos nodos hijos.

Con las probabilidades condicionales se establecen las medidas de impureza para cada uno de los nodos hijos utilizando la ecuación 19. En este momento con los datos de impureza de los nodos padre e hijos y con las proporciones de datos de cada uno de los nodos hijos, se calcula la medida de decrecimiento de la impureza dada por la ecuación 20. El objetivo es encontrar el mayor valor de decrecimiento de impureza de entre todos las variables y todos lo cortes posibles que puedan existir para cada variable. Una vez elegida la variable de corte y el valor de corte, se adicionan los nodos hijos (izquierdo y derecho) definiendo para cada uno un conjunto de datos, un conjunto de probabilidades, etc. El proceso continúa en los nodos hijos y así sucesivamente mientras no se cumplan las reglas de parada mencionadas al principio del ejemplo.

Para desarrollar el procedimiento de crecimiento de un árbol de clasificación con MATLAB se cuenta con una función llamada ‘csgrowc.m’ desarrollada en [Martinez & Martinez, 2002]. Esta función no se uso tal como aparece en la Toolbox debido a que se realizaron ciertas modificaciones para que se ajustara a los datos utilizados y en algunos casos se corrigieron algunos errores encontrados. Con base en lo anterior se creo una nueva función llamada ‘csgrowcCORREGIDO.m’ que fue la finalmente utilizada en este trabajo de grado y sus detalles se pueden observar en el Anexo E.

Los requisitos para poder aplicar esta función de crecimiento son la base de datos de aprendizaje para las variables predictoras y la variable clase, definir el número máximo de datos por nodo (maxn), definir el número de estados de la variable clase, el número de casos asociados a cada estado y finalmente las probabilidades a priori de cada estado de la variable clase. La descripción del programa para el crecimiento del árbol de clasificación se realiza en el anexo G.

6.2.2 PODANDO EL ÁRBOL

Una vez es terminado el proceso de crecimiento descrito en el numeral anterior se obtiene un árbol de clasificación de gran tamaño. Ahora el objetivo se concentra en minimizar el coste de la complejidad y esto se consigue mediante el proceso de podado del árbol.

La complejidad de un árbol se basa en el número de nodos terminales ($|\ T\ |$) y la medida del coste de la complejidad es definida como:

Ecuación 21. Coste de complejidad

$$R_\alpha(T) = R(T) + \alpha|\mathrm{T}\ |;\alpha\rangle 0$$

El proceso de podado busca un árbol que minimize el coste de complejidad dado en la ecuación anterior, donde el parámetro α es el coste de complejidad por nodo terminal. Si se tiene un gran árbol donde cada nodo terminal contiene observaciones de una sola clase, entonces R(T) será cero. Sin embargo habrá una penalización por la complejidad y la medida del coste de la complejidad sería $R_\alpha(T) = \alpha|\mathrm{T}\ |;\alpha\rangle 0$. Si α es pequeño, entonces la penalización pagada por tener un árbol complejo es pequeña y el árbol resultante es grande. El árbol que minimiza $R_\alpha(T)$ tenderá a tener pocos nodos y una gran α [Martinez & Martinez, 2002].

Reconocida la causa que origina el podado se procede a explicar el proceso, el cual se realiza recorriendo todos los nodos internos que son los que forman las ramas del árbol y evaluando una función conocida como misclassification rate [Martinez & Martinez, 2002] que se evalúa para cada nodo interno y sus dos respectivos nodos hijos (nodo izquierdo y nodo derecho) y es definida asi:

Ecuación 22. Error de malas clasificaciones

$$R(T) \geq R(t_L) + R(t_R)$$

Esta ecuación indica que el error de malas clasificaciones en el nodo padre es mayor o igual que la suma en los nodos hijos. Finalmente, el proceso busca los nodos que satisfagan $R(T) = R(t_L) + R(t_R)$ y esos nodos hijos son podados con todos sus descendientes.

En el proceso de podado de un árbol de clasificación se destacan dos fases: la primera consiste en la eliminación de todos los nodos terminales hermanos que tengan la misma clase asociada y la segunda es el cálculo del error de las malas clasificaciones mencionado anteriormente como R. Este dato de error en un nodo es calculado como el producto de la medida de impureza y la probabilidad marginal del respectivo nodo.

Para desarrollar el procedimiento de podado de un árbol de clasificación con MATLAB se cuenta con una función llamada ‘csprunec.m’ desarrollada en [Martinez & Martinez, 2002]. Esta función no se uso tal como aparece en la Toolbox debido a que se realizaron ciertas modificaciones para que se ajustara a los datos utilizados y en agunos casos se corrigieron errores encontrados. Estas modificaciones conllevan a una nueva función llamada ‘csprunecCORREGIDO.m’ que fue la finalmente utilizada en este trabajo de grado. Para observar los detalles de las correcciones ver Anexo F.

El requisito para poder aplicar esta función de podado es definir el árbol de clasificación principal que es obtenido usando la función ‘csgrowcCORREGIDO.m’. Para ver detalladamente el programa ver Anexo H.

6.2.3 ELECCIÓN DEL MEJOR SUBÁRBOL

Para la elección del mejor subárbol o validar un árbol de clasificación se usa una base de datos independiente (datos para la inferencia) a la base de datos de aprendizaje para evaluar cada uno de los subárboles.

6.3 HERRAMIENTAS COMPUTACIONALES.

Entre las herramientas computacionales usadas para el desarrollo de este trabajado de grado se encuentran Elvira y MATLAB.

6.3.1 ELVIRA.

Es un software que manipula y crea Redes Bayesianas. Posee interfaz gráfica, se puede ver el código fuente, tiene interfaz de programación y es gratuito.

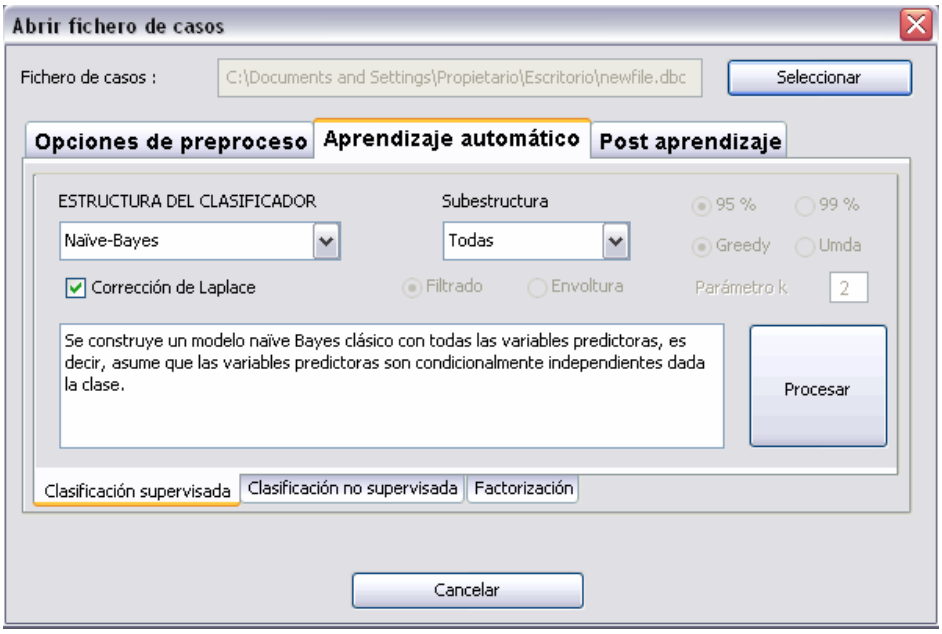
Elvira nació a partir de la idea de varias universidades españolas y entre las principales características usadas en este trabajo se destacan:

- Aprendizaje de Redes Bayesianas.
- Clasificadores Naives Bayes y TAN.
- Algoritmos de pre-procesamiento como discretización de variables, ranking de variables y selección de atributos.

En la figura 17 se observa la ventana principal para el aprendizaje automático con el software Elvira, desde esta ventana se ejecutan los clasificadores Naives Bayes y TAN que corresponden a una clasificación supervisada.

A partir de las tareas de los clasificadores Bayesianos en Elvira, se puede observar la Red Bayesiana correspondiente (grafo) y así mismo realizar la tarea de inferencia para dicho clasificador.

Figura 17. Fichero de casos para el software Elvira.



Tomado del software Elvira.

6.3.2 COMPUTATIONAL STATISTICS HANDBOOK WITH MATLAB, STATISTICAL PATTERN RECOGNITION, CLASSIFICATION TREES.

El ‘Computacional Statistics Handbook with MATLAB’ es un libro que se enfoca en los métodos estadísticos computacionales más comúnmente utilizados y en la implementación de los mismos en MATLAB. Como ventajas de este libro se

encuentra que se incluye la descripción de los algoritmos para los procedimientos y también se proveen ejemplos ilustrativos para el correcto uso de los algoritmos.

Algunos de los temas de la estadística computacional abordados por este libro son: conceptos de probabilidad, variables aleatorias, análisis de datos, inferencia estadística, etc.

El tema de estadística para Reconocimiento de patrones, más exactamente Árboles de clasificación, presente en el capitulo 9 del libro resulta útil para el desarrollo de esta tesis por las siguientes razones:

- En su función para reconocimiento de patrones, estos Árboles de Clasificación se puede emplear como un método de clasificación supervisada.

- Permite por medio de su estructura identificar cuales variables predictoras son las más importantes en el análisis de la sensibilidad del precio.

Para entender como es el funcionamiento de los Árboles de Clasificación remitirse a la sección 6.2.

6.3.3 PROGRAMAS PARA LOS CLASIFICADORES BAYESIANOS Y ÁRBOLES DE CLASIFICACIÓN CON MATLAB.

Para la programación de los métodos de clasificación se crearon las siguientes funciones en MATLAB:

- Clasificadores.m: esta función es la encargada de originar los resultados (en un archivo de texto) de las eficiencias de todos los métodos de

clasificación. Previamente se debe haber ejecutado la función discretizadores.m.

- Discretizadores.m: esta función es la encargada de discretizar los datos de aprendizaje e inferencia y entregarlos ordenados en archivos de Excel para ser posteriormente utilizados por la función clasificadores.m.

- TAN2G.m: esta es la función encargada de establecer los enlaces en el clasificador TAN, es decir, esta función establece la estructura del modelo de clasificación TAN para posteriormente usar la función inftan2.m (Ver sección 5.1.5).

- Inftan2.m: esta es la función encargada de desarrollar la inferencia para obtener la eficiencia del clasificador TAN. Se asume que previamente se ejecuto la función TAN2G.m (Ver sección 5.1.5 referente al clasificador TAN).

- Metodo1.m: esta es la función encargada de discretizar los datos de aprendizaje e inferencia de acuerdo al método 1 de discretización (Ver sección 6.3.2.3 referente a las formas de discretización usadas).

- Metodo2_999.m: esta es la función encargada de discretizar los datos de aprendizaje e inferencia de acuerdo al método 2 de discretización (Ver sección 6.3.2.3 referente a las formas de discretización usadas).

- MutualDiscreto.m: esta es la función encargada de realizar el ranking de variables predictoras [Martínez & Zárate, 2007]. (Ver definición de información mutua en la sección 5.1.1 referente a términos relacionados con métodos de clasificación).

- Nb.m: esta es la función encargada de realizar la inferencia con el método de clasificación Naives Bayes discreto (Ver sección 5.1.4 referente al clasificador Naives Bayes).

- nbContinuo.m: esta es la función encargada de realizar la inferencia con el método de clasificación Naives Bayes continuo (Ver sección 5.1.4 referente al clasificador Naives Bayes).

- rangos2.m: esta es la función encargada de definir los límites de cada grupo que se obtiene de discretizar las variables.

- selecvar.m: esta es la función que a partir del ranking de variables evalúa el método del codo para efectuar la selección de variables (Ver la sección 6.3.3 referente a la selección de variables).

- treefinal.m: esta es la función encargada de evaluar los árboles de clasificación discretos y continuos (Ver sección 5.2 referente a árboles de clasificación).

- Ficherodecasos2.m: esta es la función encargada de tomar los datos de aprendizaje y adecuarlos al formato de la aplicación Elvira para posteriormente obtener el ranking de variables predictoras, la selección de atributos por el método del codo, las estructuras de los clasificadores Naives Bayes y TAN y finalmente inferencias.

- NaivesDtres.m: Es la función encargada de realizar las dobles probabilidades condicionales entre cada uno de los estados de la variable clase y las variables predictoras.

- NaivesD.m: Usa el modelo de Naives Bayes discreto, para buscar los estados mas representativos en la fijación del precio en bolsa.

El detalle de todas las funciones anteriormente nombradas para clasificadores con Redes Bayesianas y los programas corregidos para Árboles de clasificación se encuentran en los anexos.

7. MODELO GENERAL PARA EL ANÁLISIS DE SENSIBILIDAD DE LOS PRECIOS DE OFERTA DE LOS GENERADORES.

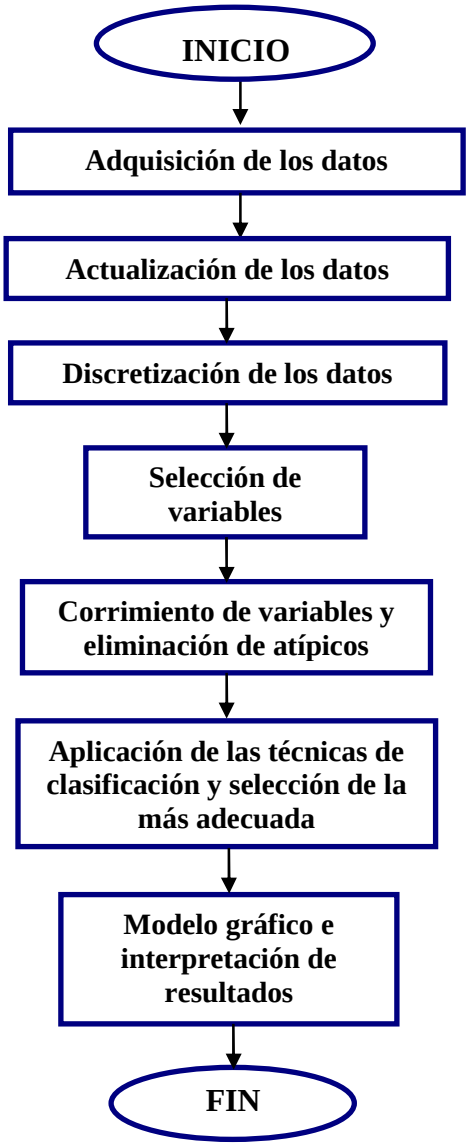
El principal objetivo del análisis de sensibilidad del precio de oferta de un generador es encontrar como se comportan los precios de oferta de los generadores ante la variación de las variables predictoras, de manera que se pueda inferir un precio teniendo certeza del estado de las demás variables y finalmente observar cómo estas pueden impactar en la competencia de la oferta.

Por esta razón este capitulo es de gran relevancia, puesto que explica detalladamente cada uno de los pasos del modelo general que se deben realizar para el análisis de la sensibilidad de los precios de oferta de un generador en particular. En este capitulo se encuentran detalles desde cómo encontrar los datos de los generadores hasta finalmente seleccionar el clasificador más adecuado para el análisis de sensibilidad de un generador basado en los resultados de los métodos de clasificación.

7.1 MODELO GRÁFICO PARA LA DESCRIPCIÓN GENERAL DEL PROCESO.

La figura 18 resume el modelo general aplicado a cada uno de los generadores para el estudio de la sensibilidad de los precios.

Figura 18. Modelo general para el análisis de sensibilidad de precios.

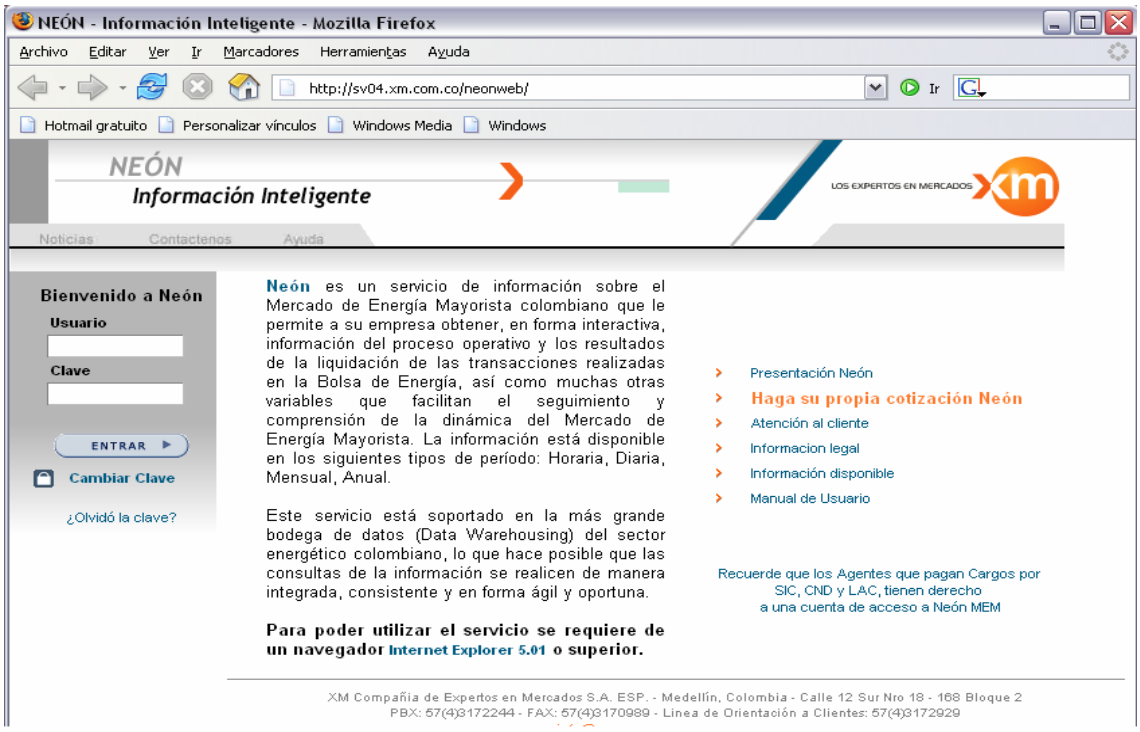


7.2 ADQUISICIÓN DE DATOS.

La manera de adquirir los datos fue a través de la base de datos de Neón, publicada de manera digital a través del sitio de Internet http://sv04.xm.com.co/neonweb/.

Neón es servicio de información sobre el Mercado de Energía Mayorista colombiano que permite obtener, en forma interactiva, información del proceso operativo y los resultados de la liquidación de las transacciones realizadas en la Bolsa de Energía, así como muchas otras variables que facilitan el seguimiento y comprensión de la dinámica del Mercado de Energía Mayorista. La información está disponible de manera horaria, diaria, mensual y Anual.

Figura 19. Sitio Web de Neón



Tomado de http://sv04.xm.com.co/neonweb/

7.3 ADECUACIÓN DE LOS DATOS.

La adecuación de los datos es el trabajo previo a la tarea de construcción de los clasificadores, la resolución de hipótesis y la inferencia de resultados acerca del estado de la variable clase.

Este proceso requiere de un tratamiento especial de la información obtenida de la base de datos digital, de tal manera que algunas variables como el tiempo, el número de grupos para la discretización, el método de discretización empleado, entre otras, modelen de una manera más adecuada el comportamiento de la oferta de los generadores y las variables estratégicas determinadas para el comportamiento de dicha oferta, de modo que la inferencia de los datos de aprendizaje resulte en los mejores resultados y comportamientos claves.

7.3.1 CORRIMIENTO DE VARIABLES Y ELIMINACIÓN DE ATÍPICOS.

7.3.1.1 CORRIMIENTO DE VARIABLES

Para explicar el comportamiento del precio de bolsa de un generador, algunas variables de la base de datos necesitan ser desfasadas en el tiempo. La razón es que hay que comportarse como el generador, es decir, se debe usar la información que el tiene disponible antes de declarar su precio antes de las 8 AM del día anterior al despacho.

El día de referencia para los desfases es el día del despacho. El precio y la disponibilidad se declaran el día anterior al despacho pero se refieren a la oferta del día siguiente, es decir, el día del despacho, por lo tanto no se deben desplazar estas variables. Los contratos generalmente son a largo plazo, por lo tanto su valor es conocido con anterioridad por el generador y no debe estar desplazado. El embalse propio del generador es medido cada día antes de las 8 AM del día

anterior al despacho, este valor se publica para el día que es medido, por lo tanto, la variable embalse propio debe desplazarse (adelantarse) un día. Para el embalse agregado o embalse del sistema el generador conoce el dato del día anterior al Predespacho ideal, es decir, se conoce el valor dos días atrás del despacho. Por lo tanto la variable embalse agregado debe adelantarse dos días. Los datos de reconciliaciones son conocidos por los generadores un día después del despacho, es decir, el día anterior al despacho el generador conoce el valor publicado el día anterior pero correspondiente a dos días atrás al Predespacho ideal, por lo tanto las variables de reconciliación deben adelantarse tres días. Las variables implicadas en la curva de demanda residual no se deben desplazar, debido a que cuando estas variables se crearon se tuvieron en cuenta los desfases.

A manera de resumen se muestra en la tabla 8 las variables y el número de días que han de desplazarse estas.

Tabla 8. Días a desplazar para las variables en estudio.

Variables	Disponibilidad, Contratos	Precios, CDR[31]	Reconciliaciones	Embalse agregado	Embalse propio
Desfase	0	0	3	2	1

7.3.1.2 ELIMINACIÓN DE ATÍPICOS.

Se entiende por atípicos a aquellos elementos presentes en las variables los cuales no representan adecuadamente el comportamiento de las variables del mercado. La importancia de la eliminación de estos elementos atípicos consiste en

[31] CDR, curvas de demanda residual. P1Dx, P2Dx, P1Dd, P2Dd, P1Dn, P2Dn

no crear grupos innecesarios (en la discretización) que no representan realmente el comportamiento dinámico del Mercado de Energía Eléctrica, los cuales reducirían la inferencia de los clasificadores Bayesianos y de los Árboles de clasificación.

El reconocimiento de estos datos atípicos se agrupó en distintos casos los cuales se muestran a continuación:

- Caso 1. Cuando la inflexibilidad sea igual a la disponibilidad: cuando un generador declara una inflexibilidad igual a su disponibilidad es un claro indicio de no desea participar en la bolsa, en el despacho por orden de méritos.

- Caso 2. Cuando no aparece datos de demanda residual: la demanda residual es una de las principales variables a analizar, por lo tanto debe estar presente en el estudio.

- Caso 3. Cuando un generador declara precio nulo: cuando un generador declara un precio nulo, es un claro indicio de que no desea salir despachado. Para aquellos días en que no hay una disponibilidad declarada o un precio nulo, o ambos, no existen datos de la curva de demanda residual.

- Caso 4. Cuando el precio de oferta de un generador hidráulico es mayor o igual al mayor precio de los generadores térmicos[32]. En este caso, el generador no desea ser despachado bajo ninguna circunstancia, además que estos elementos no son comunes dentro de las bases de datos, constituyendo un comportamiento anormal.

[32] Mayores detalles acerca de la penalización que se les impone a los generadores hidráulicos se detalla en la resolución CREG 018 de 1998.

7.3.2 ACTUALIZACIÓN DE LOS DATOS

Aquellas variables que involucran precios y que están establecidas en un cierto periodo de tiempo requieren ser actualizadas, de tal manera que puedan ser tratadas y comparadas en un mismo periodo de tiempo. Así mismo, es necesario extraer de los precios de cada uno de los agentes generadores, aquellos componentes que sean constantes para todos los precios de oferta.

En este orden de ideas, la actualización de precios que implica llevar todo a un mismo periodo de tiempo, se toma para los años 2003, 2004, 2005 y el primer semestre del 2006. El proceso de actualización de los precios, se realiza de una manera similar a la actualización de uno de los componentes presentes y obligatorios en la declaración de los precios, el FAZNI (Fondo de Apoyo Financiero de las Zonas No Interconectadas); con el cual se apoya y sostiene aquellas zonas en el país a las cuales no puede llegar el Sistema Interconectado Nacional. Este elemento se actualiza periódicamente con el Índice de Precios al Productor (IPP), un índice calculado por el Departamento Nacional de Estadística (DANE), el cual se establece mes a mes, y con el cual se asocia la variación de los precios de bienes intermedios, es decir, bienes y servicios que se utilizan para la producción de otros bienes.

Con la actualización de los precios de oferta de los generadores con el IPP, se representa de una manera análoga el papel económico de los agentes generadores, ya que se podrían clasificar estos, como productores de bienes intermedios, entendiendo a la energía eléctrica en particular como un bien o servicio, útil para la producción de otros bienes o servicios y en el confort de las personas. No se debe confundir el IPP, con el Índice de Precios al Consumidor (IPC) el cual mide el cambio a través del tiempo de la canasta familiar. Tanto el IPC, como el IPP reflejan la inflación o deflación de la economía colombiana.

Un elemento presente en todos los precios de oferta de los agentes generadores, es el Costo Equivalente de la Energía (CEE), el cual es el costo del cargo por capacidad[33] en [$/kWh] usado para la oferta diaria de los generadores en la bolsa. Este elemento es ajustado mensualmente por el Centro Nacional de Despacho (CND); debajo de este valor, los generadores no pueden declarar sus precios de oferta. Se podría decir, que este es el componente base de las ofertas de los precios de todos los generadores. Este elemento es el que se va a sustraer de las ofertas diarias de los generadores y de todas aquellas variables que involucren precio diariamente.

Los precios que se actualizan son los precios de ofertas de los generadores y los precios resultantes de la curva de demanda residual para las demandas máxima, mediana y mínima. La actualización de los precios se hace de la siguiente manera:

- Se identifican las unidades de los precios de oferta de los generadores o de la curva de demanda residual y los del CEE. Teniendo en cuenta los precios de los generadores y de la demanda residual en [$/MWh] y los del CEE en [$/kWh], el cálculo realizado teniendo en cuenta que el CEE, cambia mensualmente es:

$$A = \mathrm{Pr}\,ecioOferta - 1000 \cdot CEE$$

Ecuación 23. Actualización del precio por el CEE. A es el precio actualizado.

- Tras haber sustraído el CEE mensual para estos precios, se actualizan todos los datos al mes de Junio de 2006. Para esto se multiplica el valor del precio que se quiere actualizar por la relación entre el IPP de Junio de 2006 y el IPP del mes correspondiente al precio que se quiere actualizar.

[33] El cargo por confiabilidad entró a reemplazar la función del cargo por capacidad.

$$B = A \cdot \frac{IPP_{Junio\,2006}}{IPP_{MesActual}}$$

Ecuación 24. Actualización del precio por IPP. B es el precio actualizado

De esta manera quedan actualizados los precios por el CEE e IPP para su posterior tratamiento y análisis dentro de los modelos Bayesianos.

Otra de las actualizaciones necesarias es la disponibilidad real del generador, la cual se hace sustrayendo de la disponibilidad declarada la inflexibilidad[34] del generador para dicho día, mostrando así la disponibilidad que tendrá efecto dentro de la construcción del precio de oferta y del despacho ideal.

$$Disponibilidad\,\mathrm{Re}\,al = DisponibilidadDeclarada - Inflexibilidad$$

Ecuación 25. Disponibilidad real de un agente generador.

La última de las actualizaciones es la de los contratos en [$/MWh], obtenidos como el cociente entre los Contratos en miles de pesos y los contratos en MWh (con su respectivo análisis de unidades) y su actualización por IPP es similar a la presentada para las actualizaciones de los precios de oferta.

$$Contratos\left[\frac{\$}{MWh}\right] = \left(\frac{Contratos[Miles\$]\cdot 1000}{Contratos[MWh]}\right)\cdot\left(\frac{IPP_{Junio\,2006}}{IPP_{MesActual}}\right)$$

Ecuación 26. Actualización de la variable Contratos [$/MWh]

[34] Es la generación mínima indispensable para rodar la máquina sea por condiciones técnicas, operativas o ambientales (hacer correr un mínimo caudal para un río proveniente de un embalse)

7.3.3 DISCRETIZACIÓN DE LOS DATOS

La discretización de los datos es un paso clave para la caracterización de las ofertas y el modelado de la Red Bayesiana. La discretización consiste en la asignación de grupos o estados a un cierto conjunto de datos ya que en alguna característica propia de los mismos, poseen cierta semejanza. Para encontrar estas semejanzas existen ciertas técnicas y algoritmos que se detallan mas adelante.

La discretización de los datos se hace tanto para clasificadores Bayesianos y Árboles de clasificación discretos y continuos. Teniendo en cuenta que cuando son métodos de clasificación discretos se discretiza la variable clase y las variables predictoras mientras que cuando son métodos de clasificación continuos solo se discretiza la variable clase.

7.3.3.1 CRITERIO PARA LA SELECCIÓN DEL NÚMERO DE GRUPOS.

La elección del número de grupos necesarios para la discretización de los datos muestra unos estados característicos tanto de la variable clase como de las demás variables predictoras, logrando de esta manera ubicar dentro de los grupos seleccionados características claras del comportamiento del Mercado de Energía.

En este orden de ideas, se identificaron ciertos estados característicos que puede tomar la variable clase (precio de oferta de los generadores) dadas ciertas condiciones de la demanda y de las condiciones climáticas de un año. Por condiciones de la demanda, se entiende, el estado de algunas demandas típicas de acuerdo al consumo diario, es decir una demanda máxima, una mediana y una mínima. Para estas demandas en particular, se podría asegurar que el precio formado para satisfacer la demanda máxima será el máximo entre las tres,

siguiendo en dicho orden luego el precio de la demanda mediana y por último el precio de la demanda mínima.

Así mismo por condiciones climáticas se pueden establecer ciertos estados típicos para su caracterización. Para este caso, podría tener tres estados: condición de verano, condición normal y condición de invierno. Por condición de verano, se entiende un bajo nivel de embalses y/o una disminución significativa del caudal de los ríos a los embalses; en condición normal los embalses se encuentran en sus niveles medios y/o un aporte normal del caudal de los ríos y por condición de invierno se puede considerar un alto nivel de los embalses o un vertimiento de los mismos y/o un aporte significativo de los ríos a los embalses. Para estas condiciones, se espera unos precios altos de oferta en condiciones de sequía, precios bajos de oferta en condiciones de invierno y precios normales o típicos bajo condiciones normales de embalses y/o aporte de los ríos.

Ya habiendo caracterizado las condiciones en las cuales se puede encontrar el precio de oferta, se puede acercar de una manera más o menos clara, al comportamiento de dicha variable. A partir de la combinación de las condiciones estacionarias con las de demanda para los precios se encuentran nueve (9) combinaciones de los estados de la variable Precio o variable clase. Entre algunos de estos estados se tienen: Condición de invierno – demanda mínima, Condición de verano – demanda mediana, Condición normal – demanda máxima, entre otros.

De esta manera la selección de los grupos para la discretización se realiza escogiendo nueve (9) grupos para los distintos estados que puede tomar la variable clase según el comportamiento de la demanda y de las condiciones hidrológicas.

7.3.3.2 DESCRIPCIÓN DEL MÉTODO DE DISCRETIZACIÓN.

Con la conformación de los grupos, se pretende estructurar los datos de tal manera que aquellas observaciones con alguna característica en común estén en un mismo grupo. La primera pregunta es, ¿Cómo medir esa similitud entre los datos?, otra pregunta que cabría en este análisis previo sería, ¿cómo se conforman los grupos?

Para medir la similitud entre los datos existen diferentes técnicas de medida dentro del análisis cluster, entre estas están: medidas de correlación, medidas de distancia y medidas de asociación. Para las dos primeras, exigen datos con medidas, mientras la medida de asociación no exige datos con medidas. La medida por distancia es la más usada; para esta, la similitud entre las distancias indica semejanzas entre las observaciones; también, a medida que la distancia es mayor indica menor similitud entre los datos.

