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 Del Mercado Mayorista De Energía Eléctrica en Colombia

CÉSAR AUGUSTO MARTÍNEZ PINZÓN
SILVIA ISABEL ZÁRATE CAMACHO

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICO-MECÁNICAS
ESCUELA DE ESTUDIOS INDUSTRIALES Y EMPRESARIALES
BUCARAMANGA
2007

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 Del Mercado Mayorista De Energía Eléctrica en Colombia

CÉSAR AUGUSTO MARTINEZ PINZÓN
SILVIA ISABEL ZÁRATE CAMACHO

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

Director:
PhD. Rubén Darío Cruz Rodríguez

Codirector:
Ing. Javier Augusto Hernández Romero

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICO-MECÁNICAS
ESCUELA DE ESTUDIOS INDUSTRIALES Y EMPRESARIALES
BUCARAMANGA
2007

RESUMEN[1]

Título:

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 Del Mercado Mayorista De Energía Eléctrica en Colombia.

Autores:

Silvia Isabel Zárate Camacho

César Augusto Martínez Pinzón

Palabras Claves:

Máquinas de Soporte Vectorial (MSV), Clustering, conglomerados, Mezclas Finitas, curvas de demanda residual, patrones, clasificador bayesiano de Naives, mercado eléctrico mayorista (MEM), estadística.

Descripción:

Este trabajo de grado es una investigación en la que se adaptaron técnicas estadísticas para buscar patrones de comportamiento en la fijación de los precios de oferta en la bolsa de energía, de un conjunto de centrales generadoras de electricidad, a partir de variables consideradas estratégicas para el Mercado Eléctrico Mayorista colombiano y con el aporte de la implementación de las curvas de demanda residual de las que se extrajeron una gran cantidad de importantes descriptores o variables con las que llevaron a cabo los estudios.

El Análisis Cluster, las Máquinas de Soporte Vectorial, las Mezclas Finitas y el Clasificador Bayesiano de Naives fueron las técnicas implementadas que permitieron procesar la vasta cantidad de datos disponibles, en búsqueda de la extracción de la información contenida en ellos y de este modo encontrar elementos representativos de diferentes franjas de precios casados en la bolsa por cada una de las centrales seleccionadas en el estudio.

Para ello, se partió de las bases de datos donde se encuentran los registros históricos de las variables seleccionadas para el análisis, datos que fueron sometidos a un preprocesamiento que les permitió funcionar como insumo a las técnicas mencionadas, las cuales se aplicaron de modo que los resultados obtenidos sirven de aporte tanto a la monitorización del mercado eléctrico mayorista de Colombia con la identificación de patrones, como al estudio mismo de la estadística y en particular al de las herramientas aquí empleadas.

---

[1]Facultad de Ingenierías Físico-Mecánicas. Escuela de Estudios Industriales y Empresariales / Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones. *Director de proyecto*: Rubén Darío Cruz Rodríguez.

ABSTRACT[2]

Tittle:

Application of Clustering and Support Vector Machines for identifying patterns of behavior in the offer prices on the Spot Market of energy of electricity-generating plants at the Electricity Wholesale Market in Colombia.

Authors:

Silvia Isabel Zárate Camacho

César Augusto Martínez Pinzón

Keywords:

Support Vector Machines (SVM), Clustering, conglomerates, Finite Mixture, Residual Demand Curves, Patterns, Naives-Bayes Classifier, Electricity Wholesale Market, Statistics.

Description:

This paper is an investigation in which statistical techniques were adapted in order to search behavior patterns in the pricing offer on the Spot Market of energy in a set of electricity-generator plants, starting with variables considered as strategic for the Colombian Wholesale Electricity Market and with the support of the implementation of the residual demand curves from which were extracted some of the principal variables used in the subsequent analysis.

Clustering, Support Vector Machines, Finite Mixtures and the Naives-Bayes Classifier were the techniques implemented that allowed the processing of the vast available data to extract the information contained in them and thus find representative elements of different price settled stripes in the spot market for each of the facilities selected in the study.

To that end, the start was de data bases which contains the historical records of the selected variables, data that were subjected to a pre-process that permitted it operate as a input to the above-mentioned techniques, which were applied so that results are useful to both for the monitoring of the Colombian Wholesale Electricity Market with the pattern identification, such as the study of statistics and in particular to the tools used here.