7.3.3.2.1 DISTANCIA EUCLÍDEA.

También conocida como Distancia Euclídea simple. Este tipo de medida es la más usada entre las medidas de distancia. La distancia euclídea es la longitud de la medida más corta entre dos puntos, que pueden ser unidos mediante una línea recta. Esta medida puede ser implementada en cualquier dimensión, aunque típicamente su concepción se centra en dos dimensiones. A manera de ejemplo, para dos puntos de dos dimensiones de coordenadas (X_1, Y_1) y (X_2, Y_2) la distancia euclídea es:

$$d = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

Ecuación 27. Distancia euclídea entre dos objetos con coordenadas (X_1, Y_1) y (X_2, Y_2)

El siguiente paso tras haber determinado las distancias o separaciones entre los datos es la creación de los grupos. Para realizar la construcción de dichos grupos, existen ciertos procedimientos denominados procedimientos jerárquicos, dentro de las cuales existen unas técnicas denominadas métodos de aglomeración, para las cuales, cada objeto conforma su propio grupo o estado, luego con las evaluaciones de las distancias o una metodología determinada dichos estados o elementos se unen para formar nuevos grupos.

Existen cinco algoritmos base para desarrollar la construcción de estos grupos, a partir de la medición de la distancia euclidiana, los cuales se describen brevemente a continuación:

- Encadenamiento simple: Distancia más corta entre cualquiera de los elementos de los grupos.

- Encadenamiento completo: Distancia más larga entre los elementos más lejanos de los grupos.

- Encadenamiento medio: Distancia media entre todos los elementos de dos grupos.

- Método de Ward: Minimización de la suma de los cuadrados de las distancias entre dos grupos para todos los elementos.

- Método del centroide: Distancia euclidiana entre los centroides de dos grupos.

El algoritmo escogido para la conformación de los grupos fue el de encadenamiento medio.

7.3.3.2.2 ENCADENAMIENTO MEDIO

Este método usa las medidas de la distancia euclídea como la información base para la construcción de grupos y consiste básicamente en la agrupación de elementos o grupos a partir de la medición de la distancia media de todos los individuos de un grupo con todos los elementos de otro grupo. Esta técnica depende de cada una de las distancias que conforman dicho grupo y tiende a combinar los grupos con variaciones reducidas dentro del mismo. A continuación se mostrará la explicación de la construcción de dicho algoritmo:

- Cada uno de las observaciones, en principio se considera individualmente un grupo. A partir de estas observaciones se toman las distancias medidas entre todos los grupos y se unen aquellos grupos o elementos cuya distancia media sea más cercana.

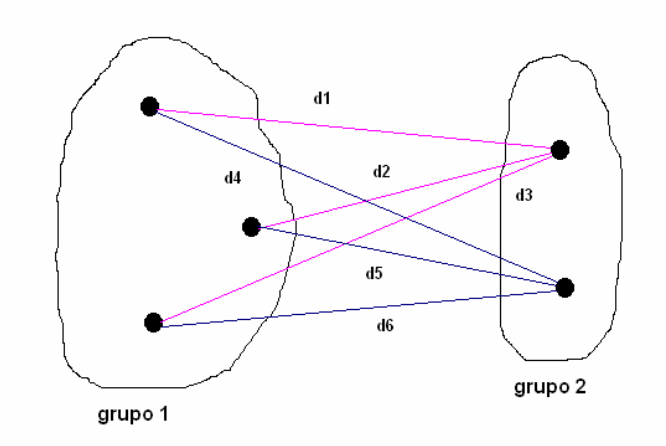
- Se evalúa si el número de grupos es el límite establecido para la conformación de los grupos, de no ser así nuevamente se toman las distancias medias entre todos los elementos de los grupos, buscando entre estos las longitudes medias mas cortas para la conformación de un nuevo grupo.

A manera de ejemplo, en la figura 20 se puede observar la medición de distancias de distancias euclidianas para dos grupos. Tras la medición de estas distancias se realiza el promedio entre estos 6 grupos en particular de la siguiente manera:

$$D_{promedio} = \frac{d1 + d2 + d3 + d4 + d5 + d6}{6}$$

Ecuación 28. Distancia media entre seis distancias distintas

Figura 20. Medición de distancias euclidianas para el método del encadenamiento medio para dos grupos



7.3.3.3 FORMAS DE DISCRETIZACIÓN USADAS.

Para analizar la sensibilidad de los precios de oferta de un generador es importante visualizar el contexto de la situación, en otras palabras, se pueden presentar dos casos al tratar de estudiar la sensibilidad de un precio. Una desde el punto de vista del monitoreo y la otra desde el punto de vista de la evaluación. Por lo tanto, una concepción integral del estudio de la sensibilidad de los precios de oferta de un generador requiere la integración del monitoreo y de la evaluación.

El monitoreo entendido como un proceso de seguimiento de lo actuado, es discretizado sin tener en cuenta los datos para la inferencia, es decir, solo se

conocen los datos de aprendizaje. Por esta razón se discretizan los datos del aprendizaje con el método explicado en la sección 7.3.2.2.2 y a partir de estos resultados se definen los límites de cada grupo. Como ejemplo observar la Tabla 9 donde se registran los resultados de discretizar la variable precio del generador Guavio usando solamente los datos del aprendizaje, es decir, desde el punto de vista del monitoreo.

Tabla 9. Ejemplo para observar el rango de los estados de una variable usando el método del monitoreo. (Realizado con el precio del generador Guavio).

Estado	Mínimo	Máximo	Número de Observaciones
1	370045	370045	1
2	381549	381549	1
3	29189	40449	201
4	14921	28680	191
5	58188	77692	87
6	40761	57870	358
7	1331	13305	251
8	133058	133058	1
9	191730	191730	2

Para discretizar un dato de inferencia simplemente se ubica en el rango adecuado obtenido del aprendizaje y se le asigna el grupo correspondiente. Se pueden presentar cuatro posibles casos que se describen a continuación con los datos del generador Guavio:

- Si un precio esta en alguno de los rangos: por ejemplo si se quiere discretizar el precio con valor 30200, de la Tabla anterior se observa que este valor se encuentra en el rango del estado o grupo 3, por lo tanto el precio 30200 forma parte del grupo 3.
- Si un precio tiene un valor más bajo que el mínimo valor de todos los rangos: por ejemplo para discretizar el precio 990, de la Tabla anterior se observa que 990 es menor que 1330.74514 que es el mínimo valor de los rangos. En este caso el precio 990 llega a formar parte de ese grupo con el mínimo valor, en este caso el estado 7.
- Si un precio tiene un valor más alto que el máximo valor de todos los rangos: por ejemplo para discretizar el precio 385000, de la Tabla anterior se observa que 385000 es mayor que 381549.436 que es el máximo valor de los rangos. En este caso el precio 385000 llega a formar parte de ese grupo con el máximo valor, en este caso el estado 2.
- Si un precio se encuentra entre dos rangos: por ejemplo para discretizar el precio 28900 que se encuentra entre los límites de los estados 3 y 4 el proceso es el siguiente: se calculan los precios promedios para los estados 3 y 4. El promedio del estado 3 es calculado con 201 datos y su valor es 35374 . El promedio del estado 4 es calculado con 191 datos y su valor es 23177. En este caso, el precio 28900 se le debe asignar el estado 4 porque es el que tiene el precio promedio más cercano al precio que se quiere discretizar.

La evaluación, por su parte, explica o se encarga de lo que ha pasado. Desde este punto de vista se realiza una discretización con el total de los datos (aprendizaje e inferencia), esta discretización también se realiza con el método descrito en la sección 7.3.2.2.2. A diferencia del método anterior no hay datos que queden sin discretizar debido a que la discretización se realizó con la totalidad de los datos, por lo tanto la discretización de datos queda totalmente elaborada.

Las formas de discretización anteriormente explicadas aplican para la variable precio en los métodos reclasificación continuos y para todas las variables en los métodos de clasificación discretos. En el desarrollo de este libro cuando se refiere al método de discretización desde el punto de vista de monitoreo se llamará método 2 y cuando se refiera al método de discretización desde el punto de vista de la evaluación se denota como método 1.

7.3.4 SELECCIÓN DE VARIABLES.

Una vez se adquiere toda la información referente a las variables consideradas como relevantes para la conformación del precio de un determinado generador, se debe proceder al uso de un criterio que permita decidir con cuantas y cuales de las variables se debe continuar en el proceso. Como idea para responder cuales variables son las que se deben utilizar, aparece el procedimiento descrito en los términos relacionados con Redes Bayesianas de la sección 5.5.5 y conocido como información mutua, ya que este método permite realizar un ranking de cuales variables arrojan mas información en relación con el precio. De este resultado se deduce que las primeras variables de este ranking, que son las que más se relacionan con el precio son las que se deben seleccionar, pero ahora surge el problema de en cual variable se debe cortar esta selección. Como solución a este problema aparece el método del codo [Molina, 2002], procedimiento que permite realizar un corte adecuado o corte más óptimo. A continuación se describe como se realiza este proceso para el generador Guatrón.

- Como primer paso para la aplicación del método del codo se requiere del ranking de variables para Guatrón. Este ranking se observa en la Tabla 10.

Tabla 10. Ranking de variables para el generador Guatrón.

Variable	Posición en el ranking	Información mutua
P1Dn	1	0.19377
P2Dn	2	0.17486
P2Dd	3	0.16609
P1Dd	4	0.12475
Contratos	5	0.092248
Embalse propio	6	0.064358
Embalse Sistema	7	0.057792
Contratos [MWh]	8	0.02453
Disponibilidad	9	0.023654
Contratos [$]	10	0.015259
P1Dx	11	0.014793
P2Dx	12	0.01405
Reconciliación +	13	0.013022
Reconciliación -	14	0.012318

Adaptado por los autores.

El resultado de aplicar la información mutua se nombra como w_i, donde el subíndice i denota la posición en el ranking. Una vez obtenidos los resultados de información mutua se procede al cálculo de la media y la varianza total estadística. De las variables que cumplan la condición de estar a dos desviaciones de la media se debe obtener la variable de corte. En este ejemplo la media es 0.0708 y la varianza es 0.0042 para obtener el siguiente rango de análisis 0.0624-0.0793.

- Calcular los s_i a partir de los w_i así:

$$s_i = w_i + w_{i-1} \qquad \text{Para i>2}$$

A partir de los s_i se definen los σ_i de la siguiente manera:

$$\sigma_i = \sum_{2}^{i} s_i \qquad \text{Para i>2}$$

- Finalmente lo que se busca es la variable que presente el K_i más alto, donde n es el número total de variables así:

$$K_i = 1 - \frac{\sigma_i}{\sigma_n} \cdot \frac{n-j}{n} \qquad \text{Para i>2}$$

Los resultados de este proceso se encuentran en la Tabla 11. La variable resaltada es la única que quedo establecida en el rango encontrado anteriormente, por lo tanto no hay necesidad de buscar el mayor K_i. Entonces la selección de variables queda definida con las siguientes variables: P1Dn, P2Dn, P2Dd, P1Dd, Contratos [$/MWh] y embalse propio.

Tabla 11. Resultados del proceso del método del codo.

Variable	Posición en el ranking	Información mutua	s_i	σ_i	K_i
P1Dn	1	0.19377	0	0	0
P2Dn	2	0.17486	0.36863	0.36863	0.82218
P2Dd	3	0.16609	0.34095	0.70958	0.68624
P1Dd	4	0.12475	0.29084	1.0004	0.59785
Contratos	5	0.092248	0.217	1.2174	0.55955
Embalse propio	**6**	**0.064358**	**0.15661**	**1.374**	**0.55813**
Embalse Sistema	7	0.057792	0.12215	1.4962	0.57899
Contratos [MWh]	8	0.02453	0.082321	1.5785	0.61928
Disponibilidad	9	0.023654	0.048184	1.6267	0.67305
Contratos [$]	10	0.015259	0.038913	1.6656	0.73218
P1Dx	11	0.014793	0.030052	1.6957	0.79551
P2Dx	12	0.01405	0.028843	1.7245	0.86136
Reconciliación +	13	0.013022	0.027072	1.7516	0.92959
Reconciliación -	14	0.012318	0.025339	1.7769	1

Adaptado por los autores.

En caso de que en este rango se encuentren dos o más variables se selecciona como variable de corte la que posea el mayor K_i.

7.4 APLICACIÓN DE LAS TÉCNICAS DE CLASIFICACIÓN Y SELECCIÓN DE LA MÁS ADECUADA.

Para cada generador seleccionado (hidráulico o térmico) se aplican los siguientes métodos de clasificación:

- Métodos de clasificación discretos: Naives Bayes discreto, TAN, Árboles de clasificación (con maxn[35] 300, 100, 50 y 5).

- Métodos de clasificación continuos: Naives Bayes continuo, Árboles de clasificación (con maxn 300, 100, 50 y 5).

Una vez aplicadas las técnicas de clasificación a un generador el siguiente paso es analizar cuál de estos métodos clasificatorios describe de una mejor manera el comportamiento del precio de oferta del generador.

7.5 MODELO GRÁFICO E INTERPRETACIÓN DE RESULTADOS.

Los modelos gráficos que se desarrollan para los generadores analizados son:

- Modelo gráfico Naives Bayes: este modelo gráfico se puede desarrollar para cualquier generador usando métodos discretos o continuos.

- Modelo gráfico TAN: este modelo gráfico se puede desarrollar para cualquier generador que emplee métodos discretos.

- Modelo gráfico para Árboles de clasificación: este modelo gráfico se desarrolla cuando los árboles obtenidos presentan un número de nodos

[35] El maxn es una variable de los árboles de clasificación que hace referencia al número máximo de observaciones permitidas en un nodo. Ver sección 5.2.1.

bajo (aproximadamente inferior a 25 nodos). Esto se debe a que la resolución de la grafica no es buena para un número de nodos superior al límite establecido.

La interpretación de resultados es independiente para cada generador y se basa en el análisis de los resultados de los diferentes métodos de clasificación.

8. RESULTADOS DEL MODELO DESCRIPTIVO

Para el análisis de los modelos gráficos a partir de los clasificadores Naives Bayes, TAN y los árboles de clasificación, se tomó un periodo de estudio de 3 años (2003, 2004 y 2005). El método de discretización usado para el análisis de estos datos fue el método 2 (monitoreo) mediante el cual se asignan los estados de los datos de inferencia de acuerdo a la discretización hecha previamente con los datos de aprendizaje[36].

Básicamente lo que se realizó durante el estudio de los años 2003, 2004 y 2005 usando como herramientas de análisis las redes bayesianas y los árboles de clasificación, es describir el comportamiento de los precios de oferta del conjunto de doce generadores seleccionados, identificando en cada uno de los agentes generadores, las variables y elementos más significativos en la formación de dichos precios.

8.1 MODELO NAIVES BAYES DISCRETO

El modelo Naives Bayes asume relaciones de independencia entre las variables en estudio; usando la probabilidad condicional como elemento de evaluación y asociación de un estado de la variable clase con un conjunto de variables predictoras. Esta expresión matemática se conoce como hipótesis MAP y se mostró anteriormente en las ecuaciones 10, 11 y 12.

[36] Mas detalles acerca de este método de discretización se ven en la sección 7.3.2.3

Para esta investigación no solo se asocia el estado más probable de la variable clase dado un conjunto de variables predictoras; también se evaluará para cada una de las variables predictoras cual es el estado más probable de cada una de ellas dado los estados de la variable clase.

Con estas dos probabilidades condicionales se realizó un producto, para analizar cual es el grado de asociación de un estado en particular de la variable clase con cada uno de los estados de las variables predictoras. Aquellos productos más significativos representan una fuerte relación entre un estado en particular de la variable clase y un estado también específico de alguna de las variables predictoras, estableciendo de esta manera la importancia de cada una de las variables predictoras en el establecimiento del precio en bolsa en los estados que puede tomar el precio de oferta del generador. Este elemento de asociación se explica con mas detalles en [Hernández, 2007]. La representación matemática de esta doble probabilidad se muestra en la ecuación 27.

$$C_{MAPDOBLE} = \max(P(A_j / C_k) * P(C_k / A_j))$$

Ecuación 29. Doble probabilidad condicional máxima entre los estados de la variable clase y cada uno de los estados de la variable predictora.

De la ecuación anterior se toman los máximos productos de cada uno de los estados de la variable clase (elementos *k-ésimos*) con cada una de las variables predictoras (elementos *j-ésimos*), asociando con el máximo producto obtenido, la variable que se destacó y que para el estudio será la mas importante en el establecimiento de este estado del precio de oferta del generador. En este mismo producto se pueden destacar también los segundos y terceros puestos de los productos asociados a un estado de la variable clase, tomando de cada uno de

estos las variables que se destacaron y que también tienen relación con el estado de la variable clase en particular[37].

A partir del producto de estas probabilidades se estableció un umbral, con el cual se escogieron los productos mas significativos en cada uno de los estados de la variable clase y en los cuales se asocian las 3 variables más significativas para dichos estados, previo cumplimiento de otras condiciones. Este umbral se estableció en 0.16. Con este umbral se establece un grado de asociación mínimo de 0.4 [38] para la variable clase dado la variable predictora y del mismo valor (0.4) para la variable predictora dada la variable clase; dicho producto da como resultado 0.16, umbral establecido. Sin embargo, no es suficiente con el establecimiento de un umbral ya que algunos de los estados asociados con la variable clase no representan significativamente un conjunto de elementos, debido al bajo número de observaciones asociadas con un estado. Así la última de las condiciones es el establecimiento de aquellos estados más significativos aquellos cuyo número de observaciones representen como mínimo un 5% del total. La evaluación de esta última condición se realizó con el clasificador Naives Bayes, sin la condición de doble probabilidad condicional conjunta, tomando como datos de aprendizaje y evaluación, los datos correspondientes a los años 2003, 2004 y 2005[39]. Un ejemplo del producto de las probabilidades condicionales para el generador Guavio se muestra en la tabla 12.

[37] El desarrollo del algoritmo de doble probabilidad conjunta se puede ver en el programa de Matlab NaivesDtres.m, el cual se encuentra en el anexo I.

[38] En probabilidad se establece un rango de 0 a 1 para asociar la probabilidad de un evento; siendo cero (0) una probabilidad nula y de uno (1) como la certeza de que dicho suceso siempre ocurra.

[39] El desarrollo del algoritmo de Naives Bayes y evaluación de aquellos estados mas significativos se encuentra en el programa de Matlab NaivesD.m, el cual se encuentra en el anexo J.

Tabla 12. Dobles probabilidades condicionales máximas para el generador Guavio por el método del monitoreo (método 2).

PRODUCTO	ESTADO VARIABLE PREDICTORA	ESTADO VARIABLE CLASE	VARIABLE
0.51247903	7	4	Precio1Dn
0.39797651	9	4	Precio2Dn
0.38672522	4	4	Precio2Dd
0.32249524	6	6	Precio1Dd
0.2779761	9	4	EmbalseGuavio
0.26566905	4	5	Precio1Dn
0.26191685	6	6	Precio2Dd
0.25348734	6	6	Precio1Dn
0.24982267	6	4	Contratos
0.23445818	4	4	Precio1Dd
0.22853279	4	5	Contratos
0.22314822	5	3	Precio1Dn
0.21853448	7	5	Precio2Dd
0.20616098	5	3	Precio1Dd
0.20148064	3	3	Precio2Dd
0.19541965	3	4	Embalseagregado

Otra de las condiciones para el establecimiento de los estados más significativos es el desempeño del clasificador Naives Bayes durante los años 2003, 2004 y 2005 (datos de aprendizaje) el cual debe ser mayor del 50% para los estados que se tendrán en cuenta en el análisis de sensibilidad de los precios. El aprendizaje y evaluación del desempeño de este clasificador se realizó con los datos

pertenecientes a los años 2003, 2004 y 2005. Un ejemplo de la construcción de este clasificador Naives Bayes se observa en la Tabla 13, en la cual se observan los grupos más significativos y los aciertos obtenidos por el clasificador.

Tabla 13. Nivel de acierto y cantidad de datos evaluados con el clasificador Naives Bayes para el generador Guavio.

ESTADO	**ACIERTOS**	**OBSERVACIONES**	**PORCENTAJE ACIERTOS CLASIFICADOR**	**PORCENTAJE DE DATOS DEL TOTAL**
1	0	1	0	0.09
2	0	1	0	0.09
3	105	201	52.23	18.41
4	156	191	81.68	17.49
5	76	87	87.36	7.97
6	250	357	70.03	32.69
7	50	251	19.92	22.99
8	0	1	0	0.09
9	0	2	0	0.18

En la Tabla 13 se puede ver como los estados 3, 4, 5 y 6 alcanzaron un nivel de acierto mayor al 50% así como que su representación es mayor al 5% del total de las observaciones. Estos estados entrarían previamente en el análisis de sensibilidad para el generador Guavio.

Las variables más significativas para este ejemplo con el generador Guavio se encuentran en la Tabla 14.

Tabla 14. Variables más significativas para el generador Guavio en los estados más importantes del generador Guavio.

GRUPO	PRIMERA VARIABLE	SEGUNDA VARIABLE	TERCERA VARIABLE
3	P1Dn	P1Dd	P2Dd
4	P1Dn	P2Dn	P2Dd
5	P1Dn	Contratos	P2Dd
6	P1Dd	P2Dd	P1Dn

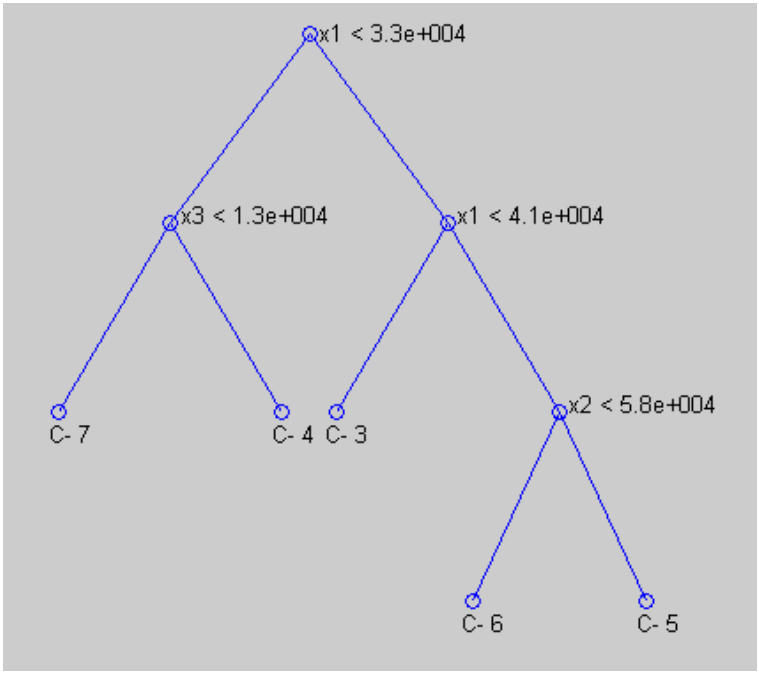
Así quedan establecidas las condiciones para la identificación de las variables más significativas, pertenecientes al conjunto de las variables predictoras con los estados mas significativos de la variable clase para cada uno de los generadores en estudio.

8.2 MODELO ÁRBOLES DE CLASIFICACIÓN CONTINUO

Con los árboles de clasificación se aproximan los estados de la variable clase de acuerdo al condicionamiento de las variables predictoras, representadas en decisiones dicotómicas[40] tomadas en cada uno de los nodos internos presentes en el árbol de clasificación. Mas detalles acerca de esta herramienta de clasificación se encuentran en la sección 6.2. De acuerdo a la desigualdad establecida en cada uno de los nodos internos los datos tomarán ciertas rutas o caminos establecidos en el árbol, hasta finalmente asociarlo a un estado de la variable clase. La primera variable con la cual se toma la primera decisión dentro del árbol de clasificación, denominado nodo raíz, es fundamental en la determinación de los estados de la variable clase. Un ejemplo de un árbol de clasificación se muestra en la figura 21.

[40] Divisiones de un elemento o conjunto de datos en dos partes.

Figura 21. Árbol de clasificación continuo con el método 2 para el generador Guavio, maxn =300.



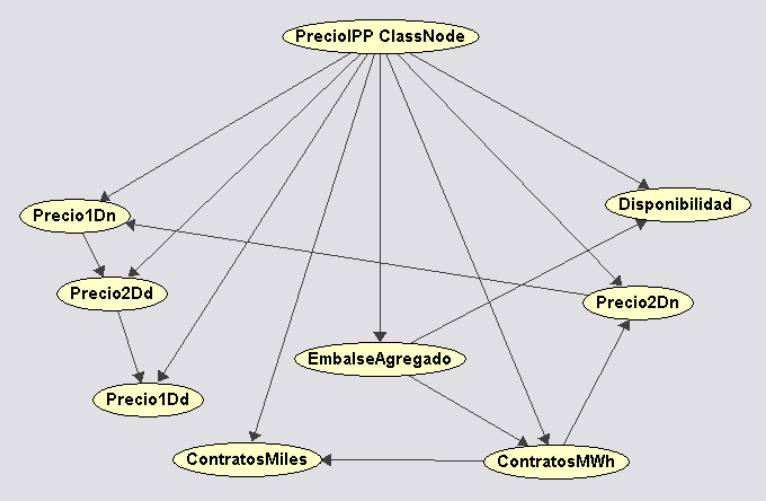
Para este ejemplo, las variables x_1, x_2 y x_3 representaron respectivamente las variables predictoras P1Dn, P2Dd y P2Dn. Estas, hicieron parte del conjunto de las más representativas para el modelo de Naives Bayes, las cuales se resumen en la Tabla 14. Así mismo, se observa como los estados presentes dentro del árbol de clasificación y representados en el árbol de clasificación como los nodos terminales, coincidieron con las variables más significativas obtenidas por el modelo Naives Bayes y resumidas en la Tabla 14.

8.3 MODELO TAN

Con el modelo TAN se encuentran las relaciones mas relevantes entre cada una de las variables predictoras, usando como algoritmo de construcción de dichas

relaciones, la *Información mútua condicionada*. Este modelo a diferencia del Naives Bayes no asume relaciones de independencia entre las variables; más detalles acerca de la construcción e inferencia de este clasificador bayesiano se pueden encontrar en la sección 6.1.5. Este modelo establece las relaciones existentes entre las variables predictoras y el sentido[41] en el que se encuentran. Un ejemplo acerca de un modelo clasificatorio TAN con el generador Chivor se encuentra en la figura 22.

Figura 22. Modelo clasificatorio TAN con el método 2 para el generador Chivor.



La importancia de este modelo, se centra en asumir las relaciones más importantes entre las variables previamente establecidas (método del codo), en función del precio de oferta del generador. Las relaciones más relevantes

[41] Se llama sentido en la construcción del modelo TAN a la dependencia entre las variables, representado gráficamente en el modelo mediante una flecha, para la cual la variable a la cual se dirige la flecha (hijo) depende de la variable de la cual parte la flecha (padre). Este sentido se ve expresado matemáticamente mediante una distribución de probabilidad.

obtenidas de este modelo, y establecidas como ejemplo con el generador Chivor, se encuentran en la Tabla 15.

Tabla 15. Relaciones establecidas en el modelo clasificatorio TAN para Chivor.

PADRE	**HIJO**
P1Dn	P2Dd
P2Dd	P1Dd
P2Dn	P1Dn
Embalse Agregado	Disponibilidad
Embalse Agregado	ContratosMWh
ContratosMiles	ContratosMWh
ContratosMWh	P2Dn

8.4 RESULTADOS POR GENERADOR

8.4.1 GENERADOR CHIVOR

Este generador fue analizado con 1078 datos de aprendizaje ubicados en el periodo del 01/01/2003 al 31/12/2005 y 181 datos de inferencia obtenidos de un rango de datos del 01/01/2006 al 30/06/2006.

Las variables predictoras escogidas para el generador Chivor usando el método del codo se encuentran en la tabla 16.

Tabla 16. Variables más significativas para el generador Chivor usando el método del codo.

PUESTO	VARIABLE
1	P1Dn
2	P2Dd
3	P2Dn
4	Embalse Agregado
5	P1Dd
6	Disponibilidad
7	Miles de pesos en contratos
8	Energía en Contratos

La distribución y los rangos de los estados de la variable clase se encuentran en la tabla 17.

Tabla 17. Rango de los estados del precio para Chivor con el método 2.

ESTADO	MÍNIMO [$/MWh]	MÁXIMO [$/MWh]	NÚMERO DE OBSERVACIONES
1	1311	7210	552
2	7815	19393.5	78
3	19605.3	28695	112
4	29688.3	41002.9	83
5	41806.6	59567.1	245
6	60362.2	82412.1	158
7	85400.1	94157.9	12
8	104662.7	115188.9	11
9	129199.9	140130.5	8

De esta tabla se puede observar una gran concentración de datos pertenecientes a los grupos 1, 5 y 6. Tres estados (7, 8 y 9) poseen un número bastante bajo de observaciones, respecto a los demás grupos.

MODELO NAIVES BAYES DISCRETO

Bajo el modelo Naives Bayes Discreto de doble probabilidad condicional, se obtuvieron las variables más significativas para los estados más representativos de los precios de oferta del generador Chivor, estas se observan en la tabla 18.

Tabla 18. Variables más importantes para los estados más representativos del generador Chivor bajo el modelo de doble probabilidad condicional.

ESTADO	**VARIABLE 1**	**PRODUCTO 1**	**VARIABLE 2**	**PRODUCTO 2**	**VARIABLE 3**	**PRODUCTO 3**
3	P2Dd	0.255	P1Dn	0.248	P2Dn	0.191
5	P1Dn	0.183	P1Dd	0.182	P2Dd	0.168
6	P1Dn	0.176				

Las variables más importantes según la representación de este modelo se dieron con las pertenecientes a la Curva de Demanda Residual, en particular aquellas relacionadas con las franjas de demanda mínima y mediana. Dentro de estas, se destacó particularmente P1Dn y en segundo lugar P2Dd.

MODELO ÁRBOLES DE CLASIFICACIÓN CONTINUO.

En este modelo se destacan las relaciones construidas por los árboles con maxn = 300 y 100. Dentro del árbol de clasificación con maxn = 300, mostrado en la figura 23, se destacan las variables x_6, x_1, x_2 y x_8 las cuales corresponden respectivamente con Disponibilidad, P1Dn, P2Dd y Energía en Contratos (Contratos MWh). En este árbol de clasificación los estados 3, 5 y 6 hicieron parte de los estados mas destacados bajo el modelo de Naives Bayes discreto. Otro estado que se presentó con frecuencia en el árbol de clasificación fue el estado 1, el cual junto con los estados 3, 5 y 6 hizo parte del árbol de clasificación, maxn = 300.

Figura 23. Árbol de clasificación continuo con el método 2 para el generador Chivor, maxn =300

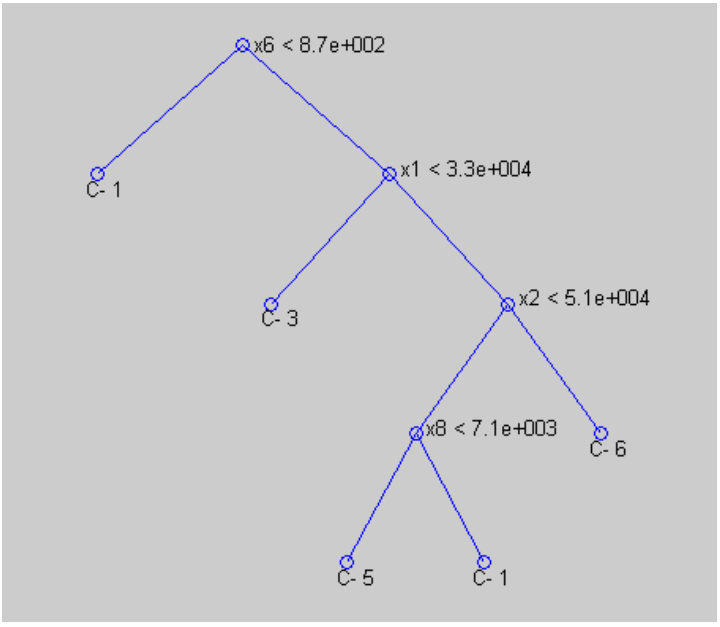
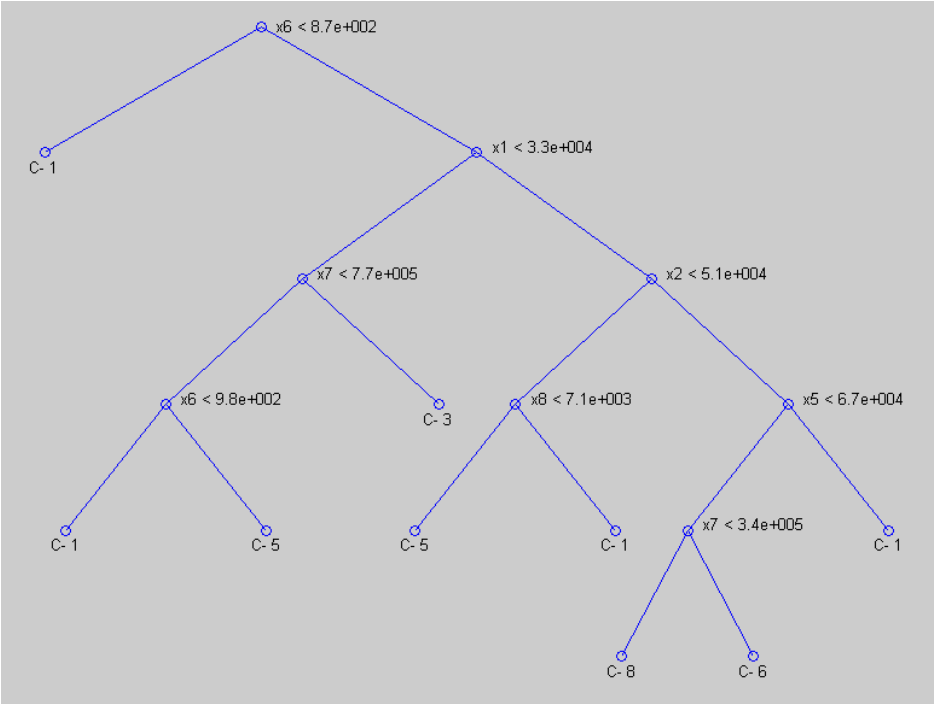


Figura 24. Árbol de clasificación continuo con el método 2 para el generador Chivor, maxn =100.

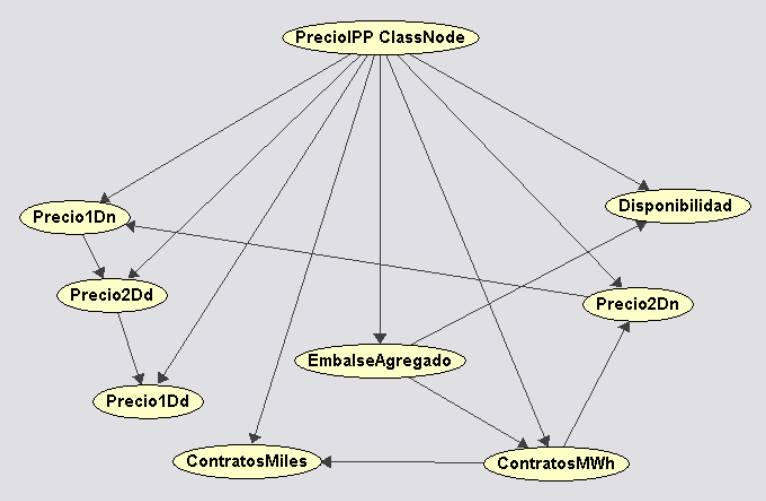


Del árbol de clasificación mostrado en la figura 24 (maxn =100), se puede observar como las relaciones establecidas a partir del árbol de clasificación con maxn = 300, mostrado en la figura 24, se construye el árbol de clasificación con maxn = 100, inclusive en los estados terminales asociados con la variable clase. Este árbol incluyó el estado 8 el cual no estaba presente dentro del árbol de clasificación con maxn = 300. Las variables disponibilidad y P2Dd sirvieron dos veces, como variables de decisión en el árbol de clasificación con maxn = 100.

MODELO TAN

Con el modelo TAN se establecen algunas de las relaciones más relevantes entre las variables predictoras. En la figura 25, se destacan dichas relaciones para el generador Chivor.

Figura 25. Modelo clasificatorio TAN con el método 2 para el generador Chivor.



Las relaciones establecidas para el modelo TAN del generador Chivor se muestran en la tabla 19.

Tabla 19. Relaciones establecidas en el modelo clasificatorio TAN para Chivor

PADRE	**HIJO**
Embalse Agregado	Disponibilidad
Embalse Agregado	Energía en Contratos
Energía en Contratos	Miles de pesos en contratos
Energía en Contratos	P2Dn
P2Dn	P1Dn
P1Dn	P2Dd
P2Dd	P1Dd

Entre las relaciones establecidas por el modelo clasificatorio TAN, la mas significativa entre las variables predictoras se dio entre las variables Energía en contratos y miles de pesos en contratos.

8.4.2 GENERADOR GUAVIO.

Este generador fue analizado con 1093 datos de aprendizaje ubicados en el periodo del 01/01/2003 al 31/12/2005 y 181 datos de inferencia obtenidos de un rango de datos del 01/01/2006 al 30/06/2006. Las variables predictoras escogidas para el generador Guavio usando el método del codo se encuentran en la tabla 20.

Tabla 20. Variables más significativas para el generador Guavio usando el método del codo.

PUESTO	**VARIABLE**
1	P1Dn
2	P2Dd
3	P2Dn
4	P1Dd
5	Precio en Contratos
6	Embalse Agregado
7	Embalse Guavio

La distribución y los rangos de los estados de la variable clase se encuentran en la tabla 21.

Tabla 21. Rango de los estados del precio para Guavio con el método 2.

ESTADO	MÍNIMO [$/MWh]	MÁXIMO [$/MWh]	NÚMERO DE OBSERVACIONES
7	1330.7	13357	349
4	14920.5	28679.9	207
3	29189.1	40448.9	226
6	40761.2	57869.8	381
5	57911.4	77692.5	106
8	133058	133058	1
9	191729.6	191729.6	2
1	370045.2	370045.2	1
2	381549.4	381549.4	1

En todos estados se puede destacar unos grupos con un número bastante pequeño de elementos como son los estados 8, 9, 1 y 2, los cuales suman solo 5 elementos. Así como otros estados como el 6, 7 y 5 con un número considerable de elementos.

MODELO NAIVES BAYES DISCRETO

Bajo el modelo Naives Bayes Discreto de doble probabilidad condicional, se obtuvieron las variables más significativas para los estados más representativos de los precios de oferta del generador Guavio, estas se observan en la tabla 22.

Tabla 22. Variables más importantes para los estados más representativos del generador Guavio bajo el modelo de doble probabilidad condicional.

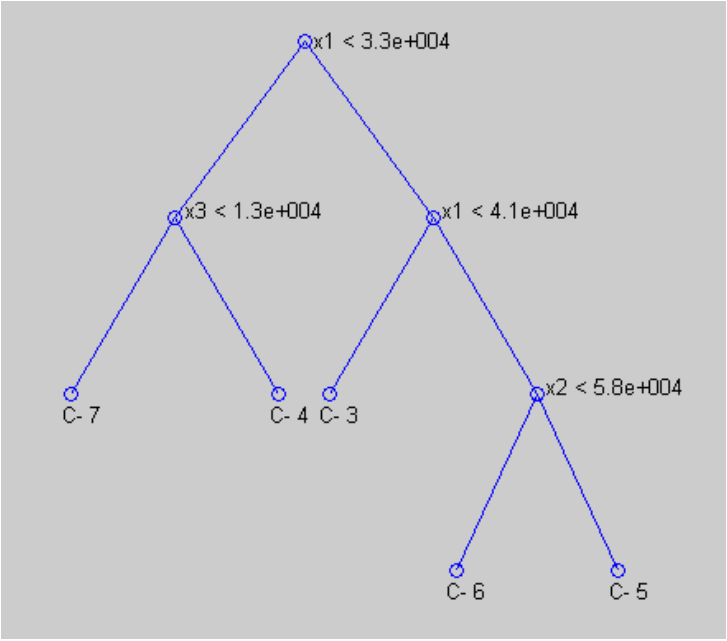
ESTADO	**VARIABLE 1**	**PRODUCTO 1**	**VARIABLE 2**	**PRODUCTO 2**	**VARIABLE 3**	**PRODUCTO 3**
3	P1Dn	0.223	P1Dd	0.206	P2Dd	0.201
4	P1Dn	0.512	P2Dn	0.398	P2Dd	0.387
5	P1Dn	0.266	Contratos	0.229	P2Dd	0.219
6	P1Dd	0.322	P2Dd	0.262	P1Dn	0.253

En la tabla anterior cabe destacar la importancia de las variables de la curva de demanda residual, particularmente de este conjunto se destacaron las variables P1Dn y P2Dd. Estos productos se destacaron de una manera significativa para el estado 4 de la variable clase.

MODELO ÁRBOLES DE CLASIFICACIÓN CONTINUO

En este modelo se destacan las relaciones construidas por los árboles con maxn = 300 y 100. Dentro del árbol de clasificación con maxn = 300, mostrado en la figura 26, las variables x_1, x_2 y x_3 sirvieron para tomar las decisiones en cada nodo interno y correspondieron respectivamente con P1Dn, P2Dd y P2Dn. En este árbol de clasificación se destacan los estados 3, 4, 5 y 6 que hicieron parte de los estados más destacados bajo el modelo de Naives Bayes discreto. También el estado 7 hizo parte de los estados de la variable clase del árbol de 300, siendo este uno de los estados más significativos en observaciones.

Figura 26. Árbol de clasificación continuo con el método 2 para el generador Guavio, maxn =300

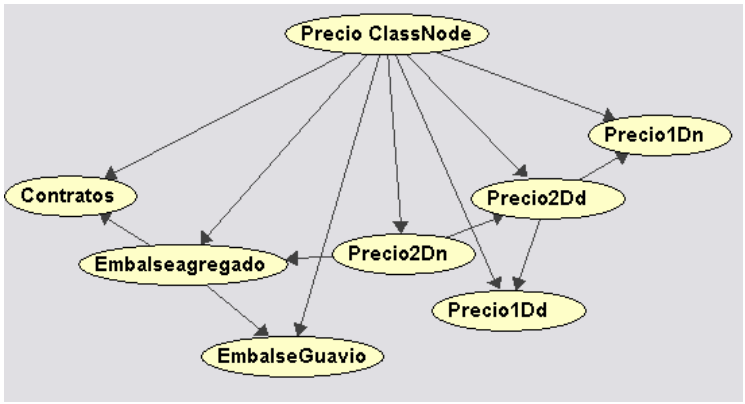


Debido al número de nodos que constituye el árbol de clasificación con maxn = 100 (25 nodos), del generador Guavio, no es fácil la visualización y el análisis de dicha estructura, por lo tanto no se incluyó.

MODELO TAN

Con el modelo TAN se establecen algunas de las relaciones más relevantes entre las variables predictoras. En la figura 27, se destacan dichas relaciones para el generador Guavio.

Figura 27. Modelo clasificatorio TAN con el método 2 para el generador Guavio.



Las relaciones establecidas para el modelo TAN del generador Guavio se muestran en la tabla 23.

Tabla 23. Relaciones establecidas en el modelo clasificatorio TAN para Guavio

PADRE	**HIJO**
P2Dn	P2Dd
P2Dn	Embalse Agregado
Embalse Agregado	Embalse Guavio
Embalse Guavio	Precio en Contratos
P2Dd	P1Dd
P2Dd	P1Dn

Entre las relaciones establecidas por el modelo clasificatorio TAN entre las variables predictoras incluidas en el generador Guavio, la más importante fue entre las variables embalse agregado y embalse propio.

8.4.3 GENERADOR GUATRÓN

Este generador fue analizado con 1089 datos de aprendizaje ubicados en el periodo del 01/01/2003 al 31/12/2005 y 178 datos de inferencia obtenidos de un rango de datos del 01/01/2006 al 30/06/2006. Las variables predictoras escogidas para el generador Guatrón usando el método del codo se encuentran en la tabla 24.

Tabla 24. Variables más significativas para el generador Guatrón usando el método del codo.

PUESTO	**VARIABLE**
1	P1Dn
2	P2Dn
3	P2Dd
4	P1Dd
5	Precio en Contratos
6	Embalse Propio

La distribución y los rangos de los estados de la variable clase se encuentran en la tabla 25.

Tabla 25. Rango de los estados del precio para Guatrón con el método 2.

ESTADO	MÍNIMO [$/MWh]	MÁXIMO [$/MWh]	NÚMERO DE OBSERVACIONES
8	5323.1	27848.7	374
5	28328.6	51548.2	599
6	51643.4	76687.2	268
7	77480.5	92328.5	11
2	95832.5	109168.2	8
1	120452.4	120452.4	3
9	139360	139360	1
3	161627.7	161627.7	1
4	181317	181317	2

De la anterior tabla, se observa una gran concentración de datos pertenecientes al grupo 5 (casi la mitad de los datos 47.28%). Los estados 8 y 6 presentan una concentración de datos significativa. Los demás estados poseen un número bajo de observaciones, respecto a los demás grupos.

MODELO NAIVES BAYES DISCRETO

Bajo el modelo Naives Bayes Discreto de doble probabilidad condicional, se obtuvieron las variables más significativas para los estados más representativos de los precios de oferta del generador Guatrón, estas se observan en la tabla 26.

Tabla 26. Variables más importantes para los estados más representativos del generador Guatrón bajo el modelo de doble probabilidad condicional.

ESTADO	**VARIABLE 1**	**PRODUCTO 1**	**VARIABLE 2**	**PRODUCTO 2**	**VARIABLE 3**	**PRODUCTO 3**
5	P2Dd	0.432	P1Dn	0.398	P2Dn	0.358
6	P1Dn	0.239	P1Dd	0.161	Embalse Propio	0.161
8	P1Dn	0.470	P2Dn	0.432	P1Dd	0.359

En la tabla anterior cabe destacar la importancia de las variables de la curva de demanda residual, particularmente de este conjunto se destacaron las variables P1Dn, P2Dn y P2Dd. Los productos más significativos se dieron en los estados 5 y 8 de la variable clase.

MODELO ÁRBOLES DE CLASIFICACIÓN CONTINUO

En este modelo se destacan las relaciones construidas por los árboles con maxn = 300 y 100. Dentro del árbol de clasificación con maxn = 300, mostrado en la figura 28, las variables x_1, sirvió para tomar las decisiones en cada nodo interno y correspondió respectivamente con P1Dn. En este árbol de clasificación se destacan los estados 5, 6 y 8 que hicieron parte de los estados más destacados bajo el modelo de Naives Bayes discreto.

Figura 28. Árbol de clasificación continuo con el método 2 para el generador Guatrón, maxn =300

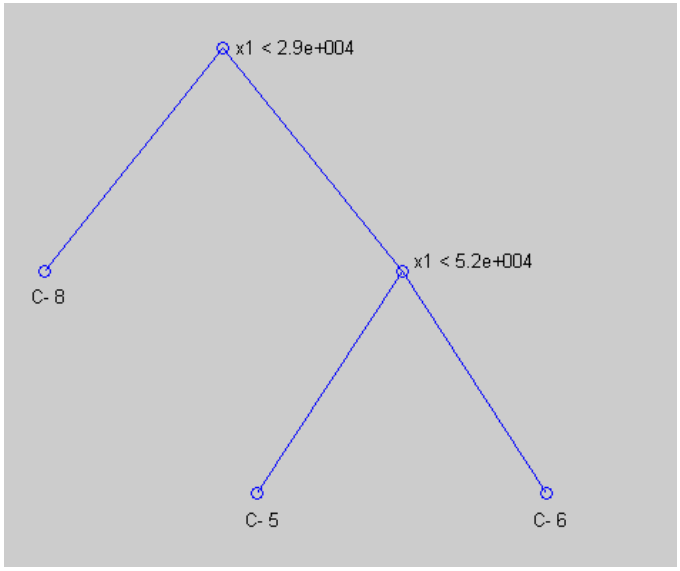
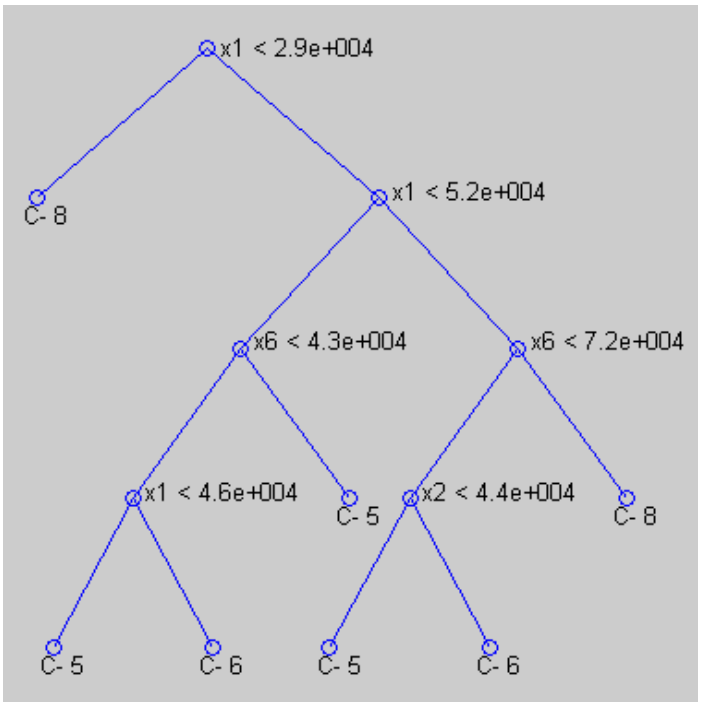


Figura 29. Árbol de clasificación continuo con el método 2 para el generador Guatrón, maxn = 100

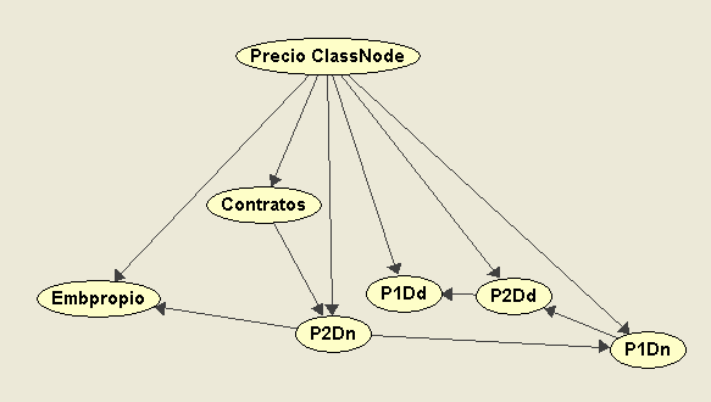


De la figura 29 se pueden destacar las variables x_1 (3 veces), x_6 (2 veces) y x_2 (1 vez), las cuales correspondieron respectivamente con P1Dn, Embalse Propio y P2Dn. Estas variables correspondieron con aquellas que mas se destacaron en el modelo Naives Bayes discreto.

MODELO TAN

Con el modelo TAN se establecen algunas de las relaciones más relevantes entre las variables predictoras. En la figura 30, se destacan dichas relaciones para el generador Guatrón.

Figura 30. Modelo clasificatorio TAN con el método 2 para el generador Guatrón.



Las relaciones establecidas para el modelo TAN del generador Guatrón se muestran en la tabla 27.

Tabla 27. Relaciones establecidas en el modelo clasificatorio TAN para Guatrón

PADRE	**HIJO**
Contratos	P2Dn
P2Dn	Embalse Propio
P2Dn	P1Dn
P1Dn	P2Dd
P2Dd	P1Dd

Entre las relaciones establecidas por el modelo clasificatorio TAN entre las variables predictoras incluidas en el generador Guatrón, la más importante fue entre las variables P2Dn y P1Dn.

8.4.4 GENERADOR SAN CARLOS

Este generador fue analizado con 1092 datos de aprendizaje ubicados en el periodo del 01/01/2003 al 31/12/2005 y 181 datos de inferencia obtenidos de un rango de datos del 01/01/2006 al 30/06/2006. Las variables predictoras escogidas para el generador San Carlos usando el método del codo se encuentran en la tabla 28.

Tabla 28. Variables más significativas para el generador San Carlos usando el método del codo.

PUESTO	VARIABLE
1	P1Dn
2	P2Dd
3	P2Dn
4	P1Dd
5	Contratos

La distribución y los rangos de los estados de la variable clase se encuentran en la tabla 29.

Tabla 29. Rango de los estados del precio para San Carlos con el método 2.

ESTADO	MÍNIMO [$/MWh]	MÁXIMO [$/MWh]	NÚMERO DE OBSERVACIONES
7	1330.4	16590.3	135
6	17021.5	27873.1	215
2	28055.2	37161.3	206
1	37281	37281	291
5	46407.4	59856.3	297
4	59896.9	71001.2	88
3	71878.8	80832.1	23
9	83939.8	95832.5	15
8	105714.3	107069.8	3

Dentro de estos estados se destacan los estados 1, 2, 5 y 6 con un número considerable de observaciones respecto a los demás grupos conformados.

MODELO NAIVES BAYES DISCRETO

Bajo el modelo Naives Bayes Discreto de doble probabilidad condicional, se obtuvieron las variables más significativas para los estados más representativos de los precios de oferta del generador San Carlos, estas se observan en la tabla 30.

Tabla 30. Variables más importantes para los estados más representativos del generador San Carlos bajo el modelo de doble probabilidad condicional.

ESTADO	**VARIABLE 1**	**PRODUCTO 1**	**VARIABLE 2**	**PRODUCTO 2**	**VARIABLE 3**	**PRODUCTO 3**
1	P2Dd	0.347	P1Dn	0.189		
2	P1Dn	0.238	P2Dd	0.236	P1Dd	0.214
5	P1Dn	0.319	P1Dd	0.227	P2Dn	0.164
6	P1Dn	0.519	P2Dd	0.190	P1Dd	0.188

En la tabla anterior se destaca la importancia de la variable P1Dn dentro de los estados más significativos de la variable clase. Dentro de estos resultados se destaca el alcanzado en el estado 6, en el cual el producto obtenido fue bastante destacado.

MODELO ÁRBOLES DE CLASIFICACIÓN CONTINUO

En este modelo se destacan las relaciones construidas por los árboles con maxn = 300 y 100. Dentro del árbol de clasificación con maxn = 300, mostrado en la figura 31, las variables x_1 sirvió para tomar las decisiones en cada nodo interno y correspondió respectivamente con la variable P1Dn. En este árbol de clasificación los estados 1, 2, 5 y 6 hicieron parte de los estados más destacados bajo el modelo de Naives Bayes discreto y del árbol de clasificación continuo también. El estado 4, también hizo parte de los estados asociados al árbol de clasificación.

Figura 31. Árbol de clasificación continuo con el método 2 para el generador San Carlos, maxn =300

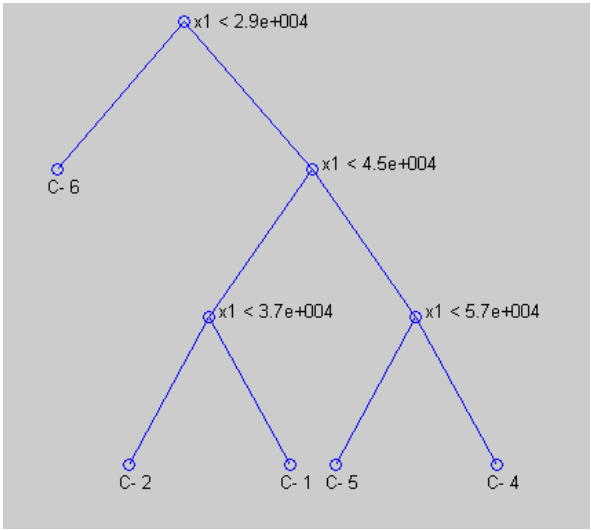
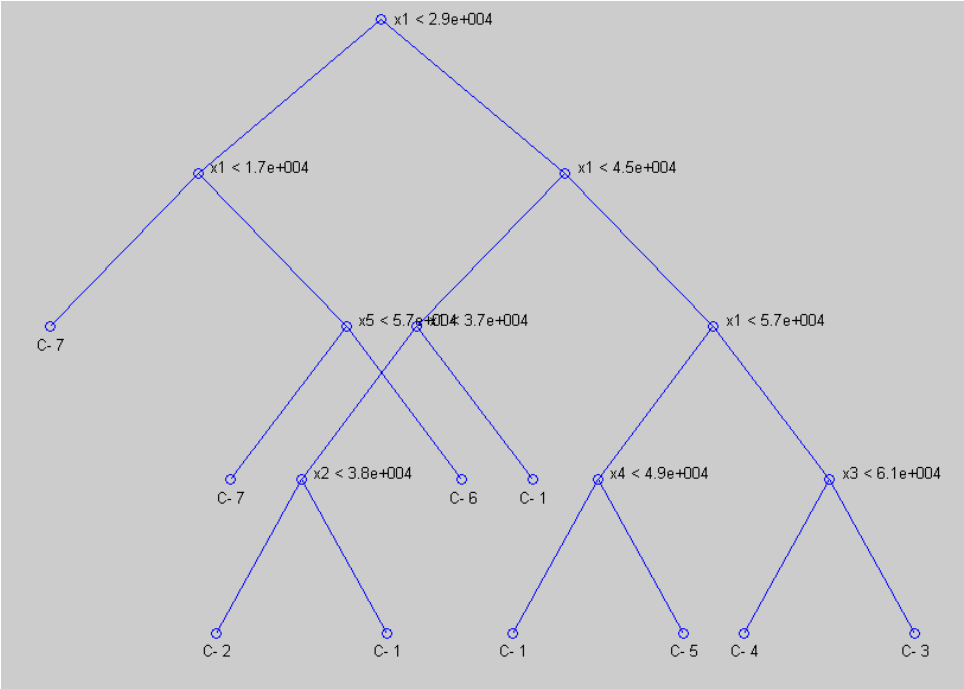


Figura 32. Árbol de clasificación continuo con el método 2 para el generador San Carlos, maxn =100

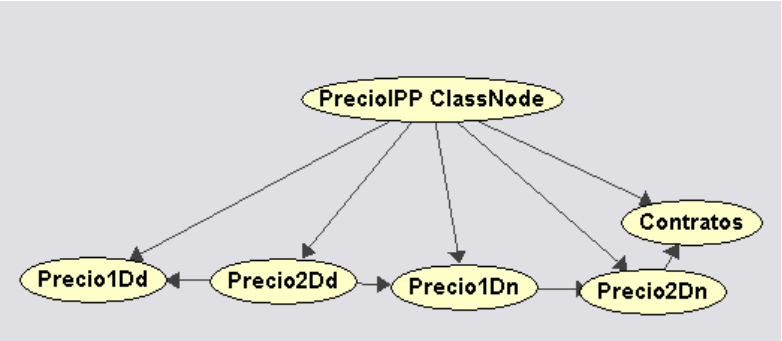


A partir de las relaciones establecidas en el árbol de maxn = 300 se construyó el de maxn = 100 con unos nodos internos más para la toma de decisiones asociadas a un estado de la variable clase. En este árbol de clasificación la variable P1Dn estuvo cuatro veces en la toma de decisiones en los nodos internos. Mientras las variables P2Dd, P2Dn, P1Dd y Contratos fueron cada uno de ellas y una sola vez variable de corte de los nodos internos.

MODELO TAN

Con el modelo TAN se establecen algunas de las relaciones más relevantes entre las variables predictoras. En la figura 33, se destacan dichas relaciones para el generador San Carlos.

Figura 33. Modelo clasificatorio TAN con el método 2 para el generador San Carlos.



Las relaciones establecidas para el modelo TAN del generador San Carlos se muestran en la tabla 31.

Tabla 31. Relaciones establecidas en el modelo clasificatorio TAN para San Carlos.

PADRE	**HIJO**
P2Dd	P1Dd
P2Dd	P1Dn
P1Dn	P2Dn
P2Dn	Contratos [$/MWh]

Entre las relaciones establecidas por el modelo clasificatorio TAN entre las variables predictoras incluidas en el generador Guavio, la más importante fue entre las variables P2Dd y P1Dn.

8.4.5 GENERADOR PORCE II

Este generador fue analizado con 1088 datos de aprendizaje ubicados en el periodo del 01/01/2003 al 31/12/2005 y 181 datos de inferencia obtenidos de un rango de datos entre el 01/01/2006 y 30/06/2006. Las variables predictoras escogidas para el generador Porce II usando el método del codo se encuentran en la tabla 32.

Tabla 32. Variables más significativas para el generador Porce II usando el método del codo.

PUESTO	VARIABLE
1	Disponibilidad
2	P1Dn
3	Contratos [$/MWh]
4	P2Dn
5	P2Dd
6	P1Dd
7	Embalse Agregado

La distribución y los rangos de los estados de la variable clase se encuentran en la tabla 33.

Tabla 33. Rango de los estados del precio para Porce II con el método 2.

ESTADO	MÍNIMO [$/MWh]	MÁXIMO [$/MWh]	NÚMERO DE OBSERVACIONES
5	1314.2	15242.5	345
2	17454.9	60912.3	713
1	61628.1	92978.2	154
4	94789.3	128870.8	38
3	140148.4	157177.1	4
8	191252.3	191252.3	7
9	270761.9	304835.9	5
7	361627.8	361627.8	2
6	407061.3	407061.3	1

MODELO NAIVES BAYES DISCRETO

Bajo el modelo Naives Bayes Discreto de doble probabilidad condicional, se obtuvieron las variables más significativas para los estados más representativos de los precios de oferta del generador Porce II, estas se observan en la tabla 34.

Tabla 34. Variables más importantes para los estados más representativos del generador Porce II bajo el modelo de doble probabilidad condicional

ESTADO	**VARIABLE 1**	**PRODUCTO 1**	**VARIABLE 2**	**PRODUCTO 2**	**VARIABLE 3**	**PRODUCTO 3**
2	Disponibilidad	0.544	P1Dn	0.285	P2Dd	0.250
5	Disponibilidad	0.324				

En la tabla anterior se destaca la importancia de las variables Disponibilidad, P1Dn y P2Dd en el establecimiento de la variable clase.

MODELO ÁRBOLES DE CLASIFICACIÓN CONTINUO

En este modelo se destacan las relaciones construidas por los árboles con maxn = 300 y 100. Dentro del árbol de clasificación con maxn = 300, mostrado en la figura 34, las variables x_1 y x_3 sirvieron para tomar las decisiones en cada nodo interno y correspondieron respectivamente con Disponibilidad y P2Dd. De este árbol de clasificación hicieron parte los estados 1, 2 y 5 de la variable clase. De estos los estados 2 y 5 se destacaron en los resultados obtenidos bajo el modelo de Naives Bayes discreto.

Figura 34. Árbol de clasificación continuo con el método 2 para el generador Porce II, maxn =300

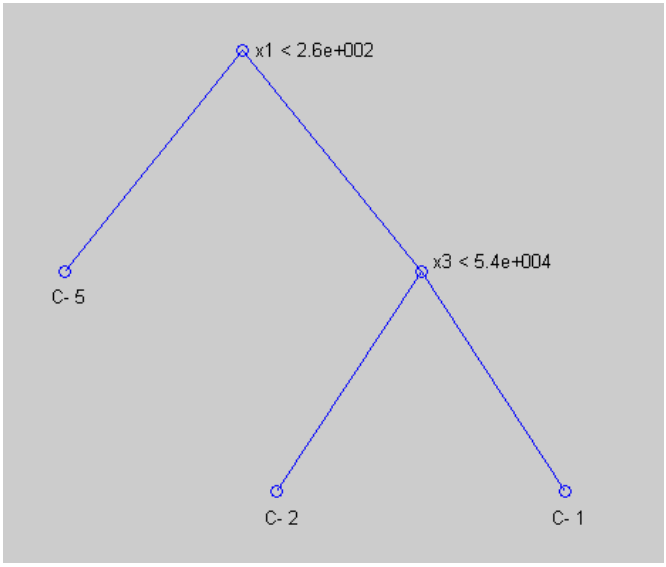
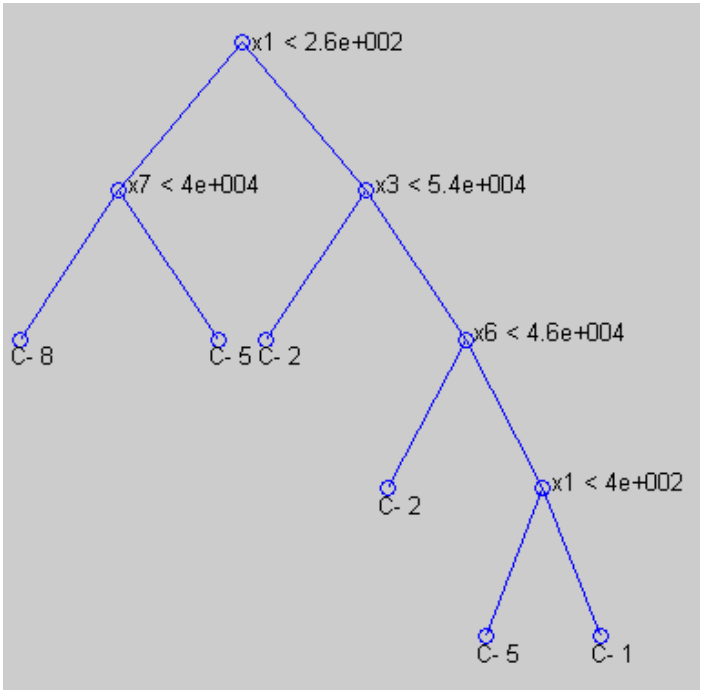


Figura 35. Árbol de clasificación continuo con el método 2 para el generador Porce II, maxn =100.