---

[2]Facultad de Ingenierías Físico-Mecánicas. Escuela de Estudios Industriales y Empresariales / Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones. *Project Director*: Rubén Darío Cruz Rodríguez**.**

TABLA DE CONTENIDO

Lista de tablas

Lista de figuras

1. INTRODUCCIÓN

En la historia reciente en los mercados eléctricos se ha venido tomando conciencia sobre la relevancia de la monitorización de la oferta de la energía haciendo urgente la necesidad de crear entes especializados para la monitorización que puedan vigilar el comportamiento de dichas ofertas, identificar comportamientos especulativos y crear mecanismos que permitan mejorar la forma como se realizan las transacciones.

En Colombia esta responsabilidad recae sobre la Superintendencia de Servicio Públicos Domiciliarios (SSPD) pero está apoyada en otras entidades como la Unidad de Planeación Minero Energética (UPME), la Comisión de Regulación de Energía y Gas (CREG) y el operador y administrador del Mercado, el XM; en conjunto, estas organizaciones contribuyen con la vigilancia, registro y búsqueda de mecanismos para obtener un mejor desempeño del Mercado de Energía Mayorista (MEM).

Algunas de las razones por las que los mercados de energía eléctrica necesitan ser monitorizados son características como la inelasticidad de la demanda de corto plazo, la obligatoriedad de que todo lo que se produzca se consuma inmediatamente (imposibilidad técnica de almacenamiento a gran escala), la dependencia de una red desarrollada, y las economías de escala, entre otras. Lo que se busca, es prevenir que se ejerza poder de mercado unilateral ya que ciertas empresas generadoras tienen la capacidad para manipular el precio del mercado en beneficio propio en razón a que los consumidores no pueden reaccionar en el corto plazo; a esto se suma que la oferta representada por las empresas generadoras está concentrada en pocos agentes de gran tamaño, lo que en definitiva hace de este mercado uno bastante complejo.

Sin embargo, conseguir esto es un reto al que se enfrenta el personal de los entes de control encargado de estos análisis y del seguimiento ya que deben transformar la inmensa cantidad de datos que genera el Mercado, en información útil que permita la toma de decisiones acertadas y de manera oportuna. En otras palabras, las entidades mencionadas deben convertir datos sin procesar en información de altísima utilidad y de esta forma liberar y aprovechar el potencial contenido en los datos para conseguir una ventaja competitiva en el mercado.

Este trabajo de grado pretende aportar a esta tarea de monitorización mediante la adaptación de técnicas estadísticas enmarcadas en la minería de datos y la inteligencia artificial, al análisis de las ofertas de un grupo de centrales generadoras en la bolsa de energía, el cual constituye uno de los elementos más importantes del Mercado Eléctrico Mayorista Colombiano.

2.GENERALIDADES DEL PROYECTO

2.1. Objetivos

2.1.1.Objetivo general

Reconocer los patrones característicos de los precios de oferta de energía eléctrica y las variables externas (precios de oferta, demanda, precio de contratos, reconciliaciones, inflexibilidades y nivel de los embalses) más relevantes que intervienen en su conformación, para un conjunto de centrales de generación en el Mercado Eléctrico Colombiano, de tal manera que permitan la clasificación de futuros grupos de datos de dichas variables en categorías de precios etiquetadas.

2.1.2.Objetivos específicos

Objetivo	**Relación cumplimiento de objetivo**
Calcular la curva de demanda residual de cada una de las centrales generadoras seleccionadas para su utilización en el análisis "cluster".	Secciones 3.4 y 7.2. Se estudiaron los conceptos, se determinaron las variables requeridas en la aplicación de esta herramienta; para cada central se calcularon tres curvas diarias representativas de donde se extraen los elementos que luego se convirtieron en variables para utilizarse en las etapas posteriores de la investigación. Estas variables estuvieron bien clasificadas en el ranking
Adaptar el análisis "cluster" y las MSV a la obtención de patrones de los precios de oferta en bolsa de energía eléctrica.	Secciones 7.3 a 7.6, 8.1y 8.2. Para adaptar estas técnicas se seleccionó un conjunto de variables que fueron preprocesadas, con las cuales se llevó a cabo la búsqueda de agrupamientos naturales que permitieran establecer las etiquetas necesarias para la implementación de las MSV, con las que se obtuvieron los patrones de los precios de oferta de cada central.
Obtener conglomerados del	Secciones 4.1, 4.3, 7.5 y 8.1. Se