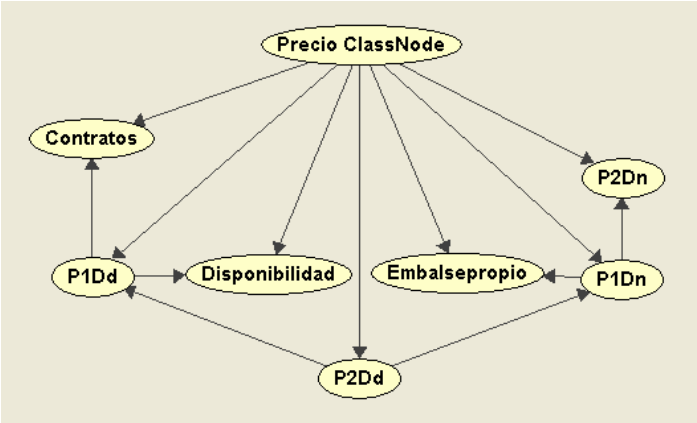


El árbol de clasificación con maxn = 100, mostrado en la figura 35 a diferencia del árbol de clasificación de 300 (figura 34), incluyó como variables de decisión el embalse propio y el precio de los contratos, (Contratos [$/MWh]).

MODELO TAN

Con el modelo TAN se establecen algunas de las relaciones más relevantes entre las variables predictoras. En la figura 36, se destacan dichas relaciones para el generador Porce II.

Figura 36. Modelo clasificatorio TAN con el método 2 para el generador Porce II.



Las relaciones establecidas para el modelo TAN del generador Porce II se muestran en la tabla 35.

Tabla 35. Relaciones establecidas en el modelo clasificatorio TAN para Porce II.

PADRE	**HIJO**
P2Dd	P1Dd
P1Dd	Contratos
P1Dd	Disponibilidad
P2Dd	P1Dn
P1Dn	Embalse propio
P1Dn	P2Dn

Entre las relaciones establecidas por el modelo clasificatorio TAN entre las variables predictoras incluidas en el generador Porce II, la más importante fue entre las variables P1Dn y P2Dn.

8.4.6 GENERADOR PARAISO-GUACA

Este generador fue analizado con 1092 datos de aprendizaje ubicados en el periodo del 01/01/2003 al 31/12/2005 y 181 datos de inferencia obtenidos de un rango de datos entre el 01/01/2006 y 30/06/2006. Las variables predictoras escogidas para el generador Pagua usando el método del codo se encuentran en la tabla 36.

Tabla 36. Variables más significativas para el generador Pagua usando el método del codo.

PUESTO	VARIABLE
1	Embalse Agregado
2	P2Dd
3	P1Dn
4	Embalse Propio
5	P1Dd
6	Contratos [$/MWh]
7	P2Dn
8	Disponibilidad

La distribución y los rangos de los estados de la variable clase se encuentran en la tabla 37.

Tabla 37. Rango de los estados del precio para Pagua con el método 2.

ESTADO	MÍNIMO [$/MWh]	MÁXIMO [$/MWh]	NÚMERO DE OBSERVACIONES
7	2364.3	23189.2	664
2	23404.9	42620.7	289
1	43133	60581.3	232
4	61046.3	76347.6	60
3	88184.4	97619.8	5
9	131052.1	171210	8
5	211210	247782.3	12
6	282614.1	282614.1	1
8	363888.5	363888.5	2

En estos estados se destacan los estados 7, 1 y 2 los cuales tuvieron un número considerable de observaciones.

MODELO NAIVES BAYES DISCRETO

Bajo el modelo Naives Bayes Discreto de doble probabilidad condicional, se obtuvieron las variables más significativas para los estados más representativos de los precios de oferta del generador Pagua, estas se observan en la tabla 38.

Tabla 38. Variables más importantes para los estados más representativos del generador Pagua bajo el modelo de doble probabilidad condicional

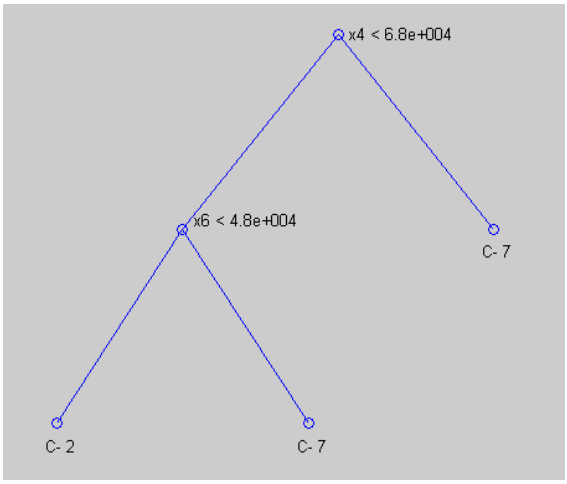
ESTADO	**VARIABLE 1**	**PRODUCTO 1**	**VARIABLE 2**	**PRODUCTO 2**	**VARIABLE 3**	**PRODUCTO 3**
1	P2Dd	0.208	P1Dd	0.195	Disponibilidad	0.194
2	Disponibilidad	0.271	P2Dd	0.161		
7	Embalse Propio	0.261	Disponibilidad	0.228	Contratos	0.206

En la tabla anterior se destaca la importancia de las variables Disponibilidad y P2Dd en el establecimiento de la variable clase.

MODELO ÁRBOLES DE CLASIFICACIÓN CONTINUO

En este modelo se destacan las relaciones construidas por los árboles con maxn = 300 y 100. Dentro del árbol de clasificación con maxn = 300, mostrado en la figura 37, las variables x_4 y x_6 sirvieron para tomar las decisiones en cada nodo interno y correspondieron respectivamente con Embalse propio y Precio en Contratos. Los estados 2 y 7 hicieron parte de este árbol de clasificación y del modelo naives bayes discreto.

Figura 37. Árbol de clasificación continuo con el método 2 para el generador Pagua, maxn =300.

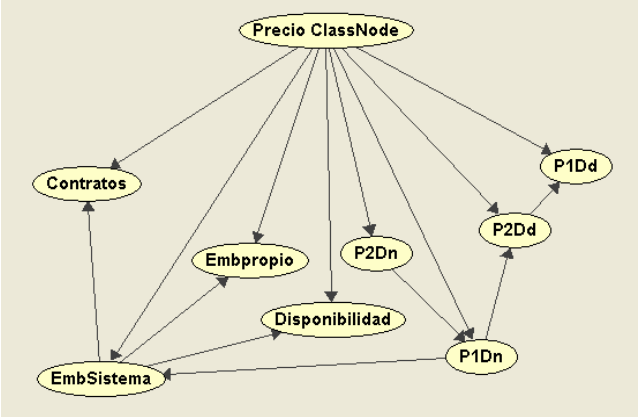


Debido al número de nodos que constituye el árbol de clasificación con maxn = 100 (20 nodos), del generador Pagua, no es fácil la visualización y el análisis de dicha estructura, por lo tanto no se incluyó.

MODELO TAN

Con el modelo TAN se establecen algunas de las relaciones más relevantes entre las variables predictoras. En la figura 38, se destacan dichas relaciones para el generador Pagua.

Figura 38. Modelo clasificatorio TAN con el método 2 para el generador Pagua.



Las relaciones establecidas para el modelo TAN del generador Pagua se muestran en la tabla 39.

Tabla 39. Relaciones establecidas en el modelo clasificatorio TAN para Pagua.

PADRE	**HIJO**
Embalse Agregado	Precio en Contratos
P1Dn	Embalse Agregado
Embalse Agregado	Embalse propio
P2Dn	P1Dn
P1Dn	P2Dd
P2Dd	P1Dd

Embalse Agregado	Disponibilidad

Entre las relaciones establecidas por el modelo clasificatorio TAN entre las variables predictoras incluidas en el generador Paraíso-Guaca, la más importante fue entre las variables P2Dn y P1Dn.

8.4.7 GENERADOR TEBSA

Este generador fue analizado con 1093 datos de aprendizaje ubicados en el periodo del 01/01/2003 al 31/12/2005 y 181 datos de inferencia obtenidos de un rango de datos del 01/01/2006 al 30/06/2006. Las variables predictoras escogidas para el generador Tebsa usando el método del codo se encuentran en la tabla 40.

Tabla 40. Variables más significativas para el generador Tebsa usando el método del codo.

PUESTO	**VARIABLE**
1	Energía en Contratos
2	Miles de pesos en contratos
3	P1Dn
4	P2Dd
5	Precio en contratos
6	P1Dd
7	P2Dn
8	Embalse Agregado
9	Reconciliación positiva

La distribución y los rangos de los estados de la variable clase se encuentran en la tabla 41.

Tabla 41. Rango de los estados del precio para Tebsa con el método 2.

ESTADO	MÍNIMO [$/MWh]	MÁXIMO [$/MWh]	NUMERO DE OBSERVACIONES
8	16512.3	19808.1	7
1	32805.4	42618.9	118
2	43767.5	60640.1	79
9	64300.3	71820.9	44
6	74271.8	97860.6	300
5	98115	112575.8	613
7	116683.2	131102.8	104
4	135754	145282.8	8
3	162329.2	162329.2	1

Para los rangos del precio de la tabla 41, se puede ver una concentración bastante acentuada de observaciones en el estado 5, seguido por el estado 6. Mientras existen otros estados con un número bastante pequeño de observaciones como el estado 8, 4 y 3.

MODELO NAIVES BAYES DISCRETO

Bajo el modelo Naives Bayes Discreto de doble probabilidad condicional, se obtuvieron las variables más significativas para los estados más representativos de los precios de oferta del generador Tebsa, estas se observan en la tabla 42.

Tabla 42. Variables más importantes para los estados más representativos del generador Tebsa bajo el modelo de doble probabilidad condicional

ESTADO	VARIABLE 1	PRODUCTO 1	VARIABLE 2	PRODUCTO 2	VARIABLE 3	PRODUCTO 3
1	RecPos	0.282	ContratosMWh	0.226	Embalse Agregado	0.186
5	Contratos	0.505	ContratosMiles	0.316	P2Dn	0.315
6	Contratos MWh	0.208	ContratosMiles	0.185		
7	Contratos MWh	0.187	Contratos	0.179		

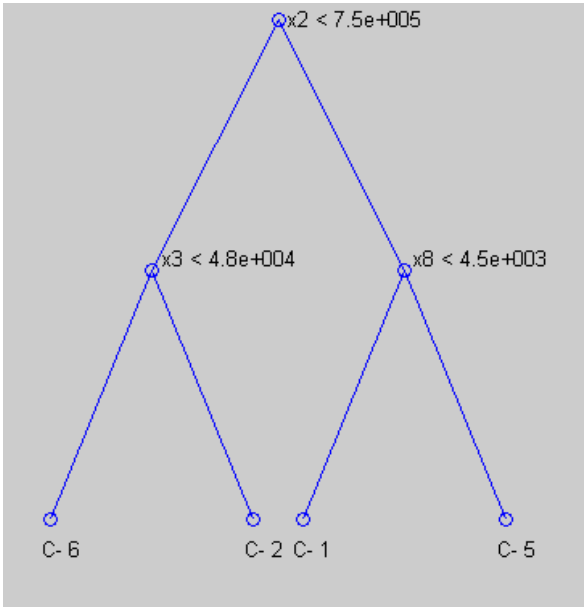
En la tabla anterior cabe destacar la importancia de las variables Energía en Contratos (ContratosMWh), Miles de pesos en contratos (ContratosMiles) y Precio en contratos (Contratos). Los productos asociados con el estado 5 se consideran bastante significativos.

MODELO ÁRBOLES DE CLASIFICACIÓN CONTINUO

En este modelo se destacan las relaciones construidas por los árboles con maxn = 300 y 100. Dentro del árbol de clasificación con maxn = 300, mostrado en la figura

39, las variables x_2, x_3 y x_8 sirvieron para tomar las decisiones en cada nodo interno y correspondieron respectivamente con Miles de pesos en contratos, P1Dn y Reconciliación positiva. En este árbol de clasificación se destacan los estados 1, 5 y 6 que hicieron parte de los estados más destacados bajo el modelo de Naives Bayes discreto y que también se encuentran dentro del árbol de clasificación continuo.

Figura 39. Árbol de clasificación continuo con el método 2 para el generador Tebsa, maxn =300

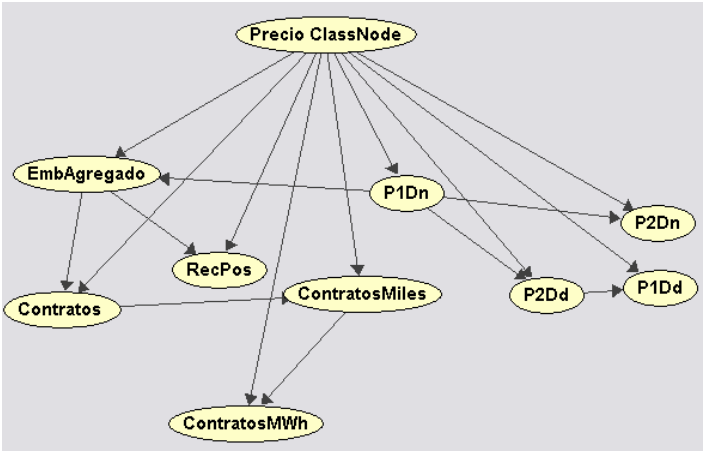


Debido al número de nodos que constituye el árbol de clasificación con maxn = 100 (25 nodos), del generador Tebsa, no es fácil la visualización y el análisis de dicha estructura, por lo tanto no se incluyó.

MODELO TAN

Con el modelo TAN se establecen algunas de las relaciones más relevantes entre las variables predictoras. En la figura 40, se destacan dichas relaciones para el generador Tebsa.

Figura 40. Modelo clasificatorio TAN con el método 2 para el generador Tebsa.



Las relaciones establecidas para el modelo TAN del generador Tebsa se muestran en la tabla 43.

Tabla 43. Relaciones establecidas en el modelo clasificatorio TAN para Tebsa

PADRE	**HIJO**
P1Dn	Embalse Agregado
P1Dn	P2Dn
Embalse Agregado	Contratos
Embalse Agregado	Reconciliación Positiva
Contratos	ContratosMiles

ContratosMiles	ContratosMWh
P1Dn	P2Dd
P2Dd	P1Dd

Entre las relaciones establecidas por el modelo clasificatorio TAN entre las variables predictoras incluidas en el generador Tebsa, la más importante fue entre las variables miles de pesos en contratos y energía en contratos.

8.4.8 GENERADOR TASAJERO

Este generador fue analizado con 1016 datos de aprendizaje ubicados en el periodo del 01/01/2003 al 31/12/2005 y 178 datos de inferencia obtenidos de un rango de datos del 01/01/2006 al 30/06/2006. Las variables predictoras escogidas para el generador Tasajero usando el método del codo se encuentran en la tabla 44.

Tabla 44. Variables más significativas para el generador Tasajero usando el método del codo.

PUESTO	**VARIABLE**
1	Precio en Contratos
2	Embalse Agregado
3	P2Dn
4	P1Dn

La distribución y los rangos de los estados de la variable clase se encuentran en la tabla 45.

Tabla 45. Rango de los estados del precio para Tasajero con el método 2.

ESTADO	MÍNIMO [$/MWh]	MÁXIMO [$/MWh]	NÚMERO DE OBSERVACIONES
5	31667.1	35675.8	11
6	37346.3	52309.8	449
1	52674.6	65506.1	442
2	66001.3	78232.6	131
9	97298.9	97298.9	1
8	139401.7	165284.6	108
7	168535.5	177165.2	16
4	181260.2	196224.1	27
3	198771.2	204652.3	9

MODELO NAIVES BAYES DISCRETO

Bajo el modelo Naives Bayes Discreto de doble probabilidad condicional, se obtuvieron las variables más significativas para los estados más representativos de los precios de oferta del generador Tasajero, estas se observan en la tabla 46.

Tabla 46. Variables más importantes para los estados más representativos del generador Tasajero bajo el modelo de doble probabilidad condicional

ESTADO	VARIABLE 1	PRODUCTO 1
6	Contratos	0.524

En la tabla anterior se destaca la importancia de las variables Precio de los contratos (Contratos) en el estado 6 de la variable clase. Esta fue la única variable que se destacó dentro de todos los posibles estados asociados con la variable clase

MODELO ÁRBOLES DE CLASIFICACIÓN CONTINUO

En este modelo se destacan las relaciones construidas por los árboles con maxn = 300 y 100. Dentro del árbol de clasificación con maxn = 300, mostrado en la figura 41, las variables x_1, x_2 y x_3 sirvieron para tomar las decisiones en cada nodo interno y correspondieron respectivamente con Precio de los contratos, Embalse Agregado y P2Dn. En este árbol de clasificación el estado 6, que hizo parte de los estados más destacados bajo el modelo de Naives Bayes discreto, también hizo parte del árbol de clasificación continuo. Así mismo la variable precio en Contratos formó parte del árbol en la toma de decisiones de dos de los nodos internos, incluyendo el nodo raíz.

Figura 41. Árbol de clasificación continuo con el método 2 para el generador Tasajero, maxn =300

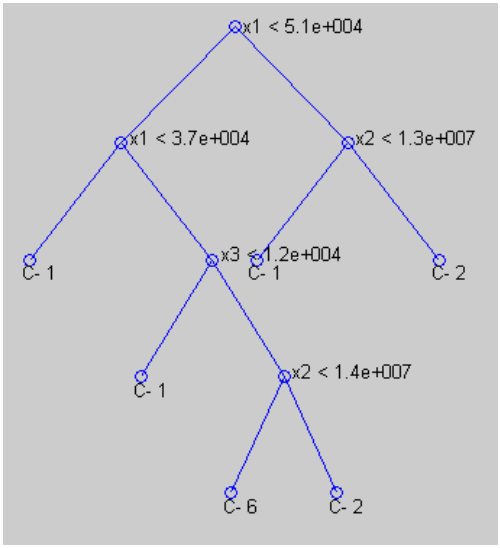
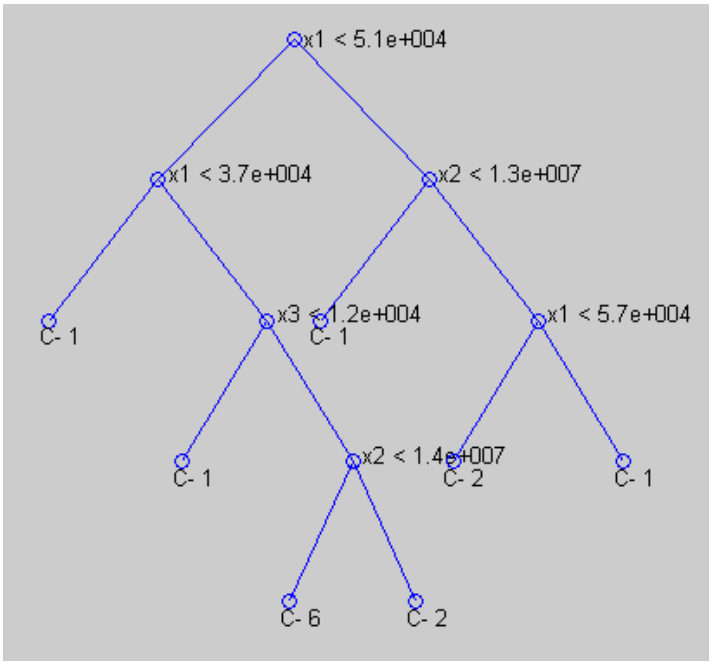


Figura 42. Árbol de clasificación continuo con el método 2 para el generador Tasajero, maxn =100.



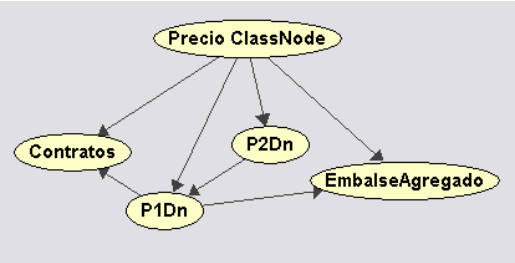
El árbol de clasificación contínuo de Tasajero con maxn = 100, incluyó un nodo interno mas comparándolo con el árbol de clasificación con maxn = 300, en donde se incluyó la variable precio en contratos en este. Dentro de los dos árboles de

clasificación se destacaron los estados 1, 2 y 6, los cuales concentraron la mayoría de los datos de aprendizaje y evidencia.

MODELO TAN

Con el modelo TAN se establecen algunas de las relaciones más relevantes entre las variables predictoras. En la figura 43, se destacan dichas relaciones para el generador Tasajero.

Figura 43. Modelo clasificatorio TAN con el método 2 para el generador Tasajero.



Las relaciones establecidas para el modelo TAN del generador Tasajero se muestran en la tabla 47.

Tabla 47. Relaciones establecidas en el modelo clasificatorio TAN para Tasajero.

PADRE	**HIJO**
P2Dn	P1Dn
P1Dn	Contratos

P1Dn	Embalse Agregado

Entre las relaciones establecidas por el modelo clasificatorio TAN entre las variables predictoras incluidas en el generador Tasajero, la más importante fue entre las variables P2Dn y P1Dn.

8.4.9 GENERADOR PAIPA IV

Este generador fue analizado con 958 datos de aprendizaje ubicados en el periodo del 01/01/2003 al 31/12/2005 y 181 datos de inferencia obtenidos de un rango de datos del 01/01/2006 al 30/06/2006. Las variables predictoras escogidas para el generador Paipa IV usando el método del codo se encuentran en la tabla 48.

Tabla 48. Variables más significativas para el generador Paipa IV usando el método del codo.

PUESTO	**VARIABLE**
1	P1Dn
2	P2Dn
3	P2Dd
4	P1Dd

5	Embalse Agregado

La distribución y los rangos de los estados de la variable clase se encuentran en la tabla 49.

Tabla 49. Rango de los estados del precio para Paipa IV con el método 2.

ESTADO	MÍNIMO [$/MWh]	MÁXIMO [$/MWh]	NÚMERO DE OBSERVACIONES
1	3738.2	21584.7	294
2	21846.2	56067	642
3	58062.7	66724	23
4	118917.2	122210	4
5	140524.8	178020.8	44
8	221022	253719	43
9	348167.3	356334.5	37
6	420804.2	420804.2	8
7	450231.8	480322.2	44

En estos estados se destacan los estados 1 y 2, los cuales tuvieron un número considerable de observaciones.

MODELO NAIVES BAYES DISCRETO

Bajo el modelo Naives Bayes Discreto de doble probabilidad condicional, se obtuvieron las variables más significativas para los estados más representativos de los precios de oferta del generador Paipa IV, estas se observan en la tabla 50.

Tabla 50. Variables más importantes para los estados más representativos del generador Paipa IV bajo el modelo de doble probabilidad condicional

ESTADO	**VARIABLE 1**	**PRODUCTO 1**	**VARIABLE 2**	**PRODUCTO 2**	**VARIABLE 3**	**PRODUCTO 3**
1	Embalse Agregado	0.199	P2Dd	0.163		
2	P1Dn	0.288	P2Dd	0.246	P2Dn	0.242

En la tabla anterior se destaca la importancia de las variables Embalse Agregado, P1Dn y P2Dd en el establecimiento de la variable clase.

MODELO ÁRBOLES DE CLASIFICACIÓN CONTINUO

En este modelo se destacan las relaciones construidas por los árboles con maxn = 300 y 100. Dentro del árbol de clasificación con maxn = 300, mostrado en la figura 44, las variables x_1 y x_4 sirvieron para tomar las decisiones en cada nodo interno y

correspondieron respectivamente con P2Dd y Embalse Agregado. Los estados 1 y 2 hicieron parte de este árbol de clasificación y del modelo naives bayes discreto.

Figura 44. Árbol de clasificación continuo con el método 2 para el generador Paipa IV, maxn =300

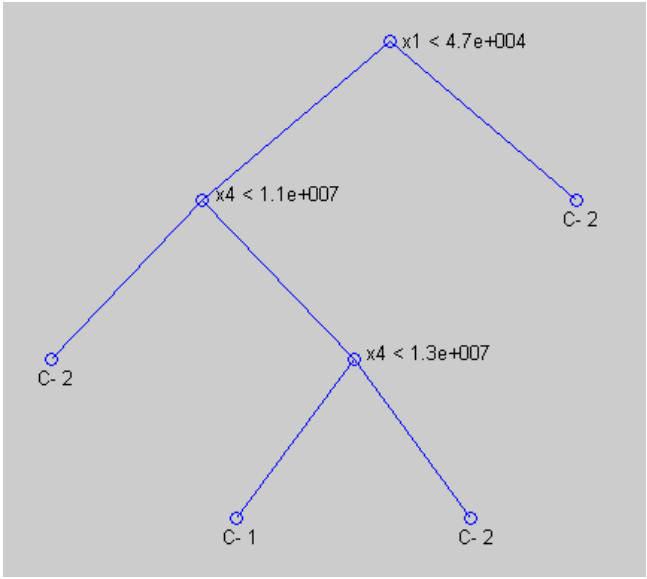
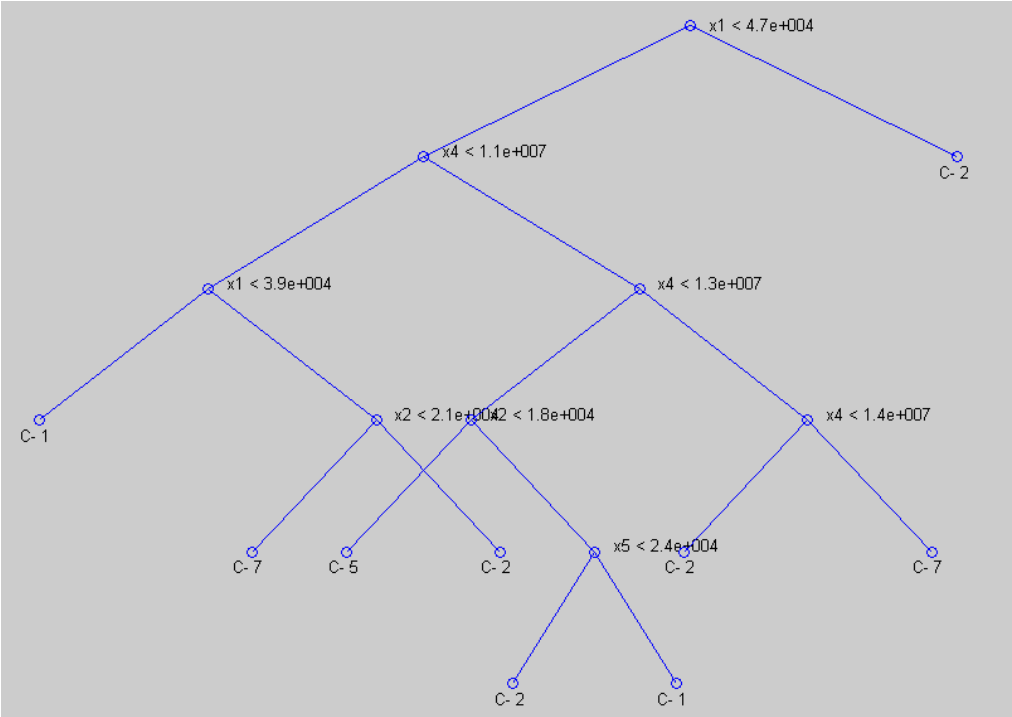


Figura 45. Árbol de clasificación continuo con el método 2 para el generador Paipa IV, maxn =300

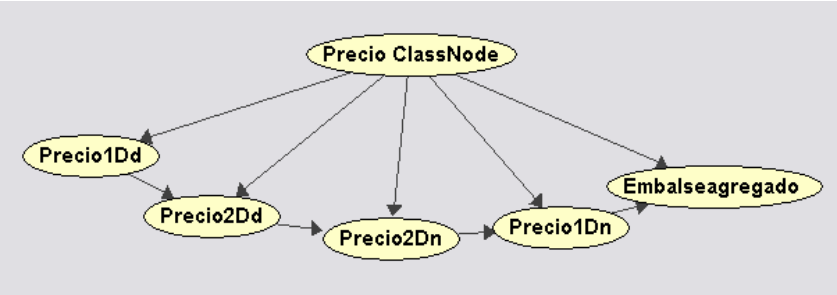


En este árbol de clasificación (maxn = 100) se observa como la variable Embalse Agregado (x_4), P1Dn (x_2) y P2Dd (x_1) estuvieron constantemente en las decisiones tomadas por el árbol de clasificación, presentándose respectivamente 4, 2 y 2 veces cada una de estas variables.

MODELO TAN

Con el modelo TAN se establecen algunas de las relaciones más relevantes entre las variables predictoras. En la figura 46, se destacan dichas relaciones para el generador Paipa IV.

Figura 46. Modelo clasificatorio TAN con el método 2 para el generador Paipa IV.



Las relaciones establecidas para el modelo TAN del generador Paipa IV se muestran en la tabla 51.

Tabla 51. Relaciones establecidas en el modelo clasificatorio TAN para Paipa IV.

PADRE	**HIJO**
P1Dd	P2Dd
P2Dd	P2Dn
P2Dn	P1Dn
P1Dn	Embalse Agregado

Entre las relaciones establecidas por el modelo clasificatorio TAN entre las variables predictoras incluidas en el generador Paipa IV, la más importante fue entre las variables P2Dn y P1Dn.

8.4.10 GENERADOR TERMOFLORES

Este generador fue analizado con 1078 datos de aprendizaje ubicados en el periodo del 01/01/2003 al 31/12/2005 y 153 datos de inferencia obtenidos de un rango de datos entre el 01/01/2006 y 30/06/2006. Las variables predictoras escogidas para el generador Termoflores usando el método del codo se encuentran en la tabla 52.

Tabla 52. Variables más significativas para el generador Termoflores usando el método del codo.

PUESTO	VARIABLE
1	Disponibilidad
2	ContratosMiles
3	ContratosMWh
4	Embalse Agregado
5	Precio Contratos
6	P2Dn
7	P1Dn

La distribución y los rangos de los estados de la variable clase se encuentran en la tabla 53.

Tabla 53. Rango de los estados del precio para Termoflores con el método 2.

ESTADO	MÍNIMO [$/MWh]	MÁXIMO [$/MWh]	NÚMERO DE OBSERVACIONES
8	32804.3	60957.9	127
4	65074.6	93311.5	237
3	94482.1	119014.2	425
1	125205.8	135639.4	197
2	138634	161834.2	124
7	178248.3	178248.3	3
9	283831.5	283832.5	2
6	460493.8	479422.1	31
5	484938	500992.4	85

Dentro de esta tabla se pueden destacar los grupos obtenidos en los estados 3 y 4, los cuales poseen un gran número de elementos comparados con los demás grupos.

MODELO NAIVES BAYES DISCRETO

Bajo el modelo Naives Bayes Discreto de doble probabilidad condicional, se obtuvieron las variables más significativas para los estados más representativos de los precios de oferta del generador Termoflores, estas se observan en la tabla 54.

Tabla 54. Variables más importantes para los estados más representativos del generador Termoflores bajo el modelo de doble probabilidad condicional

ESTADO	VARIABLE 1	PRODUCTO 1	VARIABLE 2	PRODUCTO 2	VARIABLE 3	PRODUCTO 3
3	Disponibilidad	0.338	Contratos	0.207	Energía en Contratos	0.166
4	Disponibilidad	0.450				

En la tabla anterior se destaca la importancia de las variables Disponibilidad y Precio en contratos en el establecimiento de la variable clase.

MODELO ÁRBOLES DE CLASIFICACIÓN CONTINUO

En este modelo se destacan las relaciones construidas por los árboles con maxn = 300 y 100. Dentro del árbol de clasificación con maxn = 300, mostrado en la figura 47, las variables x_1, x_3, x_4 y x_5 sirvieron para tomar las decisiones en cada nodo interno y correspondieron respectivamente con Disponibilidad (2 veces), Energía en Contratos, Embalse Agregado (2 veces) y Precio en Contratos. Los estados 2 y 7 hicieron parte de este árbol de clasificación y del modelo Naives Bayes discreto.

Figura 47. Árbol de clasificación continuo con el método 2 para el generador Termoflores, maxn =300.

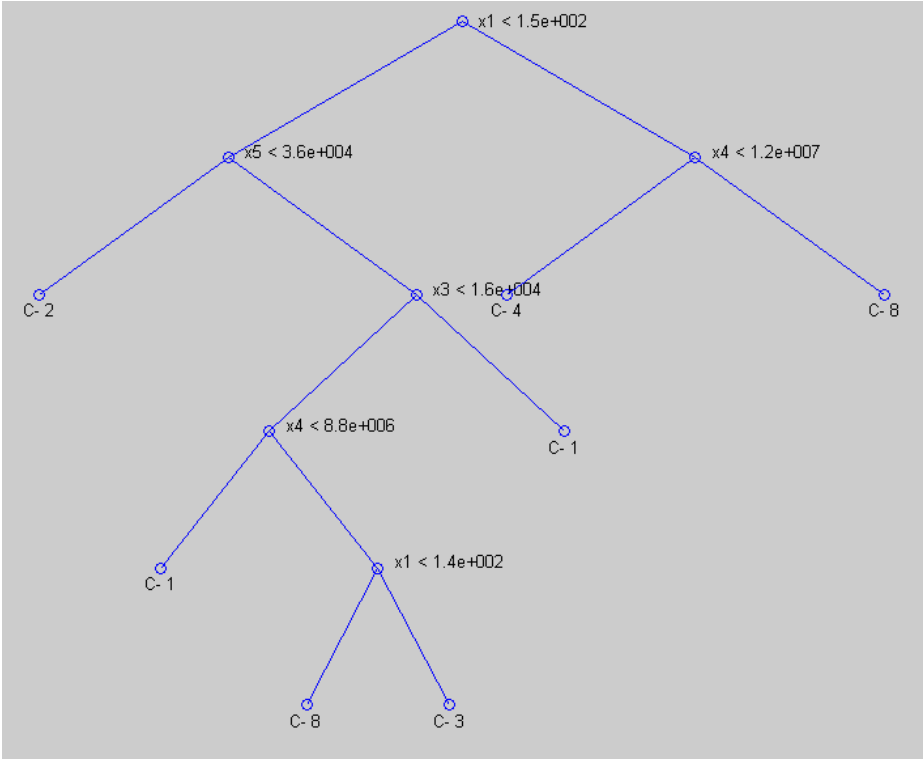
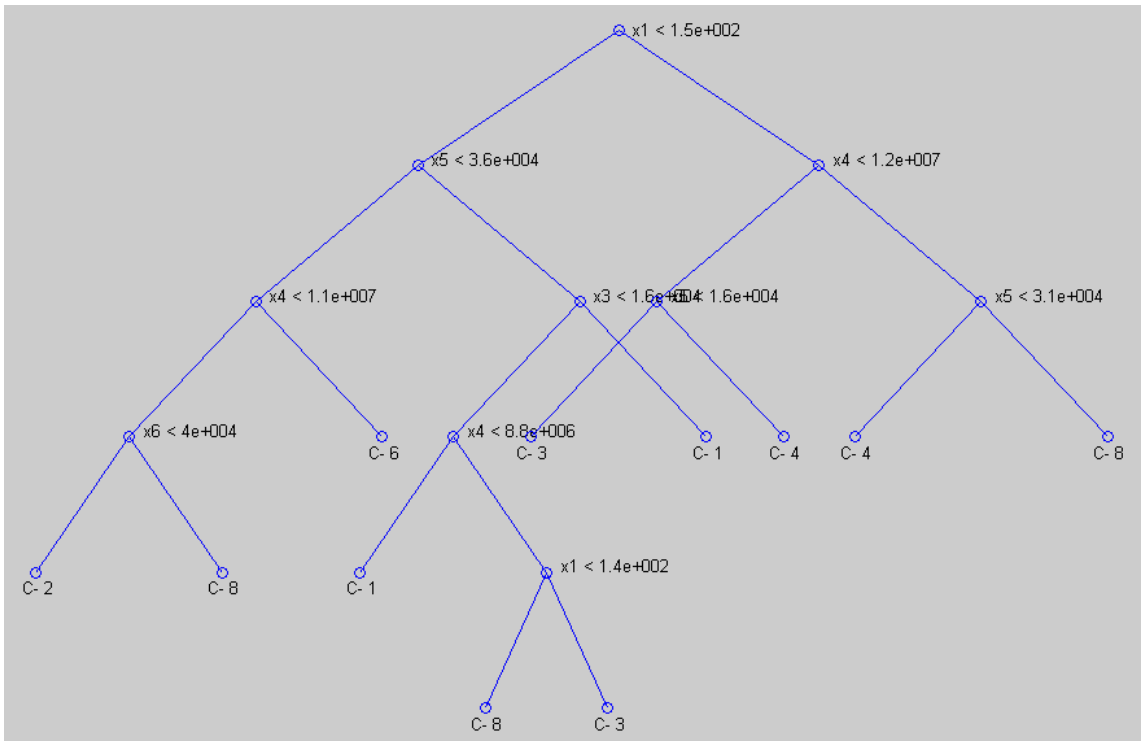


Figura 48. Árbol de clasificación continuo con el método 2 para el generador Termoflores, maxn =100.

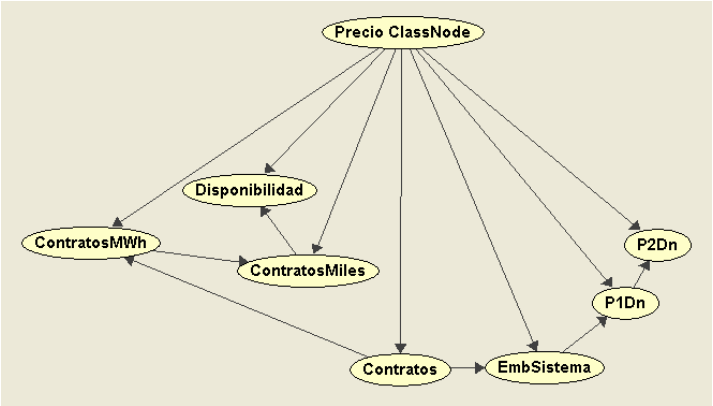


El árbol de clasificación con maxn = 100, parte de las relaciones establecidas por el árbol con maxn = 300, para la construcción del mismo. En este árbol se pueden destacar las relaciones establecidas con las variables x_4, x_5 y x_1 las cuales corresponden con las variables embalse agregado, precio en contratos y disponibilidad, y que estuvieron en los nodos 3, 3, y 2 veces respectivamente.

MODELO TAN

Con el modelo TAN se establecen algunas de las relaciones más relevantes entre las variables predictoras. En la figura 49, se destacan dichas relaciones para el generador Termoflores.

Figura 49. Modelo clasificatorio TAN con el método 2 para el generador Termoflores.



Las relaciones establecidas para el modelo TAN del generador Termoflores se muestran en la tabla 55.

Tabla 55. Relaciones establecidas en el modelo clasificatorio TAN para Termoflores.

PADRE	**HIJO**
Precio en contratos	Energía en contratos
Energía en contratos	Miles de pesos en contratos
Miles de pesos en contratos	Disponibilidad
Precio en contratos	Embalse Agregado
Embalse Agregado	P1Dn
P1Dn	P2Dn

Entre las relaciones establecidas por el modelo clasificatorio TAN entre las variables predictoras incluidas en el generador Termoflores, la más importante fue entre las variables P1Dn y P2Dn.

8.4.11 GENERADOR TERMOFLORES III

Este generador fue analizado con 1041 datos de aprendizaje ubicados en el periodo del 01/01/2003 al 31/12/2005 y 181 datos de inferencia obtenidos de un rango de datos entre el 01/01/2006 y 30/06/2006. Las variables predictoras escogidas para el generador Termoflores III usando el método del codo se encuentran en la tabla 56.

Tabla 56. Variables más significativas para el generador Termoflores III usando el método del codo.

PUESTO	VARIABLE
1	Reconciliación Positiva
2	Disponibilidad
3	Embalse Agregado
4	P2Dn
5	P1Dn

La distribución y los rangos de los estados de la variable clase se encuentran en la tabla 57.

Tabla 57. Rango de los estados del precio para Termoflores III con el método 2.

ESTADO	MÍNIMO [$/MWh]	MÁXIMO [$/MWh]	NÚMERO DE OBSERVACIONES
7	36033.4	66040.1	28
6	77052.5	108779.7	361
5	120242.4	141506.9	60
4	161210	186210.4	43
3	201243.9	233464	100
9	272827.7	285034.4	23
8	395720.2	395720.2	5
2	452139.6	482616.4	431
1	484938	500992.4	171

MODELO NAIVES BAYES DISCRETO

Bajo el modelo Naives Bayes Discreto de doble probabilidad condicional, se obtuvieron las variables más significativas para los estados más representativos de los precios de oferta del generador Termoflores III, estas se observan en la tabla 58.

Tabla 58. Variables más importantes para los estados más representativos del generador Termoflores III bajo el modelo de doble probabilidad condicional

ESTADO	**VARIABLE 1**	**PRODUCTO 1**	**VARIABLE 2**	**PRODUCTO 2**	**VARIABLE 3**	**PRODUCTO 3**
1	Disponibilidad					
2	RecPos		Disponibilidad		P1Dn	
6	RecPos		Disponibilidad		P2Dn	

En la tabla anterior se destaca la importancia de las variables Disponibilidad y Reconciliación Positiva en el establecimiento de la variable clase.

MODELO ÁRBOLES DE CLASIFICACIÓN CONTINUO

En este modelo se destacan las relaciones construidas por los árboles con maxn = 300 y 100. Dentro del árbol de clasificación con maxn = 300, mostrado en la figura 50, las variables x_1, x_2 y x_4 sirvieron para tomar las decisiones en cada nodo interno y correspondieron respectivamente con Reconciliación Positiva (2 veces), Disponibilidad (1 vez) y P2Dn (1 vez). Los estados 1, 2, 6 y 7 hicieron parte de este árbol de clasificación, así como los estados 1, 2 y 6 se destacaron dentro del modelo Naives Bayes discreto.

Figura 50. Árbol de clasificación continuo con el método 2 para el generador Termoflores III, maxn =300

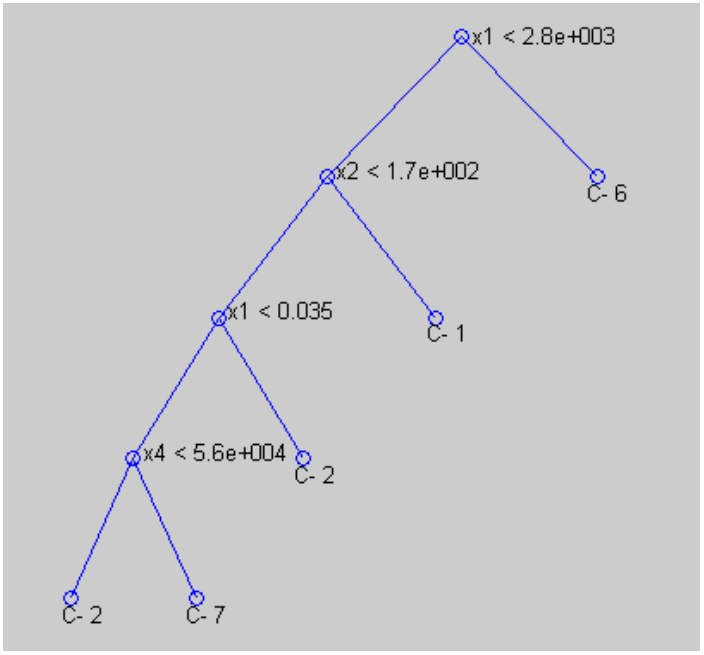
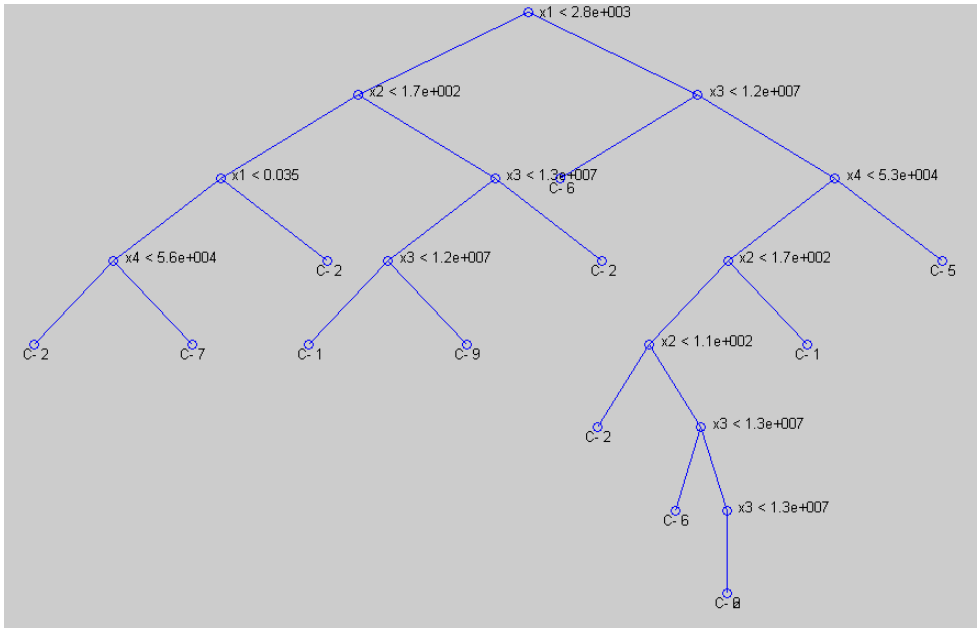


Figura 51. Árbol de clasificación continuo con el método 2 para el generador Termoflores III, maxn =100

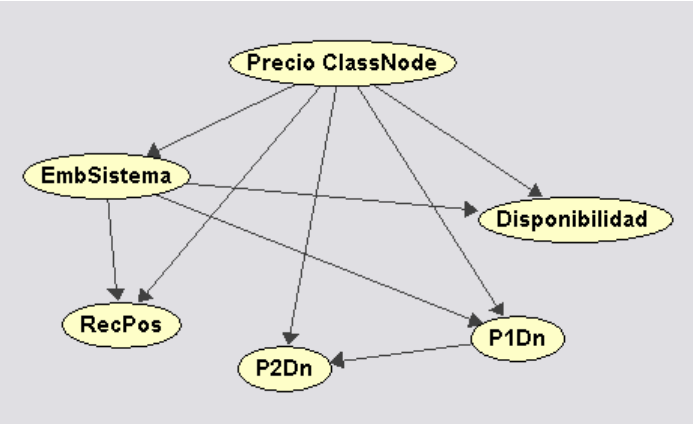


Con este árbol de clasificación se destacan las veces en que la variable x_3 (embalse agregado), x_2 (disponibilidad) y x_1 (reconciliación positiva) hicieron parte de las decisiones tomadas en los nodos internos del árbol de clasificación con maxn = 100.

MODELO TAN

Con el modelo TAN se establecen algunas de las relaciones más relevantes entre las variables predictoras. En la figura 52, se destacan dichas relaciones para el generador Termoflores III.

Figura 52. Modelo clasificatorio TAN con el método 2 para el generador Termoflores III.



Las relaciones establecidas para el modelo TAN del generador Termoflores se muestran en la tabla 59.

Tabla 59. Relaciones establecidas en el modelo clasificatorio TAN para Termoflores III.

PADRE	**HIJO**
Embalse Agregado	Disponibilidad
Embalse Agregado	Reconciliación Positiva
P1Dn	P2Dn
Embalse Agregado	P1Dn

Entre las relaciones establecidas por el modelo clasificatorio TAN entre las variables predictoras incluidas en el generador Termoflores III, la más importante fue entre las variables P1Dn y P2Dn.

8.4.12 GENERADOR TERMODORADA

Este generador fue analizado con 1091 datos de aprendizaje ubicados en el periodo del 01/01/2003 al 31/12/2005 y 181 datos de inferencia obtenidos de un rango de datos entre el 01/01/2006 y 30/06/2006. Las variables predictoras escogidas para el generador Termodorada usando el método del codo se encuentran en la tabla 60.

Tabla 60. Variables más significativas para el generador Termodorada usando el método del codo.

PUESTO	VARIABLE
1	Energía en contratos
2	Miles de pesos en contratos
3	Embalse Agregado
4	P1Dn

La distribución y los rangos de los estados de la variable clase se encuentran en la tabla 61.

Tabla 61. Rango de los estados del precio para Termodorada con el método 2.

ESTADO	MÍNIMO [$/MWh]	MÁXIMO [$/MWh]	NÚMERO DE OBSERVACIONES
4	38577.9	50809.9	76
3	55446.1	65935.4	13
7	76424.6	77692.5	46
1	98211.9	98676.3	18
2	108290.1	119730.5	271
6	131052.1	137363.8	91
5	150013.9	157893.2	531
8	285711.8	288887.9	82
9	471210	497027.3	144

MODELO NAIVES BAYES DISCRETO

Bajo el modelo Naives Bayes Discreto de doble probabilidad condicional, se obtuvieron las variables más significativas para los estados más representativos de los precios de oferta del generador Termodorada, estas se observan en la tabla 62.

Tabla 62. Variables más importantes para los estados más representativos del generador Termodorada bajo el modelo de doble probabilidad condicional

ESTADO	**VARIABLE 1**	**PRODUCTO 1**	**VARIABLE 2**	**PRODUCTO 2**	**VARIABLE 3**	**PRODUCTO 3**
2	Miles de pesos en contratos	0.40	Energía en contratos	0.290		
4	Miles de pesos en contratos	0.179	Energía en contratos	0.172		
5	Energía en contratos	0.407	Miles de pesos en contratos	0.397	P1Dn	0.205
6	Miles de pesos en contratos	0.612	Energía en contratos	0.581		
8	Energía en contratos	0.250	Miles de pesos en contratos	0.202		

En la tabla anterior se destaca la importancia de las variables Miles de pesos en contratos y Energía en contratos en el establecimiento de la variable clase.

MODELO ÁRBOLES DE CLASIFICACIÓN CONTINUO

En este modelo se destacan las relaciones construidas por los árboles con maxn = 300 y 100. Dentro del árbol de clasificación con maxn = 300, mostrado en la figura 53, las variables x_1 y x_2 sirvieron para tomar las decisiones en cada nodo interno y correspondieron respectivamente con Energía en contratos (2 veces) y Miles de pesos en contratos (1 vez). Los estados 2, 4, 5 y 6 hicieron parte de este árbol de clasificación. El análisis con el modelo Naives Bayes discreto bajo el modelo de doble probabilidad condicional incluyó los estados anteriormente nombrados, más el estado 8.

Figura 53. Árbol de clasificación continuo con el método 2 para el generador Termodorada, maxn =300

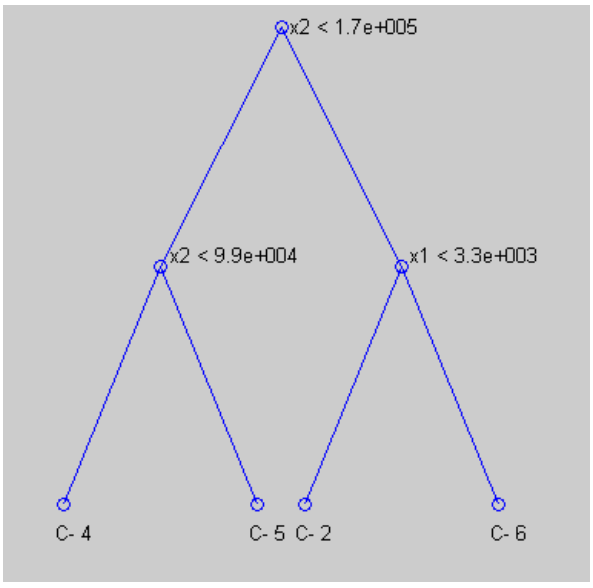
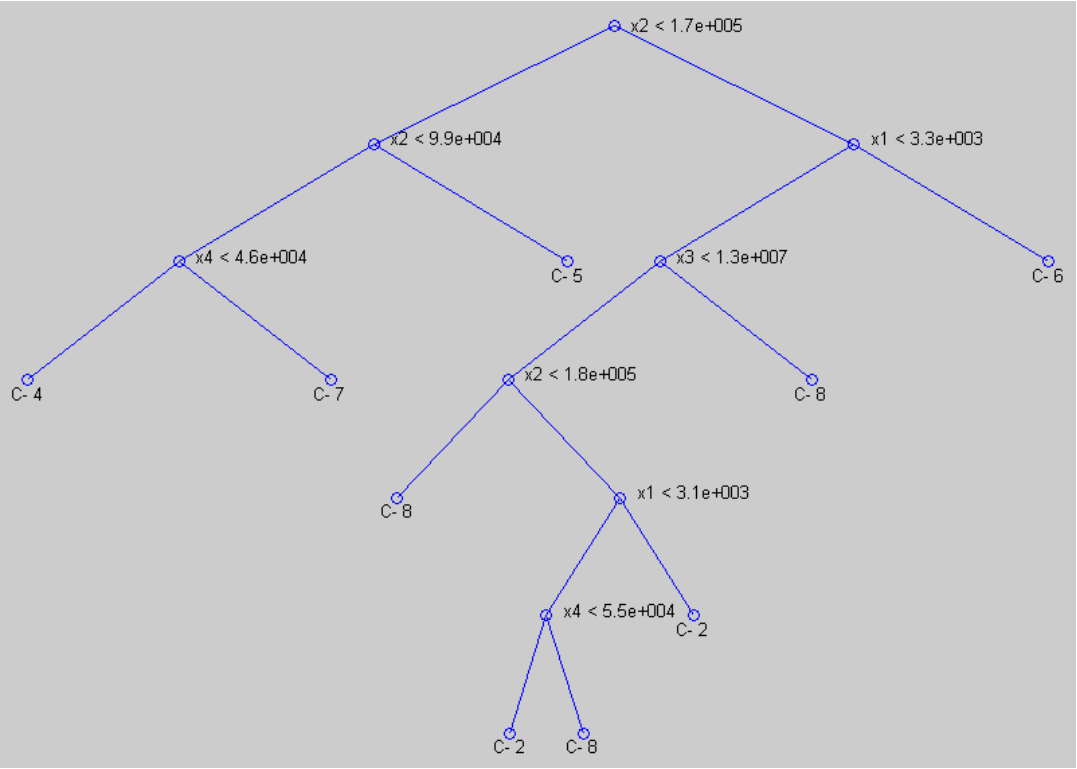


Figura 54. Árbol de clasificación continuo con el método 2 para el generador Termodorada, maxn = 100

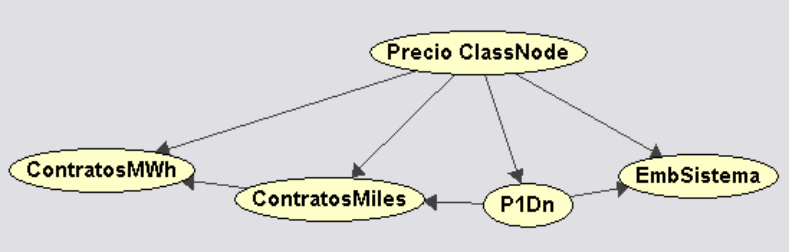


En el árbol de clasificación con maxn = 100, mostrado en la figura 54, se destacan las veces en que las variables x_2 (Miles de pesos en contratos), x_4 (P1Dn) y x_1 (Energía en contratos), presentándose respectivamente en 3, 2 y 2 ocasiones respectivamente.

MODELO TAN

Con el modelo TAN se establecen algunas de las relaciones más relevantes entre las variables predictoras. En la figura 55, se destacan dichas relaciones para el generador Termodorada.

Figura 55. Modelo clasificatorio TAN con el método 2 para el generador Termodorada.



Las relaciones establecidas para el modelo TAN del generador Termodorada se muestran en la tabla 63.

Tabla 63. Relaciones establecidas en el modelo clasificatorio TAN para Termodorada.

PADRE	**HIJO**
P1Dn	Miles de pesos en contratos
P1Dn	Embalse Agregado
Miles de pesos en contratos	Energía en contratos

Entre las relaciones establecidas por el modelo clasificatorio TAN entre las variables predictoras incluidas en el generador Termodorada, la más importante fue entre las variables miles de pesos en contratos y energía en contratos.

8.5 VARIABLES DESTACADAS EN CADA UNO DE LOS GENERADORES BAJO LOS MODELOS DESCRIPTIVOS.

8.5.1 DOBLE PROBABILIDAD CONDICIONAL (NAIVES BAYES)

Tomando cada uno de los resultados obtenidos bajo el modelo de naives bayes discreto con la doble probabilidad condicional se destacan a continuación las variables mas importantes en la determinación de los estados mas destacados de dicho modelo, estas se resumen en la tabla 64.

Tabla 64. Variables destacadas por generador bajo el modelo de doble probabilidad condicional (naives bayes).

GENERADOR	**VARIABLES**
Chivor	P1Dn, P2Dd
Guavio	P1Dn, P2Dd
Guatrón	P1Dn, P2Dn, P2Dd
San Carlos	P1Dn
Porce II	Disponibilidad, P1Dn, P2Dd
Paraíso-guaca	Disponibilidad, P2Dd
Tebsa	Energía en contratos, Miles de Pesos en contratos y precio de contratos
Tasajero	Precio en contratos
Paipa IV	Embalse Agregado, P1Dn, P2Dd
Flores	Disponibilidad, Precio en contratos
Flores III	Disponibilidad, Reconciliación positiva
Termodorada	Miles de pesos en contratos, Energía en contratos

Dentro de los resultados obtenidos en la tabla 64, se puede destacar como la variable común dentro de los generadores hidráulicos (seis primeros de la tabla) es P1Dn, sobre todo en los cuatro primeros generadores de la tabla, los cuales, poseen las mayores capacidades dentro del conjunto de generadores hidráulicos. Dentro de los dos generadores hidráulicos con menor capacidad (Porce II y Pagua) las variables destacadas dentro del conjunto son disponibilidad y P2Dd.

Entre los generadores térmicos las variables que se destacaron fueron los Precios de los contratos, los ingresos en miles de pesos por contratos y la energía transada en contratos, principalmente. Otras variables destacadas dentro de este conjunto de generadores fueron Disponibilidad, embalse agregado, reconciliación positiva, P1Dn y P2Dd.

8.5.2 ÁRBOLES DE CLASIFICACIÓN CONTINUOS

Las relaciones más relevantes establecidas con los árboles de clasificación continuos para cada uno de los generadores se resumen en la tabla 65.

Tabla 65. Relaciones destacadas bajo el modelo de árboles de clasificación continuos en cada uno de los agentes generadores.

GENERADOR	**VARIABLES**
Chivor	Disponibilidad, P1Dn, P2Dd, energía en contratos
Guavio	P1Dn, P2Dn, P2Dd
Guatrón	P1Dn, Embalse propio, P2Dn
San Carlos	P1Dn, P2Dd, P2Dn, P1Dd y precio en contratos
Porce II	Disponibilidad, P2Dd, embalse propio, precio en

	contratos
Paraíso-guaca	Embalse propio, precio en contratos
Tebsa	Miles de pesos en contratos, P1Dn, Reconciliación positiva
Tasajero	Precio en contratos, embalse agregado, P2Dn
Paipa IV	P2Dd, embalse agregado, P1Dn
Flores	Disponibilidad, energía en contratos, embalse agregado y precio en contratos
Flores III	Reconciliación positiva, disponibilidad, P2Dn y embalse agregado
Termodorada	Energía en contratos, miles de pesos en contratos y P1Dn

8.5.3 DOBLE PROBABILIDAD CONDICIONAL (NAIVES BAYES) Y ÁRBOLES DE CLASIFICACIÓN CONTINUO.

Cada una de las variables mas destacadas bajo estos dos modelos, se resume en la tabla 66. Con este resultado se quiere destacar cada una de las variables más importantes para cada agente generador en la declaración de su precio de oferta durante los años 2003, 2004 y 2005.

Tabla 66. Relaciones destacadas bajo el modelo de árboles de clasificación continuos y doble probabilidad condicional en cada uno de los agentes generadores.

GENERADOR	**VARIABLES**
Chivor	P1Dn
Guavio	P1Dn, P2Dd
Guatrón	P1Dn, P2Dn
San Carlos	P1Dn
Porce II	Disponibilidad, P2Dd
Paraíso-guaca	--------
Tebsa	Miles de pesos en contratos
Tasajero	Precio en contratos
Paipa IV	Embalse agregado, P1Dn y P2Dd
Flores	Disponibilidad y precio en contratos
Flores III	Disponibilidad y reconciliación positiva
Termodorada	Miles de pesos en contratos y energía en contratos

A excepción del generador Paraíso-guaca en el cual no se encontraron variables en común bajo los modelos de doble probabilidad condicional y el de árboles de clasificación continuos, los demás agentes generadores si presentaron semejanzas en las variables incluidas bajo estos dos modelos. En estos resultados se destacan dentro de los generadores hidráulicos la variable P1Dn en la mayoría de estos y en el primer lugar. Mientras que en los generadores

hidráulicos se destacaron en su orden las variables: miles de pesos en contratos, precio en contratos, disponibilidad, embalse agregado, entre otras.

8.5.4 TAN

Con este modelo clasificatorio se establecieron las relaciones mas importantes, entre las variables predictoras dada la variable clase, en los modelos desarrollados para cada uno de los generadores, las relaciones mas destacadas se resumen en las tablas 67 y 68. Estas se incluyeron, cuando se hayan presentado por lo menos tres veces y se muestran en el sentido en que representó el modelo TAN es decir primero la variable que fue padre y luego la que fue hijo (tabla 67) y sin importar el sentido en que se dieron (tabla 68).

Lo que se puede destacar dentro del par de variables establecidas en las siguientes tablas, es como ellas en su conjunto, son determinantes en el estado de la variable clase (precio de oferta del generador), usando como algoritmo de correlación entre las variables la Información mutua condicionada[42] .

Tabla 67. Relaciones destacadas bajo el modelo TAN de acuerdo al sentido en que se presentó.

PADRE	**HIJO**	**NÚMERO DE VECES QUE SE PRESENTÓ**
P2Dd	P1Dd	7
P2Dn	P1Dn	5
P1Dn	Embalse Agregado	5
P1Dn	P2Dn	4
P1Dn	P2Dd	4
Embalse Agregado	Disponibilidad	3

[42] Mas detalles acerca de la construcción y aplicación de este algoritmo se detallan en la sección 6.1.5

Tabla 68. Relaciones destacadas bajo el modelo TAN.

VARIABLE 1	**VARIABLE 2**	**NÚMERO DE VECES QUE SE PRESENTÓ**
P2Dd	P1Dd	9
P1Dn	Embalse Agregado	7
P1Dn	P2Dd	6
P2Dn	P1Dn	5
Energía en contratos	Miles de pesos en contratos	4
Embalse Agregado	Precio de contratos	3
Embalse Agregado	Disponibilidad	3

De las dos tablas anteriores se pueden ver como las relaciones entre las variables de las curvas de demanda residual se presenta de manera significativa, especialmente con aquellas variables correspondientes con la misma franja de demanda o aquellas que se encuentran cercanas la una de la otra, sin ser necesariamente de la misma franja de demanda (P1Dn – P2Dd). Otra relación destacable es la encontrada entre las variables Energía en contratos y Miles de pesos en contratos. Muchas de las relaciones anteriormente establecidas, tanto en las variables de la curva de demanda residual como en las relacionadas con contratos, se esperaba que se presentaran con frecuencia, debido a las similitudes o dependencias entre las variables presentes en dicha relación.

Una relación destacada dentro del conjunto obtenido es la presente entre las variables P1Dn y Embalse Agregado, la cual se presentó en todos los generadores térmicos y en uno de los hidráulicos (Paraíso-Guaca). También la

relación entre Embalse Agregado y el Precio de los contratos es destacable dentro de todo este conjunto.

Con el modelo clasificatorio TAN aplicado a cada uno de los generadores se obtuvieron las relaciones más importantes entre las variables predictoras incluidas dentro de este modelo. Las dos relaciones más destacadas entre un par de variables se detallan en la tabla 69.

Tabla 69. Relaciones mas destacadas bajo el modelo clasificatorio TAN por agente generador.

GENERADOR	**PRIMERA RELACIÓN**	**SEGUNDA RELACIÓN**
Chivor	Miles de pesos en contratos – energía en contratos	P1Dn – P2Dd
Guavio	Embalse agregado – embalse propio	P1Dn – P2Dd
Guatrón	P2Dn – P1Dn	P2Dd – P1Dd
San Carlos	P2Dd – P1Dn	P1Dn – P2Dn
Porce II	P1Dn – P2Dn	P1Dn – P2Dn
Paraíso-Guaca	P2Dn – P1Dn	P1Dn – P2Dd
Tebsa	Miles de pesos en contratos – energía en contratos	P1Dn – P2Dd
Tasajero	P2Dn – P1Dn	P1Dn – Embalse Agregado
Paipa IV	P2Dn – P1Dn	P1Dd – P2Dd
Flores	P1Dn – P2Dn	Miles de pesos en contratos – energía en contratos
Flores III	P1Dn – P2Dn	P1Dn – Embalse Agregado
Termodorada	Miles de pesos en contratos – energía en contratos	P1Dn – Embalse Agregado

Entre las relaciones mas destacadas por agente generador, establecidas en la tabla 69, se puede ver como la mayoría de estas se presentaron entre las variables de la curva de demanda residual correspondientes con la demanda mínima y mediana (P2Dn – P1Dn, P1Dd – P2Dd, P1Dn – P2Dd) y entre las variables miles de pesos en contratos y energía en contratos.

Con estos resultados se muestra, como cada una de las parejas de variables mostradas anteriormente, son importantes para el agente generador, en el establecimiento de su precio de oferta durante el periodo de estudio (años 2003, 2004 y 2005). Otra de las relaciones interesantes entre parejas de variables es la determinada por las variables P1Dn y embalse agregado, las cuales, no corresponden con las variables que normalmente dependen directamente o están correlacionadas la una de la otra, como en el caso de las establecidas en la curva de demanda residual; estas muestran una relación interesante, ya que estas variables, no comunes entre si, pero relacionadas mediante el algoritmo de información mutua condicionada, son relevantes en el establecimiento del precio de oferta de tres generadores térmicos (Tasajero, Termoflores III y Termodorada).

9. RESULTADOS DEL MODELO PREDICTIVO

Con el horizonte de los seis primeros meses del año 2006, se quieren evaluar el desempeño en general de cada uno de los clasificadores en su tarea de acertar en cada uno de los estados que puede tomar la variable clase. El aprendizaje de los clasificadores se hizo con los datos de las variables predictoras durante los años 2003, 2004 y 2005. Con esto se encuentran elementos para analizar en cada uno de los generadores los precios de oferta y ver si dichas ofertas, siguen un patrón de comportamiento dentro del mercado.

9.1 DESEMPEÑO DE LOS MÉTODOS DE CLASIFICACIÓN

Tabla 70. Eficiencias de los clasificadores evaluados durante los seis primeros meses del 2006 para los generadores hidráulicos.

		MÉTODO	SAN CARLOS	CHIVOR	GUAVIO	GUATRON	PORCE2	PAGUA
MÉTODO1	DISCRETO	NB	58.01	35.36	35.91	75.28	61.88	59.67
		TAN	44.75	28.73	39.78	74.72	65.19	60.22
	CONTINUO	NB	50.83	39.23	44.2	70.79	61.33	35.9
		ÁRBOL, maxn = 300	33.15	40.88	24.31	72.47	59.67	NA
		ÁRBOL, maxn = 100	56.35	41.44	42.54	69.1	57.46	NA
MÉTODO2	DISCRETO	NB	56.91	37.57	30.39	76.4	64.09	34.81
		TAN	45.86	25.41	40.33	66.29	56.35	43.09
	CONTINUO	NB	55.25	35.91	44.75	74.72	60.77	35.36
		ÁRBOL, maxn = 300	33.7	35.36	38.67	70.79	54.14	50.83
		ÁRBOL, maxn = 100	54.7	34.25	41.44	73.03	51.38	35.91

De la tabla 70 cabe destacar los aciertos individuales alcanzados en el generador Guatrón en el cual se alcanzó una eficiencia del 76.4% en el clasificador naives bayes discreto por el método 2. Dentro de este generador se destacaron muchos de los aciertos los cuales estuvieron por encima del 70%. También el generador Porce II se destacó con sus aciertos individuales ya que algunos de estos estuvieron por encima del 60%.

Tabla 71. Eficiencias de los clasificadores evaluados durante los seis primeros meses del 2006 para los generadores térmicos.

		MÉTODO	**TEBSA**	**TASAJERO**	**PAIPA4**	**FLORES**	**FLORES3**	**DORADA**
MÉTODO1	**DISCRETO**	NB	36.46	62.92	43.65	34.64	11.6	0
		TAN	38.12	56.74	44.75	33.99	14.92	0
	CONTINUO	NB	64.64	85.96	44.75	38.56	14.92	0
		ÁRBOL, maxn = 300	75.14	55.62	NA[43]	34.64	NA	NA
		ÁRBOL, maxn = 100	33.15	82.58	NA	34.64	NA	NA
MÉTODO2	**DISCRETO**	NB	32.04	62.92	28.18	30.07	12.15	0
		TAN	34.25	56.18	44.75	19.61	11.6	0
	CONTINUO	NB	40.88	80.34	30.94	29.41	16.02	0
		ÁRBOL, maxn = 300	27.62	83.15	42.54	16.34	14.36	0
		ÁRBOL, maxn = 100	32.6	84.83	39.23	17.65	12.15	0

De la tabla 72 se destacan los aciertos individuales alcanzados por el generador Tasajero algunos de ellos por encima del 80%, llegando a ser hasta del 85.96% con el clasificador naives bayes continuo por el método 1. Luego se destacaron los resultados alcanzados por el generador Tebsa y luego Paipa IV. Para los generadores Termoflores, Termoflores III y Termodorada, los resultados de los aciertos de los clasificadores fueron muy bajos.

[43] Para estos generadores no se pudo construir el árbol de clasificación debido al método de discretización usado, por lo tanto no aplica esta técnica de clasificación.

Tabla 72. Promedio de las eficiencias de los clasificadores para los generadores hidráulicos según el método de discretización.

	SAN CARLOS	CHIVOR	GUAVIO	GUATRON	PORCE 2	PAGUA	PROM
GENERAL	48.95	35.41	38.23	72.36	59.23	46.69	49.98
MÉTODO 1	48.62	37.13	37.35	72.47	61.11	57.83	52.77
MÉTODO 2	49.28	33.70	39.12	72.25	57.35	40.00	48.62

De la tabla 72 se destacaron en su orden por nivel de aciertos respectivamente los generadores Guatrón, Porce II, San Carlos, Paraíso-guaca, Guavio y Chivor. Los resultados de los aciertos de los clasificadores para cada uno de los generadores hidráulicos fueron destacables.

Tabla 73. Promedio de las eficiencias de los clasificadores para los generadores térmicos según el método de discretización.

	TEBSA	TASAJERO	PAIPA4	FLORES	FLORES3	DORADA	PROM
GENERAL	41.49	71.12	39.85	28.96	13.47	0.00	32.33
MÉTODO 1	49.50	68.76	44.38	35.29	13.81	0.00	35.17
MÉTODO 2	33.48	73.48	37.13	22.62	13.26	0.00	29.99

De la tabla 73 se destacaron en su orden por nivel de aciertos respectivamente los generadores Tasajero Tebsa, Paipa IV, Termoflores, Termoflores III y Termodorada. Los generadores Termoflores, Termoflores III y Termodorada tuvieron eficiencias por debajo del 30%, las cuales se consideran muy bajas para el nivel de desempeño esperado en cada clasificador. Llegando inclusive a tener aciertos del 0% en el generador Termodorada.

9.2 RESULTADOS DEL RANKING (MÉTODO DEL CODO) PARA LOS GENERADORES EN ESTUDIO.

A partir de la aplicación del método del codo, a partir del cual se obtuvieron las variables mas significativas para cada uno de los precios de oferta de los generadores, se resumen en las tablas 74 y 75, las variables mas significativas usando los métodos 1 y 2 de discretización, usando como horizonte de estudio los años 2003, 2004 y 2005. A partir de dichos resultados se encuentran el orden de las variables más importantes, a partir del algoritmo de información mutua[44] para las ofertas de cada uno de los generadores en estudio.

[44] Mas información acerca de la construcción y aplicación de este algoritmo se puede encontrar en la sección 6.1.1.

Tabla 74. Variables más destacadas para cada uno de los generadores hidráulicos por los métodos de discretización 1 y 2 usando ranking de variables (método del codo).

HIDRÁULICOS											
SAN CARLOS		**GUAVIO**		**CHIVOR**		**GUATRÓN**		**PORCE2**		**PAGUA**	
M1	**M2**	**M1**	**M2**	**M1**	**M2**	**M1**	**M2**	**M1**	**M2**	**M1**	**M2**
P1Dn	P1Dn	P1Dn	P1Dn	P1Dn	P1Dn	P1Dn	P1Dn	Disponibilidad	Disponibilidad	P1Dn	P1Dn
P2Dd	P2Dd	P2Dd	P2Dd	P2Dd	P2Dd	P2Dn	P2Dn	P1Dn	P1Dn	P2Dn	P2Dn
P2Dn	P2Dn	P2Dn	P2Dn	P2Dn	P2Dn	P2Dd	P2Dd	Contratos	Contratos	P2Dd	P2Dd
P1Dd	P1Dd	P1Dd	P1Dd	Embalse Agregado	Embalse Agregado	P1Dd	P1Dd	P2Dn	P2Dn	P1Dd	P1Dd
Precio en contratos	Precio en contratos	Precio en contratos	Precio en contratos	P1Dd	P1Dd	Contratos	Contratos	P2Dd	P2Dd	Contratos	Contratos
Embalse Agregado		Embalse Agregado	Embalse Agregado	Disponibilidad	Disponibilidad	Embalse propio	Embalse propio	P1Dd	P1Dd	Embalse propio	Disponibilidad
		Embalse propio	Embalse propio	Miles de pesos en contratos	Miles de pesos en contratos			Embalse propio	Embalse propio		Embalse Agregado
				Energía en contratos	Energía en contratos						Embalse propio

Tabla 75. Variables más destacadas para cada uno de los generadores térmicos por los métodos de discretización 1 y 2 usando ranking de variables (método del codo).

TÉRMICOS											
TEBSA		**TASAJERO**		**PAIPA4**		**FLORES3**		**FLORES**		**TERMODORADA**	
M1	**M2**	**M1**	**M2**	**M1**	**M2**	**M1**	**M2**	**M1**	**M2**	**M1**	**M2**
P1Dn	Energía en contratos	Precio en contratos	Precio en contratos	P1Dn	P1Dn	Reconciliación Positiva	Reconciliación Positiva	Disponibilidad	Disponibilidad	Precio en contratos	P1Dn
Energía en contratos	Miles de pesos en contratos	Embalse Agregado	Embalse Agregado	P2Dn	P2Dn	Embalse Agregado	Embalse Agregado	Miles de pesos en contratos	P1Dn	Energía en contratos	Precio en contratos
Miles de pesos en contratos	P1Dn	P2Dn	P2Dn	P2Dd	P2Dd	Disponibilidad	Disponibilidad	Energía en contratos	P2Dn	Embalse Agregado	Energía en contratos
P2Dd	P2Dd	P1Dn	P1Dn	P1Dd	P1Dd	P1Dn	P1Dn	P1Dn	Contratos		
P1Dd	Contratos	P1Dd		Embalse Agregado	Embalse Agregado	P2Dn	P2Dn	P2Dn	Miles de pesos en contratos		
Embalse Agregado	P1Dd							P2Dd	Energía en contratos		
Reconciliación Positiva	P2Dn								Embalse Agregado		
Precio en contratos	Embalse Agregado										
P2Dn	Reconciliación Positiva										

Con estas tablas se construyó otro ranking para establecer por los generadores hidráulicos y térmicos el conjunto de variables mas importantes, y cuales de estas se encuentran mas relacionadas con la presentación de los precios de oferta en bolsa; estos resultados se detallan en las tablas 76 y 77.

Tabla 76. Variables más destacadas dentro de los generadores hidráulicos establecidas por el método del codo.

POSICIÓN	**VARIABLE**
1	P1Dn
2	P2Dn
3	P2Dd
4	P1Dd
5	Precio en contratos
6	Disponibilidad
7	Embalse Agregado
8	Embalse Propio
9	Miles de pesos en contratos
10	Energía en contratos

En la tabla 76, cabe destacar, como dentro de los 4 primeros puestos se encuentran las variables correspondientes con la curva de demanda residual, tanto con la correspondiente con la franja de demanda mínima como con la demanda mediana.

Tabla 77. Variables más destacadas dentro de los generadores térmicos establecidas por el método del codo.

POSICIÓN	**VARIABLE**
1	P1Dn
2	Embalse Agregado
3	P2Dn
4	Precio en contratos
5	Energía en contratos
6	Disponibilidad
7	P2Dd
8	Miles de pesos en contratos
9	P1Dd
10	Reconciliación Positiva

En la tabla 77, cabe destacar, como en el primer y tercer puesto respectivamente estuvieron las variables de la curva de demanda residual correspondientes con la demanda mínima, en segundo lugar se encontró el embalse agregado, mientras en cuarto y quinto puesto estuvieron las variables precio en contratos y energía en contratos.

10. RESULTADOS GENERALES

Con base en lo modelos obtenidos del clasificador Naives Bayes discreto y el análisis de la doble probabilidad condicional durante los años 2003, 2004 y 2005 se concluye que:

- ✓ Para los generadores hidráulicos y sobre todo aquellos con mayor capacidad de generación en el país (San Carlos, Chivor, Guavio y Guatrón), la variable de la demanda residual P1Dn es la más relacionada con el precio de oferta de dichos agentes generadores. La segunda variable más importante en este conjunto de generadores fue P2Dd.

- ✓ Para los generadores térmicos, las variables relacionadas con contratos fueron las más relacionadas con el precio de oferta.

Con base en los modelos obtenidos con árboles de clasificación continuos durante los años 2003, 2004 y 2005, se concluye que:

- ✓ Las variables más importantes para el establecimiento de los precios de oferta de los generadores hidráulicos son las originadas de la curva de demanda residual, en especial aquellas relacionadas con las franjas de demanda mínima y mediana (P1Dn, P2Dn, P2Dd y P2Dd) junto con los contratos, la disponibilidad y el embalse propio.

- ✓ Las variables más relevantes para el establecimiento de los precios de oferta de los generadores térmicos son los contratos, el embalse agregado y el precio de la curva de demanda residual, P1Dn.

A partir de las relaciones que se establecieron con el modelo TAN para los generadores en estudio, durante los años 2003, 2004 y 2005, se concluye:

- ✓ Los pares de variables que se relacionaron fuertemente entre ellas y el precio de oferta de los generadores, correspondieron en primer lugar con aquellas variables afines entre ellas como las correspondientes con la curva de demanda residual y entre los contratos

- ✓ Para los generadores térmicos se observó un conjunto particular de pares de variables relacionadas fuertemente entre si y con su precio de oferta, fueron las variables embalse agregado con P1Dn y embalse agregado con el precio en contratos.

11. CONCLUSIONES

✓ El comportamiento del precio de bolsa en la demanda mínima es un elemento estratégico en la determinación de los precios de oferta de los generadores hidráulicos en estudio, así estos agentes generadores participen o no en el mercado spot durante esta franja de demanda. En la franja mínima es donde existe mayor competencia y en la cual se concentran fuertemente las ofertas de todos los generadores hidráulicos.

✓ Para la fijación de los precios de oferta de los generadores térmicos se toman como referencia el valor de sus contratos, lo cual podría ser usado como estrategia para cubrir o asegurar ingresos por los mismos.

✓ El embalse agregado es un elemento de referencia y vigilancia por parte de los generadores térmicos en estudio para la determinación de sus ofertas, siendo inclusive más importante que los precios de oferta de los demás generadores participantes en el mercado spot.

✓ La función de demanda residual es un elemento de vigilancia y decisión en la fijación de los precios de oferta, tanto de los generadores hidráulicos como térmicos evaluados; en ésta, se refleja el comportamiento de la competencia ante la estrategia del generador en estudio para atender una cantidad de energía determinada.

✓ Las variables predictoras más importantes por agente generador y por grupos de generadores hidráulicos y térmicos, fueron semejantes, aplicando distintas técnicas de clasificación. Con ésto se refleja la importancia de un conjunto particular de las variables predictoras en la fijación del precio en bolsa de dichos generadores, en cada uno de los modelos descriptivos.

✓ Los precios de oferta de los generadores Tasajero I, Guatrón y Porce II mostraron una alta asociación del conjunto de variables predictoras con el precio de oferta, lo cual nos indica que sus precios de oferta varía con la dinámica del mercado.

✓ Los precios de oferta de los generadores Termoflores, Termoflores III y Termodorada presentaron una muy baja asociación del conjunto de variables predictoras con el precio de oferta. Estos generadores no presentaron ningún comportamiento característico con las variables más importantes del mercado eléctrico.

✓ Deteniéndose en los desempeños de las técnicas de clasificación aplicadas a un agente generador, se destaca que los estados del precio que alcanzan mayores aciertos, son aquellos que presentan un buen porcentaje de datos de aprendizaje y de inferencia en la construcción del modelo de clasificación.

✓ Entre las técnicas de clasificación usadas, no hubo una que se destacara notablemente de las otras en el análisis de la sensibilidad de los precios de oferta en bolsa de los generadores del mercado eléctrico. Cada generador requiere de un análisis particular con las diferentes técnicas de clasificación para encontrar características particulares en el comportamiento en los precios de oferta.

12. BIBLIOGRAFÍA

[1] [Armañanzas, 2004]. Rubén Armañanzas Arnedillo. “Medidas de filtrado de selección de variables mediante la plataforma Elvira”. Universidad del País Vasco. 2004.

[2] [Felgaer, Sicre, 2000] Pablo Felgaer. Enrique Eduardo Sicre y otros, “Optimización de redes bayesianas basado en técnica de aprendizaje por inducción”, Buenos Aires, 2000.

[3] [Hernández, Ramírez, 2004] Hernández Orallo; Ramírez Quintana; Ferri Ramírez, “Introducción a la minería de datos”, Pearson Educación S.A., Madrid, 2004.

[4] [Hernández, 2007] Javier Augusto Hernández Romero. “Comportamiento de las ofertas de los generadores en mercados mayoristas de electricidad”, Trabajo de Investigación en Maestría de Ingeniería – Área Ingeniería Eléctrica, Universidad Industrial de Santander. Dirección: Dr. Rubén Dario Cruz Rodríguez. Septiembre, 2007.

[5] [GISEL – UIS & CPC, 2005] Grupo de Investigación en Sistemas de Energía Eléctrica – Universidad Industrial de Santander, Centro de Productividad y Competitividad del Oriente, “Modelo de Análisis de Mercados de Energía Eléctrica mediante la Aplicación de una metodología que involucra Inteligencia Competitiva y Agentes Inteligentes”. 2005.

[6] [Lacave, 2002] Carmen Lacave Rodero. “Explicación en redes bayesianas causales”, Departamento de Inteligencia Artificial, Universidad Nacional de Educación a Distancia. Madrid. 2002.

[7] [Larrañaga, Inza, 2000]. Pedro Larrañaga, Iñaki Inza. “Clasificadores bayesianos”, Departamento de ciencias de la computación e inteligencia artificial, Universidad del País Vasco. 2000.

[8] [López de Castilla, 2005] Carlos López de Castilla, “Clasificadores por Redes Bayesianas”, Tesis Maestría, Universidad de Puerto Rico, Mayagüez Campus, 2005

[9] [López de Castilla, 2005] Carlos López de Castilla, “Clasificadores por Redes Bayesianas”, Tesis Maestría, Universidad de Puerto Rico, Mayagüez Campus, 2005

[10] [Martínez & Martínez, 2002] Wendy L. Martínez, Angel R. Martínez, “Computacional Statistics Handbook with MATLAB”. Champan & Hall/CRC, 2002

[11] [Martínez & Zárate, 2007] Cesar Augusto Martínez Pinzón, Silvia Isabel Zárate Camacho, “Aplicación de Técnicas de Agrupamiento (Clustering) y Máquinas de Soporte Vectorial para la Identificación de Patrones de Comportamiento en los Precios de Oferta en Bolsa de los Generadores en el Mercado Mayorista de Energía Eléctrica en Colombia”. Trabajo de Grado en Ingeniería Eléctrica e Industrial, Universidad Industrial de Santander. Dirección: Dr. Rubén Dario Cruz Rodríguez. Codirección: Ing. Javier Augusto Hernández Romero. Agosto, 2007.

[12] [Molina, 2002] Luis Carlos Molina, “Feature selection algorithms: a survey and experimental evaluation”, Universidad Politécnica de Cataluña, 2002.

[13] [UPME, 2004] Unidad de Planeación Minero Energética – UPME, “Una Visión del Mercado Eléctrico Colombiano”, Parte del ejercicio del Plan de Expansión, 2004.

[14] [X.M. S.A. E.S.P, 2007] XM Compañía de Expertos en Mercados S.A. E.S.P., Características del Sistema Eléctrico colombiano, Gerencia Centro Nacional de Despacho, Seminario Introducción a la operación del SIN y a la Administración del Mercado, 2007.

[15] [X.M. S.A. E.S.P, 2006] XM Compañía de Expertos en Mercados S.A. E.S.P. Contratos de Energía a largo plazo, Gerencia Mercado de Energía Mayorista, Seminario Taller Transacciones en Bolsa, 2006.

[16] [X.M. S.A. E.S.P, 2006]. XM Compañía de Expertos en Mercados S.A. E.S.P. "Restricciones", Gerencia Mercado de Energía Mayorista, Seminario Taller Transacciones de Bolsa, 2006.

[17] [X.M. S.A. E.S.P, 2006]. Compañía de Expertos en Mercados S.A. E.S.P. Cargo por capacidad, Transacciones en bolsa, Gerencia Mercado de Energía Mayorista. 2006.

http://sv04.xm.com.co/neonweb/ Sitio Web de Neón

http://www.dane.gov.co Sitio Web del Departamento Administrativo Nacional de Estadísticas (DANE)

http://www.superservicios.gov.co/energiagas/energia_ind_comp_mem.htm
Número de coincidencias del precio de oferta con el precio de bolsa – Superintendencia de Servicios Públicos Domiciliarios

ANEXOS

Anexo A. Función utilizada para discretizar los datos.

```
clear all
clc
tic;
metodo=input('Digite la ruta del archivo de excel donde esta la base de datos\n');
%'C:\MATLAB7\work\GUATRON\metodo1\metodo1.xls'
ae_con=input('Digite el nombre de la pestaña donde esta el Aprendizaje y la inferencia\n');
%'matlab'
archivo=input('Digite el nombre del Generador en minusculas\n');

%Para guardar resultados:
warning off;
mkdir('C:\MATLAB7\work','ResultadosGeneradores');
mkdir('C:\MATLAB7\work\ResultadosGeneradores',archivo);



numg=[9];
for ii=1:length(numg)
   numgk=numg(ii);
   %Para el metodo 1 de discretizacion
   [O,De,titulos,limite]=metodo1(metodo,ae_con,numgk);%O=base de datos discreta sin fecha y De base de datos continua sin fecha
   [IpOrd] = MutualDiscreto(O);
   [varfin] = selecvar(IpOrd);
   varfin=varfin';
   vdr=length(varfin);
```

```
    %Para generar el formato elvira para el generador
    clc

nomarc=['C:\MATLAB7\work\ResultadosGeneradores\',archivo,'\metodo1paraelvira
_',num2str(numgk)];
    ficherodecasos2(O,titulos,nomarc);
    %Para generar el archivo de los rangos y el resultado del metodo del codo
    clc
    diary off;

delete(['C:\MATLAB7\work\ResultadosGeneradores\',archivo,'\metodo1rangos_',nu
m2str(numgk),'.txt']);

diary(['C:\MATLAB7\work\ResultadosGeneradores\',archivo,'\metodo1rangos_',nu
m2str(numgk),'.txt']);
    diary on;
    disp('*****************************************************************');
    disp(['RESULTADOS DE LOS RANGOS DISCRETIZADO CON EL METODO 1
PARA: ',num2str(archivo)]);
    disp(['El numero de grupos es ',num2str(numgk)]);
    disp('Se discretizo con el metodo 1, osea teniendo en cuenta el aprendizaje y la
inferencia');
    disp(['Estos resultados son para las ',num2str(vdr),' mejores variables del
ranking empleando el metodo del codo ']);
    disp('*****************************************************************');
    disp(' ');
    titulos=titulos([1,varfin]);
    a_dis=O(1:limite,[1,varfin]);
    e_dis=O(limite+1:end,[1,varfin]);
    a_con=[O(1:limite,1) De(1:limite,varfin)];
```

```
    e_con=[O(limite+1:end,1) De(limite+1:end,varfin)];
    varprecio=De(:,1);
    rangos2(a_dis,e_dis,a_con,e_con,titulos,varprecio);
    libro=['C:\MATLAB7\work\ResultadosGeneradores\',archivo,'\',archivo];
    pestana=['m1','g',num2str(numgk),'a_con'];
    xlswrite(libro,titulos,pestana);
    xlswrite(libro,a_con,pestana,'A2');
    pestana=['m1','g',num2str(numgk),'e_con'];
    xlswrite(libro,titulos,pestana);
    xlswrite(libro,e_con,pestana,'A2');
    pestana=['m1','g',num2str(numgk),'a_dis'];
    xlswrite(libro,titulos,pestana);
    xlswrite(libro,a_dis,pestana,'A2');
    pestana=['m1','g',num2str(numgk),'e_dis'];
    xlswrite(libro,titulos,pestana);
    xlswrite(libro,e_dis,pestana,'A2');
    diary off
end


numgk=9;
%Para el metodo 2 de discretizacion
[O,De,titulos,limite]=metodo2_999(metodo,ae_con,numgk);%O=base de datos
discreta sin fecha y De base de datos continua sin fecha
[IpOrd] = MutualDiscreto(O);
[varfin] = selecvar(IpOrd);
varfin=varfin';
vdr=length(varfin);
%Para generar el formato elvira para el generador
Clc
```

```
nomarc=['C:\MATLAB7\work\ResultadosGeneradores\',archivo,'\metodo2paraelvira
_',num2str(numgk)];
ficherodecasos2(O,titulos,nomarc);
%Para generar el archivo de los rangos y el resultado del metodo del codo
clc
diary off;
delete(['C:\MATLAB7\work\ResultadosGeneradores\',archivo,'\metodo2rangos_',nu
m2str(numgk),'.txt']);
diary(['C:\MATLAB7\work\ResultadosGeneradores\',archivo,'\metodo2rangos_',nu
m2str(numgk),'.txt']);
diary on;
disp('**********************************************************************');
disp(['RESULTADOS DE LOS RANGOS DISCRETIZADO CON EL METODO 2 PARA: ',num2str(archivo)]);
disp(['El numero de grupos es ',num2str(numgk)]);
disp('Se discretizo con el metodo 2, osea teniendo en cuenta SOLO el aprendizaje');
disp(['Estos resultados son para las ',num2str(vdr),' mejores variables del ranking empleando el metodo del codo ']);
disp('**********************************************************************');
disp(' ');
titulos=titulos([1,varfin]);
a_dis=O(1:limite,[1,varfin]);
e_dis=O(limite+1:end,[1,varfin]);
a_con=[O(1:limite,1) De(1:limite,varfin)];
e_con=[O(limite+1:end,1) De(limite+1:end,varfin)];
varprecio=De(:,1);
rangos2(a_dis,e_dis,a_con,e_con,titulos,varprecio);
libro=['C:\MATLAB7\work\ResultadosGeneradores\',archivo,'\',archivo];
pestana=['m2','g',num2str(numgk),'a_con'];
```

```
xlswrite(libro,titulos,pestana);
xlswrite(libro,a_con,pestana,'A2');
pestana=['m2','g',num2str(numgk),'e_con'];
xlswrite(libro,titulos,pestana);
xlswrite(libro,e_con,pestana,'A2');
pestana=['m2','g',num2str(numgk),'a_dis'];
xlswrite(libro,titulos,pestana);
xlswrite(libro,a_dis,pestana,'A2');
pestana=['m2','g',num2str(numgk),'e_dis'];
xlswrite(libro,titulos,pestana);
xlswrite(libro,e_dis,pestana,'A2');
diary off

%--------------------------------------------------------------------------
tiempo=(toc/60);
disp(['El tiempo gastado para correr este programa fue ',num2str(tiempo),' minutos']);
```

Anexo B. Función utilizada para los métodos de clasificación.

```
clear all
clc
tic;
archivo=input('Digite el nombre del Generador en minusculas\n');
metodo=['C:\MATLAB7\work\ResultadosGeneradores\',archivo,'\',archivo,'.xls'];
jnvp1=input('Digite 1 si el metodo de discretizacion es el 1 o digite 2 si el metodo es el 2\n');
jnvp2=input('Digite el numero de grupos de la discretizacion, puede ser 8 o 10\n');

a_dis=['m',num2str(jnvp1),'g',num2str(jnvp2),'a_dis'];
e_dis=['m',num2str(jnvp1),'g',num2str(jnvp2),'e_dis'];
a_con=['m',num2str(jnvp1),'g',num2str(jnvp2),'a_con'];
e_con=['m',num2str(jnvp1),'g',num2str(jnvp2),'e_con'];

[eficiencia, PorcAciertos, PorcApre, PorcInfe, N1, N2]= nb(metodo,a_dis,e_dis);
clc

warning off;
diary off;
delete(['C:\MATLAB7\work\ResultadosGeneradores\',archivo,'\',archivo,'clasificadores','_m',num2str(jnvp1),'_g',num2str(jnvp2),'.txt']);
diary(['C:\MATLAB7\work\ResultadosGeneradores\',archivo,'\',archivo,'clasificadores','_m',num2str(jnvp1),'_g',num2str(jnvp2),'.txt']);
diary on;
disp(['RESULTADOS CON LOS CLASIFICADORES PARA: ',num2str(archivo)]);
disp(['Datos para el Aprendizaje  ',num2str(N1)]);
```

```
disp(['Datos para la Inferencia   ',num2str(N2)]);
disp(['Metodo de discretizacion: metodo ',num2str(jnvp1)]);
disp(['Numero de grupos de la discretizacion: ',num2str(jnvp2)]);
disp(' ');
disp('**********************************************************************');
disp('Clasificadores Discretos');
disp('**********************************************************************');
disp('***NAIVES BAYES DISCRETO');
disp(['eficiencia   ', num2str(eficiencia)]);
disp('% de casos para el precio por estado en el aprendizaje   ');
disp(num2str(PorcApre));
disp('% de casos para el precio por estado en la inferencia   ');
disp(num2str(PorcInfe));
disp('% de aciertos para cada estado del precio   ');
disp(num2str(PorcAciertos));
disp('----------------------------------------------------------------');

[eficiencia]= nbMODIFICADO(metodo,a_dis,e_dis);
disp(' ');
disp('***NAIVES BAYES teniendo en cuenta el Cmap y la 2da probabilidad mas
alta como un acierto');
disp(['eficiencia   ',num2str(eficiencia)]);
disp('----------------------------------------------------------------');
disp(' ');

disp('***CLASIFICADOR TAN');
enlaces = TAN2G(metodo,a_dis);
[resultado,eficiencia,PorcInfe,PorcAciertos]                                    =
infTAN2(metodo,a_dis,e_dis,enlaces);
disp(['eficiencia   ',num2str(eficiencia)]);
```

```
disp('% de casos para el precio por estado en la inferencia   ');
disp(num2str(PorcInfe));
disp('% de aciertos para cada estado del precio   ');
disp(num2str(PorcAciertos));
disp('----------------------------------------------------------------');
cms=1;
maxn=[300 100 50 5];
for ii=1:length(maxn)
   maxnk=maxn(ii);

[subarbol_efic,subarbol,subarbol_de,subarbol_numnodos,subarbol_var,arbol_origi
nal_numnodos,arbol_original_var,PorcInfe,PorcAciertos,texto,subarbol_mejor]     =
treefinal(metodo,a_dis,e_dis,maxnk);
   mejoresdiscretos{cms}=subarbol_mejor;
   cms=cms+1;
   disp(' ');
   disp('***ARBOLES DE CLASIFICACION CON DATOS DISCRETOS');
   disp(['Estos resultados son para maxn= ',num2str(maxnk)]);
   disp(['eficiencia del mejor subarbol:   ',num2str(subarbol_efic)]);
   disp(['El subarbol mas eficiente fue el ',num2str(subarbol),' de
',num2str(subarbol_de),' y su numero de nodos es
',num2str(subarbol_numnodos)]);
   disp(['Numero de veces que la variable predictora fue nodo interno en el mejor
subarbol fue']);
   disp(texto);
   disp(['       ',num2str(subarbol_var)]);
   disp(['El numero de nodos del arbol original es
',num2str(arbol_original_numnodos)]);
   disp(['Numero de veces que la variable predictora fue nodo interno en el arbol
original fue']);
```

```
    disp(texto);
    disp(['        ',num2str(arbol_original_var)]);
    disp('% de casos para el precio por estado en la inferencia   ');
    disp(num2str(PorcInfe));
    disp('% de aciertos para cada estado del precio   ');
    disp(num2str(PorcAciertos));
    disp('-----------------------------------------------------------------');
end

disp(' ');
disp('*********************************************************************');
disp('Clasificadores Continuos');
disp('*********************************************************************');
[eficiencia,PorcInfe,PorcAciertos] = nbContinuo(metodo,a_con,e_con);
disp('***NAIVES BAYES CONTINUO');
disp(['eficiencia   ',num2str(eficiencia)]);
disp('% de casos para el precio por estado en la inferencia   ');
disp(num2str(PorcInfe));
disp('% de aciertos para cada estado del precio   ');
disp(num2str(PorcAciertos));
disp('-----------------------------------------------------------------');

[eficiencia] = nbContinuoMODIFICADO(metodo,a_con,e_con);
disp(' ');
disp('***NAIVES BAYES CONTINUO teniendo en cuenta el Cmap y la 2da
probabilidad mas alta como un acierto');
disp(['eficiencia   ',num2str(eficiencia)]);
disp('-----------------------------------------------------------------');

disp(' ');
```

```
cms=1;
maxn=[300 100 50 5];
for ii=1:length(maxn)
   maxnk=maxn(ii);

[subarbol_efic,subarbol,subarbol_de,subarbol_numnodos,subarbol_var,arbol_original_numnodos,arbol_original_var,PorcInfe,PorcAciertos,NO_SE_USA,subarbol_mejor] = treefinal(metodo,a_con,e_con,maxnk);
   mejorescontinuos{cms}=subarbol_mejor;
   cms=cms+1;
   disp(' ');
   disp('***ARBOLES DE CLASIFICACION CON DATOS CONTINUOS');
   disp(['Estos resultados son para maxn= ',num2str(maxnk)]);
   disp(['eficiencia del mejor subarbol:   ',num2str(subarbol_efic)]);
   disp(['El subarbol mas eficiente fue el ',num2str(subarbol),' de ',num2str(subarbol_de),' y su numero de nodos es ',num2str(subarbol_numnodos)]);
   disp(['Numero de veces que la variable predictora fue nodo interno en el mejor subarbol fue']);
   disp(texto);
   disp(['    ',num2str(subarbol_var)]);
   disp(['El numero de nodos del arbol original es ',num2str(arbol_original_numnodos)]);
   disp(['Numero de veces que la variable predictora fue nodo interno en el arbol original fue']);
   disp(texto);
   disp(['       ',num2str(arbol_original_var)]);
   disp('% de casos para el precio por estado en la inferencia   ');
   disp(num2str(PorcInfe));
   disp('% de aciertos para cada estado del precio   ');
```

```
    disp(num2str(PorcAciertos));
    disp('------------------------------------------------------------');
end
tiempo=toc/60;
disp(' ');disp(' ');
disp(['El tiempo total para correr el programa fue ', num2str(tiempo),' minutos']);
diary off;
```

Anexo C. Funciones utilizadas por la función ‘clasificadores’.

A continuación se muestran cada uno de las funciones implementadas dentro de la función ‘clasificadores’.

- TAN2G.m: esta es la función encargada de establecer los enlaces en el clasificador TAN, es decir, esta función establece la estructura del modelo de clasificación TAN para posteriormente usar la función inftan2.m (Ver sección 5.1.5).

```
function enlaces = TAN2G(nombre,pestana)
%ALGORITMO PARA CALCULO DE TAN
%LECTURA DE LOS DATOS DISCRETOS DE EXCEL
DATOS=xlsread(nombre,pestana);
%DATOS=ABC;
%seleccionar los datos de aprendizaje
DATOS=DATOS(1:end,:);       % 978 CHIVOR
% TAMAÑO DE DATOS A FILAS B COLUMNAS
[A,B]=size(DATOS);
% busqueda del mayor numero de grupo dado la discretizacion
%ojo teniendo en cuenta que las primera fila no contiene letras
%calculo del maximo de las clases
for m=1:B
   [varmax(m) a]=max(DATOS(:,m));
end
%hipermatriz en donde quedaran lasbb frecuencias de los datos
%frecuencia conjunta de dos variables
```

```
COSO=zeros(max(varmax),max(varmax),B,B);
%frecuencia conjunta de tres variables
conju=zeros(varmax(1),max(varmax),max(varmax),B,B);
%frecuencia conjunta de todos contra todos incluida la clase
for fila=1:A
        for vari=1:B
                for colum=vari+1:B
                        COSO(DATOS(fila,vari),DATOS(fila,colum),colum,vari)=
COSO(DATOS(fila,vari),DATOS(fila,colum),colum,vari)+1;
         if (vari~=1)
            %conju(estado de la variable clase, estado de vari, estado
            %de colum, variable colum , variable vari)

conju(DATOS(fila,1),DATOS(fila,vari),DATOS(fila,colum),colum,vari)=conju(DATO
S(fila,1),DATOS(fila,vari),DATOS(fila,colum),colum,vari)+1;
          end
                end
        end
end
IMC=zeros(B,B);
% calculo de informacion mutua condicionada
for varix=2:B
        for variy=varix+1:B
                %sumatoria triple
                for n=1:varmax(varix)
                   for m=1:varmax(variy)
                      for c=1:varmax(1)
                              %probabilidad marginal y conjunta
                        pxyc=conju(c,n,m,variy,varix)/A;
                              pxy=COSO(n,m,variy,varix)/A;
```

```
                                pyc=COSO(c,m,variy,1)/A;
                                pxc=COSO(c,n,varix,1)/A;
                                py=sum(COSO(:,m,variy,variy-1))/A;
                                pc=sum(COSO(c,:,2,1))/A;
                                if (pxc*pyc~=0 && pxyc*pc~=0)

%IMC(varix,variy)=IMC(varix,variy)+(pxyc/pc)*log10(pxyc/(pxc*pyc*pc));

IMC(varix,variy)=IMC(varix,variy)+(pxyc)*log10(pxyc*pc/(pxc*pyc));
                                end
                            end
                        end
                    end
                end
            end
        end
```

```
%armar el grafo con los datos de IMC
%estructura de control para saber si el grafo esta listo
estrG=zeros(B);
estrG=estrG(1,:);
%estructura para conocer la uniones realizadas
%la ultima columna indica el la dirección de la unión si vale 1 es dirigida de la 1 a
la 2
%si vale 2 es dirigida de forma inversa
unioG=zeros(B-2,3);
%conteo de enlaces
cont=1;
IMC2=IMC;

while (length(find(estrG==0))>1 | max(estrG(2:end))~=min(estrG(2:end)) )
        [vMAX uMAX]=max(IMC2);
```

```
%busca el enlace de mayor peso
    mayor=max(vMAX);
    [fila colum]=find(IMC2==mayor);
%examina si la unión es valida o forma un enlace ciclico
    for ii=1:length(fila)
  %busca si los puntos a unir tienen otras uniones
          if (estrG(fila(ii))~=estrG(colum(ii)) )
                %adiciona el nuevo enlace
     unioG(cont,1)=fila(ii);
                unioG(cont,2)=colum;
                cont=cont+1;
                if (estrG(fila(ii))>estrG(colum(ii)))
                      dom=fila(ii);
                      camb=colum(ii);
                else
                      dom=colum(ii);
                      camb=fila(ii);
      end
      %si es valida la unión combierte a todos los puntos unidos a un
      %valor
      if (estrG(camb)~=0)
         cfi=find(estrG==estrG(camb));
         for jj=1:length(cfi)
            estrG(cfi(jj))=estrG(dom);
         end
      else
         estrG(camb)=estrG(dom);
      end
      %si los puntos a unir no tienen otros enlaces crea una nueva
      %cadena
```

```
                elseif ((estrG(fila(ii))==0) & (estrG(colum(ii))==0) )
                        %adiciona el nuevo enlace
            unioG(cont,1)=fila(ii);
                        unioG(cont,2)=colum(ii);

                        estrG(fila(ii))=cont;
                        estrG(colum(ii))=cont;
                        cont=cont+1;
        end
        %borra el enlace que se examino
                IMC2(fila(ii),colum(ii))=0;
    end

end
%selección del mejor mutual para utilizar en el arbol
for var=2:B
        %busca el máximo de cada grupo de datos discretos
        rx=varmax(var);
        %matrix de frecuencias de los datos entre la variable clase y una variable
    Iptot=0;
    for jj=1:rx
        %Probabilidad marginal de la variable a analizar
        PmargX=sum(COSO(:,jj,var,var-1))/A;
        for k=1:varmax(1)
            %Probabilidad marginal de la variable clase
            PmargC=sum(COSO(k,:,2,1))/A;

            %conjunta
            pxc=COSO(k,jj,var,1)/A;
            if ((pxc~=0) & (PmargX*PmargC~=0))
```

```
            Ip=pxc*log10(pxc/(PmargX*PmargC));
            Iptot=Iptot+Ip;
          end
        end
    end
    IpFinal(var-1,1:2)=[var Iptot];
end
%el mejor mutual
mmutual=max(max(IpFinal(:,2)));
fila=find(IpFinal(:,2)==mmutual);
mmutual=IpFinal(fila(1),1);

%direccionar el grafo
fila=find(unioG(:,1)==mmutual);
unioG=buscaCamb(unioG,fila,1);
fila=find(unioG(:,2)==mmutual);
unioG=buscaCamb(unioG,fila,2);
%contador de lances
cont=1;
enlaces=zeros(B-2,2);
%selecciona los hijos que tienen padre
for ii=2:B
        fila=find(unioG(:,1)==ii);
        for jj=1:length(fila)
                if (unioG(fila(jj),3)==2)
                        enlaces(cont,1)=ii;
                        enlaces(cont,2)=unioG(fila(jj),2);
                        cont=cont+1;
                        break;
                end
```

```
        end
        fila=find(unioG(:,2)==ii);
        for jj=1:length(fila)
                if (unioG(fila(jj),3)==1)
                        enlaces(cont,1)=ii;
                        enlaces(cont,2)=unioG(fila(jj),1);
                        cont=cont+1;
                        break;
                end
        end
end


%se crea un ciclo anidado que recorre y cambia los enlaces del grafo
function unioG = buscaCamb(unioG,fila,colm)
        if (colm==1)
                ocolm=2;
        else
                ocolm=1;
        end
        for ii=1:length(fila)
                %la ultima columna indica el la dirección de la unión si vale 1 es dirigida de la 1 a la 2
                %si vale 2 es dirigida de forma inversa
                if (unioG(fila(ii),3)==0)
                        unioG(fila(ii),3)=colm;
                        %hay que modificar los que cambian buscando primero en la colm
                        ofila=find(unioG(:,colm)==unioG(fila(ii),ocolm));
                        unioG=buscaCamb(unioG,ofila,colm);
```

```
                %modificar los que busca en la ocolm
                ofila=find(unioG(:,ocolm)==unioG(fila(ii),ocolm));
                unioG=buscaCamb(unioG,ofila,ocolm);
            end
        end
```

- Inftan2.m: esta es la función encargada de desarrollar la inferencia para obtener la eficiencia del clasificador TAN. Se asume que previamente se ejecuto la función TAN2G.m (Ver sección 5.1.5 referente al clasificador TAN).

```
function [resultado,eficiencia,PorcInfe,PorcAciertos] = infTAN2(nombre,pestana1,pestana2,enlace)
%LECTURA DE LOS DATOS DISCRETOS DE EXCEL
total=xlsread(nombre,pestana1); % ojo este el que originalmente esta con solo pestana
otros=xlsread(nombre,pestana2);
%seleccionar los datos de aprendizaje
DATOS=total(1:end,:);
%Leer datos de la evidencia
mat2=otros(1:end,:);          % (1:49,:) %ojo este es el que originalmente esta
[N2,K2] = size(mat2);
% TAMAÑO DE DATOS A FILAS B COLUMNAS
[A,B]=size(DATOS);
% busqueda del mayor numero de grupo dado la discretizacion
%ojo teniendo en cuenta que las primera fila no contiene letras
%calculo del maximo de las clases
for m=1:B
```

```
   [varmax(m) a]=max(DATOS(:,m));
end
%************************************************************************
for ii = 1:varmax(1)%Para hallar el % de casos para el precio
   PorcInfe(ii)=length(find(mat2(:,1)==ii));
end
%************************************************************************
%hipermatriz en donde quedaran las frecuencias de los datos
%frecuencia conjunta de dos variables
COSO=zeros(max(varmax),max(varmax),B,B);
%frecuencia conjunta de dos variables con la variable clase
conju=zeros(varmax(1),max(varmax),max(varmax),B,B);
%frecuencia conjunta de todos contra todos incluida la clase
for fila=1:A
        for vari=1:B
                for colum=vari+1:B
         %al utilizar COSO hay que tener en cuenta que colum siempre es
         %mayor que vari
                        COSO(DATOS(fila,vari),DATOS(fila,colum),colum,vari)=
COSO(DATOS(fila,vari),DATOS(fila,colum),colum,vari)+1;
         if (vari~=1)
            %conju(estado de la variable clase, estado de vari, estado
            %de colum, variable colum , variable vari)

conju(DATOS(fila,1),DATOS(fila,colum),DATOS(fila,vari),colum,vari)=conju(DATO
S(fila,1),DATOS(fila,colum),DATOS(fila,vari),colum,vari)+1;
         end
                end
        end
end
```

```
aciertos=0;
%Matriz que indica la estructura del arbol
%la primera columna indica los hijo y la segunda el padre
est=enlace;
PorcAciertos=zeros(1,varmax(1));
for ii = 1:N2 %recorre la evidencia
   ii;
   %Calculo del Cmap
   cmap=0;
   norma=0;
   for c = 1:varmax(1) %recorre todos los estados de C para hallar cmap
      fc=sum(COSO(c,:,2,1));%frecuencia marginal de la variable clase
      pc=fc/A;%probabilidad marginal de la variable clase
      productoria=pc;
      for kk = 2:K2 %recorre todas las variable excepto la clase buscando la
evidencia
         %busca los padres de la variable
         pad=find(est(:,1)==kk);
         if (length(pad)>0)
            %variables a utilizar
            varx=max(est(pad,:));
            varn=min(est(pad,:));
            %frecuencia conjunta
            fxyc=conju(c,mat2(ii,varx),mat2(ii,varn),varx,varn);
            fcv=COSO(c,mat2(ii,est(pad,2)),est(pad,2),1);
            %número de casos para la laplace
            nc=length(find(COSO(:,:,est(pad,2),1)>0));
            %productoria=productoria*(fxyc+1)/(fcv+varmax(1)+varmax(est(pad,2)));
%con la laplace
            %productoria=productoria*(fxyc+1)/(fcv+varmax(kk)); %con la laplace
```

```
            %productoria=productoria*(fxyc+1)/(fcv+nc); %con la laplace
            productoria=productoria*(fxyc+1)/(fcv+5); %con la laplace

            %productoria=productoria*(fxyc+1)/(fcv+varmax(1)); %con la laplace
            %productoria=productoria*fxyc/fcv; %sin la laplace
         else
            fxc=COSO(c,mat2(ii,kk),kk,1);
            productoria=productoria*(fxc+1)/(fc+varmax(1)); %con la laplace
            %productoria=productoria*fxc/fc; %sin la laplace
         end
      end
      vector(ii,c)=productoria;
      norma=norma+productoria;
      if (productoria>cmap)
         cmap=productoria;
         estado=c;
      end
   end
        vector(ii,:)=vector(ii,:)./norma;
   resultado(ii,1)=estado;
   resultado(ii,2)=cmap/norma;
   if (resultado(ii,1) == mat2(ii,1))%Para contar los aciertos
      aciertos=aciertos+1;
      PorcAciertos(mat2(ii,1))=PorcAciertos(mat2(ii,1))+1;
   end
end
eficiencia=(aciertos/N2)*100;
PorcAciertos=(PorcAciertos./PorcInfe)*100;
PorcAciertos(find(isnan(PorcAciertos)==1))=0;
PorcInfe=(PorcInfe/N2)*100;
```

- MutualDiscreto.m[45]: esta es la función encargada de realizar el ranking de variables predictoras (Ver definición de información mutua en la sección 6.1.1 referente a términos relacionados con métodos de clasificación).

```
function [IpOrd] = MutualDiscreto(datos)
% arma_fichero('M:\temp\solo estados.xls','porce2','M:\temp\ejemplo')
%la primera columna es el precio y luego las variables
%toma los datos de excel
De=datos;
[F V]=size(De);
O=De;
%busca el máximo valor de discretizacion de la variable clase
%realiza el proceso de ranking con información mutua
rc=max(O(:,1));
%la primera columna de O es el precio
for var=2:V
    %El precio debe estar en la primera columna
  M=[O(:,var) O(:,1)];
    %busca el máximo de cada grupo de datos discretos
  rx=max(M(:,1));
    %matrix de frecuencias de los datos entre la variable clase y una variable
  New=zeros(rx,rc);
  for j=1:F
    a=M(j,1);
    b=M(j,2);
    New(a,b)= New(a,b)+(1/F);
  End
    %Probabilidad marginal de la variable a analizar
```

[45] Esta función fue desarrollada en la tesis de grado de Cesar Martínez y Silvia Zarate.

```
    PmargX=sum(New,2);
        %Probabilidad marginal de la variable clase
    PmargC=sum(New,1);
    Iptot=0;
    for j=1:rx
       for k=1:rc
          if New(j,k)~=0
             Ip=New(j,k)*log10(New(j,k)/(PmargX(j)*PmargC(k)));
             Iptot=Iptot+Ip;
          end
       end
    end
    IpFinal(var-1,1:2)=[var Iptot];
end
%ordena de mayor a menor en función de la segunda columna
IpOrd=sortrows(IpFinal,-2);
for ii=1:length(IpOrd(:,1))
        IpOrd(ii,3)=ii;
end
temp1=IpOrd(:,2);
IpOrd(:,2)=IpOrd(:,3);
IpOrd(:,3)=temp1;
```

- Nb.m: esta es la función encargada de realizar la inferencia con el método de clasificación Naives Bayes discreto (Ver sección 5.1.4 referente al clasificador Naives Bayes).

```
function [eficiencia, PorcAciertos, PorcApre, PorcInfe, N1, N2] =
nb(archivo,aprendizaje,inferencia)
%La funcion nb realiza la clasificacion de naives bayes.
%archivo: archivo de excel q contiene el aprendiaje y la inferencia.
```

```
%aprendizaje:pestaña donde estan los datos del aprendizaje.
%inferencia:pestaña donde estan los datos de inferencia.
%Aprendizaje e inferencia son discretas y las variables estan
%discreetizadas en el mismo numero de grupos

%Leer datos para el aprendizaje
mat1=xlsread(archivo,aprendizaje);
[N1,K1] = size(mat1);
%Leer datos para la inferencia
mat2=xlsread(archivo,inferencia);
[N2,K2] = size(mat2);
%Calculo del Cmap
aciertos=0;
maxc=max(mat1(:,1));%número de estados de la variable clase

for ii = 1:maxc%Para hallar el % de casos para el precio
   PorcApre(ii)=length(find(mat1(:,1)==ii));
   PorcInfe(ii)=length(find(mat2(:,1)==ii));
end
PorcAciertos=zeros(1,maxc);

for ii = 1:N2 %recorre la evidencia
   norma=0;cmap=0;
   for jj = 1:maxc %recorre todos los estados de C para hallar la probabilidad de cada estado y luego el cmap
      vcd=find(mat1(:,1)==jj);%donde esta la variable clase jj
      p=length(vcd);%número de datos por segmento de la variable clase
      pp=p/N1;%probabilidad marginal de la segmento jj que se evalua de la variable clase
      vector2(jj)=pp;
```

```
        productoria=1;

        for kk = 2:K2 %recorre todas las variable excepto la clase para hallar la probabilidad condicional
            ved=find(mat1(:,kk)==mat2(ii,kk));%localiza las posiciones de la variable de la evidencia en la base de datos del aprendizaje
            p4=0; %guarda el número de conjuntas
            for mm = 1:p %Busca las coincidencias esto es la conjunta
               p4=p4+length(find(ved==vcd(mm)));%calcula la conjunta
                     end
            omega=max(mat1(:,kk)); %correcion Laplace
            %productoria=productoria*(p4)/(p);% sin suavizante
            productoria=productoria*(p4+1)/(p+omega);% el uno que suma tambien es correcion Laplace
        end
        pp=pp*productoria;
        vector(ii,jj)=pp;
        norma=norma+pp;
        if (pp>cmap)
            cmap=pp;
            estado=jj;
        end
    end
         vector(ii,:)=vector(ii,:)./norma;
    resultado(ii,1)=estado;
    resultado(ii,2)=cmap/norma;
    if (resultado(ii,1) == mat2(ii,1))%Para contar los aciertos
        aciertos=aciertos+1;
        PorcAciertos(mat2(ii,1))=PorcAciertos(mat2(ii,1))+1;
    End
```

```
end
eficiencia=(aciertos/N2)*100;
PorcAciertos=(PorcAciertos./PorcInfe)*100;
PorcAciertos(find(isnan(PorcAciertos)==1))=0;
PorcApre=(PorcApre/N1)*100;
PorcInfe=(PorcInfe/N2)*100;
```

- nbContinuo.m: esta es la función encargada de realizar la inferencia con el método de clasificación Naives Bayes continuo (Ver sección 5.1.4 referente al clasificador Naives Bayes).

```
function [eficiencia,PorcInfe,PorcAciertos] = nbContinuo(archivo,aprendizaje,inferencia)
%La funcion nb realiza la clasificacion de naives bayes.
%archivo: archivo de excel q contiene el aprendiaje y la inferencia.
%aprendizaje:pestaña donde estan los datos del aprendizaje.
%inferencia:pestaña donde estan los datos de inferencia.
%Aprendizaje e inferencia son discretas y las variables estan
%discreetizadas en el mismo numero de grupos

%Leer datos para el aprendizaje
mat1=xlsread(archivo,aprendizaje);
[N1,K1] = size(mat1);
%Leer datos para la inferencia
mat2=xlsread(archivo,inferencia);
[N2,K2] = size(mat2);

%Calculo del Cmap
```

```
aciertos=0;errores=0;noerrores=0;
maxc=max(mat1(:,1));%número de estados de la variable clase

%------------------------------------------------------------------------
%Para estandarizar las variables predictoras
% Z=zscore(mat1(:,2:end));
% mat1=[mat1(:,1) Z];
% mat1=abs(mat1);
%------------------------------------------------------------------------
%------------------------------------------------------------------------
%Para estandarizar las variables predictoras
% Z=zscore(mat2(:,2:end));
% mat2=[mat2(:,1) Z];
% mat2=abs(mat2);
%------------------------------------------------------------------------

for w = 1:maxc %Recorre todo los estados de la variable clase
   z=find(mat1(:,1) == w); %Encuentra posiciones de cada estado de la variableclase
   pc(w)=length(z)/N1;
   for x = 2:K1 %Calcula la media y la desviacion asociada al estado de la variable clase
                media(w,x)=mean(mat1(z,x));
      desviacion(w,x)=std(mat1(z,x));
   end
end
aciertos=0;
for ii = 1:maxc%Para hallar el % de casos para el precio
   PorcInfe(ii)=length(find(mat2(:,1)==ii));
End
```

```
PorcAciertos=zeros(1,maxc);
for ii = 1:N2 %recorre la evidencia
   norma=0;cmap=0;
   for kk = 1:maxc %para recorrer todo los estados de C buscando el maximo
      productoria=1;
      for jj = 2:K1 %para hallar la probabilidad de las variables de la evidencia
         if (desviacion(kk,jj) == 0)
            errores=errores+1;
            productoria=0;
         else
            noerrores=noerrores+1;
            productoria=productoria*(1/(sqrt(2*pi)*desviacion(kk,jj)))*exp(-
((mat2(ii,jj)-media(kk,jj))^2)/(2*desviacion(kk,jj)*desviacion(kk,jj)));
         end
      end
      productoria=productoria*pc(kk);
      norma=norma+productoria;
      if (productoria > cmap)
         cmap=productoria;
         estado=kk;
      end
   end
   resultado(ii,1)=estado;
   resultado(ii,2)=cmap/norma;
   if (resultado(ii,1) == mat2(ii,1))
      aciertos=aciertos+1;
      PorcAciertos(mat2(ii,1))=PorcAciertos(mat2(ii,1))+1;
   end
end
eficiencia=(aciertos/N2)*100;
```

```
PorcAciertos=(PorcAciertos./PorcInfe)*100;
PorcAciertos(find(isnan(PorcAciertos)==1))=0;
PorcInfe=(PorcInfe/N2)*100;
```

- rangos2.m: esta es la función encargada de definir los límites de cada grupo que se obtiene de discretizar las variables.

```
function rangos2(a_dis,e_dis,a_con,e_con,titulos,varprecio)
%Hallar los limites de cada grupo.....Crea una tabla
discreto=[a_dis;e_dis];
continuo=[a_con;e_con];
continuo(:,1)=varprecio;
[H,V]=size(discreto);
numg=max(discreto(:,1));

for ii=1:V %recorre cada variable
   disp(titulos(ii));
   disp(' Estado Minimo Maximo NumerodeObsevaciones');
   for jj=1:numg %halla la tabla de limites para cada variable
      temp1=find(discreto(:,ii) == jj);
      tabla(jj,1)=jj;
      tabla(jj,2)=min(continuo(temp1,ii));
      tabla(jj,3)=max(continuo(temp1,ii));
      tabla(jj,4)=length(temp1);
   end
   tabla=sortrows(tabla,2);
   for jj=1:numg
      disp(['     ',num2str(tabla(jj,:))]);
```

```
    end
    disp('------------------------------------------------------------');
    disp(' ');
end
```

- selecvar.m: esta es la función que a partir del ranking de variables evalúa el método del codo para efectuar la selección de variables (Ver la sección 6.3.3 referente a la selección de variables).

```
function [varfin] = selecvar(IpOrd)
IpOrd(1,4)=0;
IpOrd(1,5)=0;
IpOrd(1,6)=0;
for ii=2:(max(IpOrd(:,2)))
    IpOrd(ii,4)=IpOrd(ii,3)+IpOrd(ii-1,3);
    IpOrd(ii,5)=IpOrd(ii-1,5)+IpOrd(ii,4);
end
for ii=2:(max(IpOrd(:,2)))
    IpOrd(ii,6)=1-((IpOrd(ii,5)/IpOrd(end,5))*((IpOrd(end,2)-
IpOrd(ii,2))/IpOrd(end,2)));
end
media=mean(IpOrd(:,3));
varianza=var(IpOrd(:,3),1);
limsup=media+2*varianza;
liminf=media-2*varianza;
kmax=0;
contador=0;
for ii=2:(max(IpOrd(:,2)))
    if (IpOrd(ii,3)<=limsup && IpOrd(ii,3)>=liminf)
        contador=contador+1;
```

```
        if (IpOrd(ii,6)>kmax)
            kmax=IpOrd(ii,6);
            corte=ii;
        end
    end
end
if (contador >= 1)
    varfin=IpOrd(1:corte,1);
else
    for ii=2:(max(IpOrd(:,2)))
        if (IpOrd(ii,3) > limsup)
            corte=ii;
        end
    end
    varfin=IpOrd(1:corte,1);
end
```

- treefinal.m: esta es la función encargada de evaluar los árboles de clasificación discretos y continuos (Ver sección 5.2 referente a árboles de clasificación).

```
function
[subarbol_efic,subarbol,subarbol_de,subarbol_numnodos,subarbol_var,arbol_original_numnodos,arbol_original_var,PorcInfe,PorcAciertos,texto,subarbol_mejor] = treefinal(archivo,aprendizaje,inferencia,maxn,nvp)
%La funcion tree realiza la clasificacion por arboles de clasificacion.
%archivo: archivo de excel q contiene el aprendiaje y la inferencia.
%aprendizaje:pestaña donde estan los datos del aprendizaje.
%inferencia:pestaña donde estan los datos de inferencia.
```

```
%Aprendizaje e inferencia son discretas y las variables estan
%discreetizadas en el mismo numero de grupos

%Leer datos para el aprendizaje
[X1,texto]=xlsread(archivo,aprendizaje);
temp=X1(:,1);
X1(:,1)=[];
X1=[X1 temp];
[n1,nc1]=size(X1);
texto=texto(2:end);
%Leer datos para la inferencia
X2=xlsread(archivo,inferencia);
temp=X2(:,1);
X2(:,1)=[];
X2=[X2 temp];
[n2,nc2]=size(X2);
%Encontrar para la variable clase cuantos casos hay para cada estado y
%encontrar el maximo y minimo de cada estado
% precio=xlsread('Precios2','Precios2','a2:a1093');%lee precios continuos
% precio=precio/1000;%precio en $/kWh
Nk=zeros(1,max(X1(:,end)));%se usa el max de data porq puede ser que X1 no tenga el estado 10 del precio
for ii=1:max(X1(:,end))
   ind=find(X1(:,end)==ii);
%     vector=precio(ind,end);
%     estados(ii,1)=ii;
%      if length(vector) ~= 0%si no se usa marcaria errores para hallar el min y el max
%         estados(ii,2)=min(vector);
%         estados(ii,3)=max(vector);
```

```
%        estados(ii,4)=(length(vector)/n1)*100;
%    end
   Nk(ii)=length(ind);
   PorcInfe(ii)=length(find(X2(:,end)==ii));
end
%Adecuar las variables para crecer el arbol
%maxn=300;Ahora es un dato de entrada
pies=Nk/n1;
clas=1:max(X1(:,end));%Ahora la columna discreta es la ultima
tree = csgrowcCORREGIDO(X1,maxn,clas,Nk,pies);
%csplotreec(tree) %Cuando son muchos casos no tiene claridad la grafica.
%treeseq = csprunecCORREGIDO(tree);
arbol_original=tree;
treeseq{1}=tree;
treeseq1=csprunecCORREGIDO(tree);
K=length(treeseq1);
for ii=2:K+1
   treeseq{ii}=treeseq1{ii-1};
end
K=length(treeseq);  %para conocer el numero de sub-árboles a evaluar
Rk=zeros(1,K-1);              %vector para guardar la misclassication rate
PorcAciertos1=zeros(K-1,length(Nk));
for  k=1:K-1          %recorre todos los sub-árboles excepto el último que es el
nodo raíz.
  nmis=0;             %contador de observaciones mal clasificadas.
   treek=treeseq{k}; %guarda la información del árbol actual.
   for i=1:n2         %recorre cada observación de la base de datos para la
inferencia X2.
      [clas,pclass,node]=cstreec(X2(i,1:end-1),treek);  %infiere en el árbol actual.
```

```
        if clas ~= X2(i,end)  %compara el resultado de la inferencia con el real de la base de datos.
          nmis=nmis+1;       %contador de observaciones mal clasificadas.
        end
        if clas == X2(i,end)  %compara el resultado de la inferencia con el real de la base de datos.
          PorcAciertos1(k,X2(i,end))=PorcAciertos1(k,X2(i,end))+1;  %contador de observaciones mal clasificadas.
        end
  end
  Rk(k)=nmis/n2;             %vector que contiene el % observación Mal clasificadas en cada árbol.
end
[mrk,ind]=min(Rk);  %halla el mínimo error
semrk=sqrt(mrk*(1-mrk)/n2);       %halla el error estándar
Rk2=mrk+semrk;
efic=1-Rk;
%Para tener en cuenta las variables mas importantes
subarbol_efic=0;
subarbol_de=length(Rk);
for ii=1:length(Rk)
  if (Rk(ii) <= Rk2 && efic(ii) >= subarbol_efic)%se pone Rk(ii) <= Rk2 y no < para q entre al if
     subarbol_efic=efic(ii);
     subarbol=ii;
     subarbol_mejor=treeseq{ii};
     PorcAciertos=PorcAciertos1(ii,:);
  end
end
nvp=nc2 -1;%numero de variables predictoras
```

```
subarbol_var=zeros(1,nvp);
subarbol_numnodos=0;
for ii=1:subarbol_mejor.numnodes
   if (subarbol_mejor.node(1,ii).term == 0)

subarbol_var(subarbol_mejor.node(1,ii).var)=subarbol_var(subarbol_mejor.node(1,
ii).var)+1;
      subarbol_numnodos=subarbol_numnodos+1;
   end
   if (subarbol_mejor.node(1,ii).term == 1)
      subarbol_numnodos=subarbol_numnodos+1;
   end
end
%-------------------------------------------
%Para graficar el mejor subárbol
if subarbol_numnodos <= 15
   figure
   csplotreec(subarbol_mejor);
end
%-------------------------------------------
arbol_original_var=zeros(1,nvp);
arbol_original_numnodos=0;
for ii=1:arbol_original.numnodes
   if (arbol_original.node(1,ii).term == 0)

arbol_original_var(arbol_original.node(1,ii).var)=arbol_original_var(arbol_original.n
ode(1,ii).var)+1;
      arbol_original_numnodos=arbol_original_numnodos+1;
   end
   if (arbol_original.node(1,ii).term == 1)
```

```
        arbol_original_numnodos=arbol_original_numnodos+1;
    end
end
subarbol_efic=subarbol_efic*100;
PorcAciertos=(PorcAciertos./PorcInfe)*100;
PorcAciertos(find(isnan(PorcAciertos)==1))=0;
PorcInfe=(PorcInfe/n2)*100;
```

Anexo D. Funciones utilizadas por la función ‘discretizadores’.

- Metodo1.m: esta es la función encargada de discretizar los datos de aprendizaje e inferencia de acuerdo al método 1 de discretización (Ver sección 6.3.2.3 referente a las formas de discretización usadas).

```
function [O,De,titulos,limite]=metodo1(archivo,todos,numg);
[De,titulos]=xlsread(archivo,todos);%BASE DE DATOS TODO CONTINUO APRENDIZAJE y EVIDENCIA
limite=find(De(:,1)==38717);
titulos(1)=[];
De=De(:,2:end);
%-----------------------------------------------------------------------
%numg=10; %numero de grupos
[F V]=size(De);
for colX=1:1:V        % 6:8
        B=De(:,colX);
        distax='euclidean';
        metodx='average';

        Y=pdist(B, distax);
        Z=linkage(Y, metodx);
        O(:,colX)=cluster(Z,numg);%BASE DE DATOS TODO DISCRETO APRENDIZAJE
end
```

- Metodo2_999.m: esta es la función encargada de discretizar los datos de aprendizaje e inferencia de acuerdo al método 2 de discretización (Ver sección 6.3.2.3 referente a las formas de discretización usadas).

```
function [O,temp81,titulos,limite]=metodo2_999(archivo,todos,numg);
[De,titulos]=xlsread(archivo,todos);%BASE DE DATOS TODO CONTINUO APRENDIZAJE y EVIDENCIA
limite=find(De(:,1)==38717);
titulos(1)=[];
De=De(:,2:end);
temp81=De;
Ev=De(limite+1:end,:);%evidencia
[fev cev]=size(Ev);
De=De(1:limite,:);%aprendizaje
[F V]=size(De);
%numg=8; %numero de grupos
for colX=1:1:V        % 6:8
        B=De(:,colX);
        distax='euclidean';
        metodx='average';

        Y=pdist(B, distax);
        Z=linkage(Y, metodx);
        O(:,colX)=cluster(Z,numg);%BASE DE DATOS TODO DISCRETO APRENDIZAJE
end

%Hallar los limites de cada grupo.....Crea una tabla
```

```
for ii=1:V %recoorre cada variable
   for jj=1:numg %halla la tabla de limites para cada variable
      temp1=find(O(:,ii) == jj);
      tabla(jj,1)=jj;
      tabla(jj,2)=min(De(temp1,ii));
      tabla(jj,3)=max(De(temp1,ii));
      tabla(jj,4)=mean(De(temp1,ii));
      tabla(jj,5)=length(temp1);
   end
   tabla1=sortrows(tabla,2);
   if (ii == 1)
      tablaprecios=tabla1;
   end
   %discretización de la evidencia para 9 grupos
   for kk=1:fev
     temp2=Ev(kk,ii);
     if (temp2 <= tabla1(1,3))
        Evd(kk,ii)=tabla1(1,1);
     elseif (temp2 <= tabla1(2,2))
         t1=abs(temp2-tabla1(1,4));
         t2=abs(temp2-tabla1(2,4));
         if (t1 < t2)
            Evd(kk,ii)=tabla1(1,1);
         else
            Evd(kk,ii)=tabla1(2,1);
         end
     elseif (temp2 <= tabla1(2,3))
        Evd(kk,ii)=tabla1(2,1);
     elseif (temp2 <= tabla1(3,2))
         t1=abs(temp2-tabla1(2,4));
```

```
    t2=abs(temp2-tabla1(3,4));
    if (t1 < t2)
       Evd(kk,ii)=tabla1(2,1);
    else
       Evd(kk,ii)=tabla1(3,1);
    end
elseif (temp2 <= tabla1(3,3))
   Evd(kk,ii)=tabla1(3,1);
elseif (temp2 <= tabla1(4,2))
    t1=abs(temp2-tabla1(3,4));
    t2=abs(temp2-tabla1(4,4));
    if (t1 < t2)
       Evd(kk,ii)=tabla1(3,1);
    else
       Evd(kk,ii)=tabla1(4,1);
    end
elseif (temp2 <= tabla1(4,3))
   Evd(kk,ii)=tabla1(4,1);
elseif (temp2 <= tabla1(5,2))
    t1=abs(temp2-tabla1(4,4));
    t2=abs(temp2-tabla1(5,4));
    if (t1 < t2)
       Evd(kk,ii)=tabla1(4,1);
    else
       Evd(kk,ii)=tabla1(5,1);
    end
 elseif (temp2 <= tabla1(5,3))
   Evd(kk,ii)=tabla1(5,1);
 elseif (temp2 <= tabla1(6,2))
    t1=abs(temp2-tabla1(5,4));
```

```
    t2=abs(temp2-tabla1(6,4));
    if (t1 < t2)
       Evd(kk,ii)=tabla1(5,1);
    else
       Evd(kk,ii)=tabla1(6,1);
    end
 elseif (temp2 <= tabla1(6,3))
   Evd(kk,ii)=tabla1(6,1);
elseif (temp2 <= tabla1(7,2))
    t1=abs(temp2-tabla1(6,4));
    t2=abs(temp2-tabla1(7,4));
    if (t1 < t2)
       Evd(kk,ii)=tabla1(6,1);
    else
       Evd(kk,ii)=tabla1(7,1);
    end
 elseif (temp2 <= tabla1(7,3))
   Evd(kk,ii)=tabla1(7,1);
 elseif (temp2 <= tabla1(8,2))
    t1=abs(temp2-tabla1(7,4));
    t2=abs(temp2-tabla1(8,4));
    if (t1 < t2)
       Evd(kk,ii)=tabla1(7,1);
    else
       Evd(kk,ii)=tabla1(8,1);
    end
 elseif (temp2 <= tabla1(8,3))
   Evd(kk,ii)=tabla1(8,1);
 elseif (temp2 <= tabla1(9,2))
    t1=abs(temp2-tabla1(8,4));
```

```
            t2=abs(temp2-tabla1(9,4));
            if (t1 < t2)
                Evd(kk,ii)=tabla1(8,1);
            else
                Evd(kk,ii)=tabla1(9,1);
            end
        elseif (temp2 <= tabla1(9,3))
            Evd(kk,ii)=tabla1(9,1);
%         elseif (temp2 <= tabla1(10,2))
%             t1=abs(temp2-tabla1(9,4));
%             t2=abs(temp2-tabla1(10,4));
%             if (t1 < t2)
%                 Evd(kk,ii)=tabla1(9,1);
%             else
%                 Evd(kk,ii)=tabla1(10,1);
%             end
        else
%             Evd(kk,ii)=tabla1(10,1);
            %Evd(kk,ii)=tabla1(8,1);
        Evd(kk,ii)=tabla1(9,1);
       end


    end
end
O=[O;Evd];
```

Anexo E. Correcciones a la Toolbox para Árboles de clasificación en la sección de crecimiento de un Árbol de clasificación.

Descripción del problema 1: en la función 'csgrowc.m', que es la función utilizada para el crecimiento de un árbol de clasificación se encuentra una función llamada 'splitnode' que es la encargada de encontrar la variable de corte (almacenada en la variable 'dim') y el valor de ese corte (almacenado en la variable 'split').

El problema se presenta cuando no se encuentra una variable de corte y un valor de corte[46]. Debido a que estos valores son los argumentos de entrada de la función 'addnode' (que es la función encargada de adicionar un nuevo nodo al árbol encontrado previamente la variable de corte y el valor del corte) no pueden ser nulos.

Corrección del problema 1: para solucionar este problema después que se usa la función 'splitnode' se pregunta si la variable 'split' es vacía[47]. Si es vacía entonces se omite el uso de la función 'addnode' y no se crea un nuevo nodo mientras que si no es vacía el nuevo nodo es creado con la función 'addnode' como ocurre normalmente. A continuación se muestra la programación:

```
[split,dim]=splitnode(tree.node(ind(i)).data,tree.node(ind(i)).impurity,
tree.class,tree.Nk,tree.pies);
    % split the node---CORRIGE CUANDO SPLIT ES VACIO
    if ~isempty(split)
```

[46] La razón para no encontrar una variable de corte y un valor de corte es que la base de datos actual en el nodo que se esta analizando presente valores iguales para cada variable. No debe ser necesariamente el mismo valor para todas las variables.

[47] No es necesario preguntar por la variable 'dim' debido a que si la variable 'split' es vacía la variable 'dim' también lo es y viceversa.

```
  tree = addnode(tree,ind(i),dim,split);
end
if isempty(split)
   cs=cs+1;
   tree.node(ind(i)).term=0;
   casoespecial(cs)=ind(i);
end
```

Anexo F. Corrección de la Toolbox para Árboles de clasificación en la sección de podado de un Árbol de clasificación.

Descripción del problema 1: cuando se esta en la etapa de podado, ésta se realiza de dos formas, la primera podada del árbol se hace de manera particular mientras las demás siguen un proceso repetitivo. En la primera podada del árbol, el autor de la Toolbox recomienda evaluar una función llamada 'misclassification rate' para todos los pares de nodos (se supone que cada par de nodos son nodos hermanos) y para sus respectivos nodos padres. Si la 'misclassification rate' del padres es igual a la suma de la de los hijos se recomienda eliminar estos nodos.
La función 'csprunec' incorpora lo anteriormente mencionado pero solo es valido para un solo par de nodos hijos. Cuando se presenta más de un caso donde la 'misclasiffication rate' del padre es igual a la suma de los hijos se presenta error.

Corrección del problema 1: para corregir el problema si se presentan más de un caso se realizó la siguiente corrección donde en la variable 'indparent' se guardan los nodos padres q cumplen la condición antes nombrada y en la variable 'indchildren' se guardan los nodos hijos que cumplen la misma condición:

```
if ~isempty(indchild)
  % change the parents to terminal nodes
  %PARTE DE LA CORRECCION debido a q no acepta tree.node(indpar).term =
1;
  for ii=1:length(indpar)
      tree.node(indpar(ii)).term = 1;
      tree.node(indpar(ii)).children = [];
```

```
    tree.node(indpar(ii)).var = [];
    tree.node(indpar(ii)).split = [];
end

%TERMINA PARTE DE LA CORRECCION
```

Descripción del problema 2: la técnica usada por el autor de la Toolbox durante el podado es deshabilitar los nodos que se van podando, es decir, los nodos no se eliminan completamente. Cada nodo tiene una variable llamada 'term', que puede tomar el valor de 1 cuando es un nodo terminal y 0 cuando es un nodo interno. Durante el proceso de podado esta variable es modificada a un valor de 100 si el nodo es podado.
El autor tiene en cuenta este proceso a partir de la segunda podada en adelante, pero sino no se tiene en cuenta en la primer podada y por alguna razón se presenta el problema 1 se presenta un error durante la segunda etapa de podado.

Corrección del problema 2: para corregir este problema se modifica el programa durante la primera podada y en vez de eliminar los nodos que cumplen con la condición del problema 1 se deshabilita estableciendo la variable 'term' al valor de 100. Las líneas modificadas se muestran a continuación:

```
for ii=1:length(indchild)
    tree.node(indchild(ii)).term=100;
end
% delete the children
%tree.node(indchild) = [];
```

Anexo G. Función para el crecimiento de un árbol de clasificación corregida.

```
function tree = csgrowcCORREGIDO(X,maxn,clas,Nk,pies)
% CSGROWC   Classification Tree.
%
%   TREE = CSGROWC(X,MAXN,CLAS,NK,PRIORS)
%
%   This function grows a classification tree.
%   X is a matrix containing the cases, along with a class label.
%   Each row contains a case. The first d columns of X correspond
%   to the variables/features and the last column is the class
%   label.
%   MAXN is the maximum number of cases allowed in a terminal node
%   if they do not all belong to the same class.
%   CLAS is a vector of numeric class labels: 1, 2, ...
%   NK is the number of cases in each class.
%   PRIORS is a vector of prior probabilities for each class. If this
%   is not provided, then the priors are estimated based on the
%   number in each class.
%
%   See also CSPRUNEC, CSTREEC, CSPLOTREEC, CSPICKTREEC


%   W. L. and A. R. Martinez, 9/15/01
%   Computational Statistics Toolbox

[n,dd] = size(X);
%d = d-1;
if nargin == 4% then estimate the pies
```

```
        pies = Nk/n;
end

% The tree will be implemented as a structure
% get the initial tree - which is the data set itself
tree.pies = pies;
tree.class = clas;  % need for node impurity calcs
tree.Nk = Nk;
tree.maxn = maxn;  % maximum number to be allowed in the terminal nodes
tree.numnodes = 1; % number of nodes in the tree - total
tree.termnodes = 1; % vector of terminal nodes
tree.node.term = 1; % (1) 1=terminal node, 0=not terminal
tree.node.nt = sum(Nk);  % (2) total number of points in the node
tree.node.impurity = impure(pies);   % (3)
tree.node.misclass = 1-max(pies);         % 1 - max(tree.node.pclass)  (4)
tree.node.pt = 1;      % prob it is node t  (5)
tree.node.parent = 0; % root node has no parent
% this will be a 2 element vector of node numbers to the children
tree.node.children = [];
tree.node.sibling = [];         % pointer to sibling node
tree.node.class = [];  % the class membership associated with this node
tree.node.split = []; % the splitting value
tree.node.var = [];   % the variable or dimension that will be split
tree.node.nkt = Nk; % number of points from each class in this node
% joint prob it is class k and it falls into node t
tree.node.pjoint = pies;
tree.node.pclass = pies;    %prob it is class k given node t
tree.node.data = X; % the root node contains all of the data

% Now get started on growing the tree very large
```

```
% first we have to extract the number of terminal nodes that
% qualify for splitting.
[term,nt,imp]=getdata(tree);          % get the data needed to decide to split the node

% find all of the nodes that qualify for splitting
ind = find( (term==1) & (imp>0) & (nt>maxn) );
contador=0;
cs=0;%CONTADOR DE CASOS ESPECIALES
% now start splitting
while ~isempty(ind) % while there are terminal nodes that qualify for split
%for k=1:2
  for i=1:length(ind) % check all of them
    % get split
    contador=contador+1;
    %if contador == 342
     %   pedro=14;
    %end
    [split,dim]=splitnode(tree.node(ind(i)).data,tree.node(ind(i)).impurity,...
      tree.class,tree.Nk,tree.pies);
    % split the node---CORRIGE CUANDO SPLIT ES VACIO
    if ~isempty(split)
     tree = addnode(tree,ind(i),dim,split);
    end
    if isempty(split)
      cs=cs+1;
      tree.node(ind(i)).term=0;
      casoespecial(cs)=ind(i);
    end
    %FIN DE LA CORRECION PARCIAL
  end  % end for loop
```

```
  [term,nt,imp]=getdata(tree);
  tree.termnodes = find(term==1);
  ind = find( (term==1) & (imp>0) & (nt>maxn) );
  length(tree.termnodes);
  itmp = find(term==1);
 %end
end  % end while loop
%PARA VOLVER A  PONER COMO NODOS TERMINALES LOS CASOS ESPECIALES
for ii=1:cs
  tree.node(casoespecial(ii)).term=1;
end
```

Anexo H. Función para el podado de un árbol de clasificación corregida.

```
function [treeseq]= csprunecJAVIER(tree)
% CSPRUNEC  Prune a classification tree.
%
%   TREESEQ = CSPRUNEC(TREE)
%   This function prunes a classifcation tree. The input to the
%   function is a large tree grown using CSGROWC. The output is
%   a cell array of tree structures in TREESEQ.
%
%   See also CSGROWC, CSTREEC, CSPLOTREEC, CSPICKTREEC

%   W. L. and A. R. Martinez, 9/15/01
%   Computational Statistics Toolbox

%PODA LOS NODOS HERMANOS EN LOS CUALES EL ERROR ES MAYOR O
IGUAL AL ERROR EN EL NODO PADRE
w=1;
while w>0
   [tree,w]= peorError(tree);
end
%LOS TERMINALES QUE TIENEN IGUAL CLASE SON PODADOS
w=1;
while w>0
   [tree,w]= repetidas(tree);
end
```

```
tree.alpha = 0;
k=1;
% treeseq{k} = tree;% this is the first tree in the sequence
treeseq = tree;        % this is the first tree in the sequence
% % NOTE: The variable tree should always have the same number
% % of nodes as the original tree(1). We will just be re-setting
% % the flags to 1 if terminal and the children to empty.
%
% numnodes = tree.numnodes;
%
% while numnodes > 1              % find weakest link branches
%    k=k+1;
%    [term,misclass,pt]=getinfo(tree);
%    % now get the other trees in the sequence
%    node = find(term==0); % all non-term nodes
%    numbranch = length(node);  % number of non-term nodes
%    gt = zeros(1,numbranch);
%    % loop over all of the branches (non-terminal nodes)
%    for i = 1:numbranch
%       % get the weakest link function for all of the branches
%       gt(i) = brnchmisclas(node(i),tree);
%    end
%    [mg,mi]=min(gt);        % find the weakest link
%
%    weak=find(gt==mg);
%    mi=max(weak);% prune the maximum node, if more than one weak link
%    % prune those branches
%    branch=getbranch(node(mi),tree);
%    tree.node(node(mi)).term=1;
%    tree.node(node(mi)).children=[];
```

```
%    tree.node(node(mi)).split=[];
%    tree.node(node(mi)).var=[];
%    for i = 1:length(branch)
%        tree.node(branch(i)).term=100;
%        % A flag of 100 means it has been pruned off
%    end
%    numnodes = numnodes-length(branch);    % number of nodes pruned off
%    tree.alpha = mg;
%    treeseq{k}=tree;
% end


function [tree,w]= repetidas(tree)
[term,misclass,pt]=getinfo(tree);
ind = find(term == 1);
indchild=[];
indpar = [];
% decide which ones to prune
w = 0;
for i = 1:length(ind)  % these are the terminal nodes
  parnode = tree.node(ind(i)).parent;
  sibnode = tree.node(parnode).children;
       %EVALUA SI LOS NODOS TERMINALES HERMANOS TIENEN LA
MISMA CLASE
  if all(term(sibnode)) && tree.node(sibnode(1)).class ==
tree.node(sibnode(2)).class
     w = w+1;
     %JAVIER linea original
     %indchild(w) = ind(i);
     %LOS terminales hijos son dos y no uno como el caso del parent
```

```
      indchild(w,:) = sibnode;
      indpar(w) = parnode;

   end
end
if ~isempty(indpar)
   tree=borrar(tree,indchild,indpar);
end



function tree =borrar(tree,indchild,indpar)
        %PARTE DE LA CORRECCION debido a q no acepta
tree.node(indpar).term = 1;
  for ii=length(indpar):-1:1
     if ii==length(indpar) || indpar(ii)~=indpar(ii+1)
        if tree.node(indchild(ii,2)).term==0
           tree=borrar(tree,tree.node(indchild(ii,2)).children,indchild(ii,2));
        end
        tree.node(indchild(ii,2)).term=100;
        if tree.node(indchild(ii,1)).term==0
           tree=borrar(tree,tree.node(indchild(ii,1)).children,indchild(ii,1));
        end
        tree.node(indchild(ii,1)).term=100;
     end
     tree.node(indpar(ii)).term = 1;
%       tree.node(indpar(ii)).children = [];
%       tree.node(indpar(ii)).var = [];
%       tree.node(indpar(ii)).split = [];

  End
```

```
            [term,misclass,pt]=getinfo(tree);
    tree.numnodes = length(find(term==1))+length(find(term==0));
    tree.termnodes = find(term==1);

function [tree,w]= peorError(tree)
[term,misclass,pt]=getinfo(tree);
ind = find(term == 1);
indchild=[];
indpar = [];
% decide which ones to prune
w = 0;
for i = 1:length(ind)  % these are the terminal nodes
   parnode = tree.node(ind(i)).parent;
   sibnode = tree.node(parnode).children;
    %if all(term(sibnode))
        % if they are both terminal (i.e., equal to 1), then check for pruning
      rt = misclass(parnode)*pt(parnode);
      rtl = misclass(sibnode(1))*pt(sibnode(1));
          % JAVIER LA SIGUIENTE ES LA LINEA ORIGINAL
      %rtr = misclass(sibnode(2))*pt(sibnode(1));
          rtr = misclass(sibnode(2))*pt(sibnode(2));
          % JAVIER LA SIGUIENTE ES LA LINEA ORIGINAL
          %if rt == (rtl+rtr)
        if rt <= (rtl+rtr) && parnode~=1
            w = w+1;
          %JAVIER linea original
            %indchild(w) = ind(i);
          %LOS terminales hijos son dos y no uno como el caso del parent
          indchild(w,:) = sibnode;
            indpar(w) = parnode;
```

```
        end
    %end
end
if ~isempty(indpar)
    tree=borrar(tree,indchild,indpar);
end
```

Anexo I. Función para el modelo de doble probabilidad condicional

- ✓ NaivesDtres.m : Con esta función se asocian pesos en cada uno de los estados de la variable clase (precio de oferta) con cada una de las variables predictoras, identificando aquellos productos mas significativos que luego se asociarán como variables de estudio.

```
function [COSO] = naivesDtres(nombre,pestana)

%tic;
%ESTA FUNCION calcula los estados más significativos asociados a cada estado de la variable clase
%LA VARIABLE CLASE ESTA EN LA PRIMERA COLUMNA
%TODOS LOS DATOS SON DISCRETOS
%LECTURA DE LOS DATOS DISCRETOS DE EXCEL
total=xlsread(nombre,pestana);

% SELECCIONAR VARIABLES MEJOR RANKING

% total(:,2)=[];%Eliminar fecha
% total(:,2)=[];%Eliminar fecha

mat1=total;

[N1,K1] = size(mat1);

%Calculo del Cmap
aciertos=0;
```

```
maxc=max(mat1(:,1));%número de estados de la variable clase

%BUSCA LA CANTIDAD DE GRUPOS POR VARIABLE
%en cuantos grupos esta discreta cada variable
for m=2:K1
%     varmax es el valor maximo de las otra variables
   [varmax(m) a]=max(mat1(:,m));
end
%Calcular la conjunta de la variable clase versus las otras

%frecuencia conjunta de la variable clase versus las otras
fconj=zeros(maxc,max(varmax),K1);
%marginal de las variables
mar=zeros(K1,max(varmax));
%maxima probabilidad condicional
condx=zeros(maxc*(K1-1),2);
%probabilidad condicional
COSO=zeros(maxc,max(varmax),K1);

%frecuencia conjunta de todos contra la clase
for fila=1:N1
        %ACUMULA LA FRECUENCIA MARGINAL DE LA VARIABLE CLASE
        mar(1,mat1(fila,1))=mar(1,mat1(fila,1))+1;
        for vari=2:K1
                %ACUMULA LA FRECUENCIA CONJUNTA
          fconj(mat1(fila,1),mat1(fila,vari),vari)=
fconj(mat1(fila,1),mat1(fila,vari),vari)+1;
         %ACUMULA LA FRECUENCIA MARGINAL POR VARIABLE
                mar(vari,mat1(fila,vari))=mar(vari,mat1(fila,vari))+1;
        end
```

```
end

%PRODUCTO DE LAS CONDICIONALES DADA LA CLASE POR LA CONDICIONAL DADA LA VARIABLE
for varclase=1:maxc


            %PRODUCTO condicional
        COSO(varclase,:,:)= (fconj(varclase,:,:).^2)./(mar(1,varclase));
            %COSO(acierto(fila),:,:)= fconj(acierto(fila),:,:);

end

for vari=2:K1
      for fila=1:maxc
            COSO(:,fila,vari)=COSO(:,fila,vari)./(mar(vari,fila));
      end
end
contar=1;
%CALCULA EL MÁXIMO producto condicional asociado a cada clase.
for varclase=1:maxc
   %probabilidad máxima
   for vari=2:K1
      [condx(contar,1) condx(contar,2)]=max(COSO(varclase,:,vari));
      contar=contar+1;
   end

end
%contar=1;
%se buscan las probabilidades condicionales de los mejores aciertos
```

```
%acierto=[4 8 10];
%acierto=[2 8 4 3 7];
%for fila=1:5
```

Anexo J. Función para el modelo de Naives Bayes discreto

- ✓ NaivesD.m: Con este programa se busca encontrar con el clasificador Naives Bayes discreto los estados más relevantes que se asociarán con los obtenidos por el modelo de doble probabilidad conjunta.

```
function [eficiencia] = naivesD(nombre,pestana)

%El precio esta en la primera columna
%LECTURA DE LOS DATOS DISCRETOS DE EXCEL
total=xlsread(nombre,pestana);

% SELECCIONAR VARIABLES MEJOR RANKING
%total=total(:,[1,4,5,7,8,14,15]);
%total=total(:,[1,4,5,7,8]);%cuatro MEJORES RANKING
%total=total(:,[1,5,7]);%DOS MEJORES RANKING
%total=total(:,[1,3,4]);%DOS MEJORES RANKING
%total=total(:,[1,3]);%una MEJOR RANKING
%total=total(:,[1,4,5,7,8,14]); %SIN EMBALSE PROPIO
%total=total(:,[1,5,7]);
%SIN DELTAS
% [A,B]=size(total);
% total=total(:,1:B-3);

%seleccionar los datos de aprendizaje
%mat1=total(364:1062,:);
%mat1=total(51:end,:);
%mat1=total(1:993,:);
```

```
%mat1=total(1:1042,:);
%mat1=total(1:1093,:);%2006
%mat1=total(364:1093,:);%2004 y 2005
%mat1=total(728:1093,:);%2005
%mat1=total(364:1333,:);%2004 a 2006 jul
%mat1=total(728:1333,:);%2005 a 2006 jul
%mat1=total(1:1333,:);%2003 a 2006 jul
%mat1=[total(100:246,:) ; total(473:577,:) ];
%mat1=total(1:577,:);
%mat1=[total(1:100,:) ; total(246:473,:) ]
mat1=total;

%[mat1,var1] = xlsread('aprendizaje','SANCARLOS');
[N1,K1] = size(mat1)
%Leer datos de la evidencia
%[mat2,var2] = xlsread('evidencia','SANCARLOS');
mat2=total;
%mat2=total(1:50,:);
%mat2=total(994:end,:);            % (1:49,:)
%mat2=total(1043:end,:);
%mat2=total(1334:end,:);%2006
%mat2=total(1063:end,:);
%mat2=total(577:834,:);
[N2,K2] = size(mat2)



%Calculo del Cmap
aciertos=0;
maxc=max(mat1(:,1));%número de estados de la variable clase
```

```
%Calcular la conjunta de la variable clase versus las otras
for m=2:K1
%     varmax es el valor maximo de las otra variables
   [varmax(m) a]=max(mat1(:,m));
end
%frecuencia conjunta de la variable clase versus las otras
COSO=zeros(maxc,max(varmax),K1);
%marginal de la variable clase
marc=zeros(maxc,1);
%frecuencia conjunta de todos menos la clase
fconj=zeros(10,10,10);
fconjT=zeros(10,10,10,10);
%maxima probabilidad condicional
condx=zeros(maxc*(K1-1),2);
%frecuencia conjunta de todos contra la clase
for fila=1:N1
        marc(mat1(fila,1))=marc(mat1(fila,1))+1;

fconj(mat1(fila,2),mat1(fila,3),mat1(fila,4))=fconj(mat1(fila,2),mat1(fila,3),mat1(fila,4
))+1;

fconjT(mat1(fila,1),mat1(fila,2),mat1(fila,3),mat1(fila,4))=fconjT(mat1(fila,1),mat1(fil
a,2),mat1(fila,3),mat1(fila,4))+1;
        for vari=2:K1

         COSO(mat1(fila,1),mat1(fila,vari),vari)=
COSO(mat1(fila,1),mat1(fila,vari),vari)+1;

        end
end
```

```
% contar=1;
% for fila=1:maxc
%
%
%                %probabilidad condicional
%            COSO(fila,:,:)= COSO(fila,:,:)./marc(fila);
%     %probabilidad máxima
% %     for vari=2:K1
% %         [condx(contar,1) condx(contar,2)]=max(COSO(fila,:,vari));
% %         contar=contar+1;
% %     end
%
% end

for ii = 1:N2 %recorre la evidencia
   norma=0;cmap=0;
   for jj = 1:maxc %recorre todos los estados de C para hallar cmap
      vcd=find(mat1(:,1)==jj);%donde esta la variable clase jj
      p=length(vcd);%número de datos por segmento de la variable clase
      pp=p/N1;%probabilidad marginal de la segmento jj que se evalua de la
variable clase
      vector2(jj)=pp;
      productoria=1;

      for kk = 2:K2 %recorre todas las variable excepto la clase buscando la
evidencia
         ved=find(mat1(:,kk)==mat2(ii,kk));%localiza la variable de la evidencia
         p4=0; %guarda el número de conjuntas
         for mm = 1:p %Busca las coincidencias esto es la conjunta
            p4=p4+length(find(ved==vcd(mm)));%calcula la conjunta
```

```
                    end
            %omega=max(mat1(:,kk)); %correcion Laplace
            %productoria=productoria*(p4)/(p);% sin suavizante
            productoria=productoria*(p4+1)/(p+maxc);% el uno que suma tambien es
correcion Laplace
        end
        pp=pp*productoria;
        vector(ii,jj)=pp;
        norma=norma+pp;
        if (pp>cmap)
            cmap=pp;
            estado=jj;
        end
    end
        vector(ii,:)=vector(ii,:)./norma;
    resultado(ii,1)=estado;
    resultado(ii,2)=cmap/norma;
    resultado(ii,3)=mat2(ii,1);
    if (resultado(ii,1) == mat2(ii,1))%Para contar los aciertos
        aciertos=aciertos+1;
    end
end
eficiencia=(aciertos/N2)*100;

%Nombre de la pestaña en la cual guarda los resultados en excel


    xlswrite(nombre, resultado,'naives','B2');
```