conjunto de datos de las variables externas (precios de oferta, demanda, precio de contratos, reconciliaciones, inflexibilidades y nivel de los embalses) al generador que intervienen en la formación de los precios de oferta de energía eléctrica en bolsa de las centrales seleccionadas y su precio característico, si existe.	aplicó el análisis cluster con el conjunto completo de variables, pero no se encontraron agrupamientos naturales que puedan reconocerse como etiquetas, por lo tanto fue necesario emplear cluster univariado aplicado al precio de oferta.
Validar los patrones obtenidos para los precios de oferta en bolsa de energía de las centrales seleccionadas con datos del período comprendido entre el 1° de Junio de 2005 y el 31 de Mayo de 2006.	Sección 8.2. Luego de realizar el entrenamiento de las MSV se clasificaron los datos correspondientes al período en mención. Para realizar la validación se crearon índices que permitieron determinar la eficiencia de las máquinas y la existencia de patrones particulares para cada una de las centrales.
Encontrar valores atípicos en los precios de la energía en bolsa ofertados por los generadores seleccionados y presentar un primer acercamiento a sus posibles causas.	Sección 8.1.2. Los datos atípicos fueron encontrados luego de aplicar el cluster univariado con el que se obtuvieron las etiquetas. Se encontraron características comunes en los casos extraídos para las centrales hidráulicas y térmicas.
Obtener la función de densidad de probabilidad de las variables seleccionadas mediante mezclas finitas para implementar el Clasificador Bayesiano Naive.	Sección 6 y 8.3. Para cada una de las centrales generadoras se implementó el algoritmo EA que permitió encontrar la función de densidad de probabilidad para los grupos previamente etiquetados. A partir de los parámetros que determinan estas funciones se validaron los datos usando el Clasificador Naives-.Bayes y se calcularon los índices creados para las MSV.

Tabla 2.1 Tabla de cumplimiento de objetivos.

3. MERCADO ELÉCTRICO MAYORISTA

3.1. El mercado colombiano de energía

El mercado mayorista de energía colombiano, tal y como se le conoce hoy, es el resultado de un proceso que lleva ya varios años desarrollándose, sin embargo, corresponde a un sector que sigue en etapa de aprendizaje. La legislación, los cambios en el consumo, los nuevos competidores, los cambios macroeconómicos, entre otras muchas cosas, hacen que este mercado se mantenga en continuo movimiento, de modo que lo que se busca con este capítulo es ubicar dentro de un marco de referencia al lector interesado en este proyecto.

3.1.1. Generalidades del MEM

A partir de la puesta en vigencia de las Leyes 142 y 143 de 1994 se han presentado en el ámbito nacional una serie de hechos que han marcado el rumbo del desarrollo del mercado de energía en Colombia.

La Ley 143 denominada Ley Eléctrica (LE) reglamentó de manera específica y complementaria el servicio de electricidad y la Ley 142 de Servicios Públicos Domiciliarios (LSPD) estableció un marco general para los servicios públicos domiciliarios (incluidos el gas natural por redes y el Gas Licuado del Petróleo). Esta última transformó a la Comisión de Regulación Energética (CRE) en la Comisión de Regulación de Energía y Gas (CREG), la cual tomando como base los desarrollos regulatorios realizados por la CREG, diseña, reglamenta e implementa el nuevo marco institucional del sector eléctrico y de gas.

Con el fin de garantizar la eficiencia en la prestación del servicio de energía eléctrica se creó un mercado mayorista basado en ofertas, denominado MEM (Mercado de Energía Mayorista), en el cual pueden participar los agentes que realizan actividades de generación, transmisión, distribución y comercialización; las reglas y requisitos para la participación en el MEM se estipulan dentro de las Leyes 142 y 143 de 1994 y a través de estos años se han ido haciendo modificaciones por parte de la CREG, mediante resoluciones que buscan mejorar las deficiencias que se han identificado.

Así mismo se crea la UPME (Unidad de Planeación Minero Energética) en 1994, organizada como Unidad Administrativa Especial adscrita al Ministerio de Minas y Energía, que tiene entre sus funciones elaborar y actualizar el Plan de Expansión de Referencia del sector eléctrico, de tal manera que los planes para atender la demanda sean lo suficientemente flexibles para que se adapten a los cambios que determinen las condiciones técnicas, económicas, financieras y ambientales.

Además de estos órganos regulatorios, de control y planeación se creó una serie de entidades encargadas de la operación y administración del Sistema Interconectado Nacional (SIN) y del MEM, respectivamente, tales como el CND (Centro Nacional de Despacho) y el ASIC (Administrador del Sistema de Intercambios Comerciales); dichas entidades actualmente forman parte de XM,

Compañía de Expertos en Mercados S.A. E.S.P, quien es la encargada de operar el sistema y administrar el mercado.

El XM es una empresa del grupo ISA, y su estructura organizacional puede ser visualizada en el esquema de la figura 3.1 y de manera general en la figura 3.2, se muestra la estructura institucional del Mercado Eléctrico.

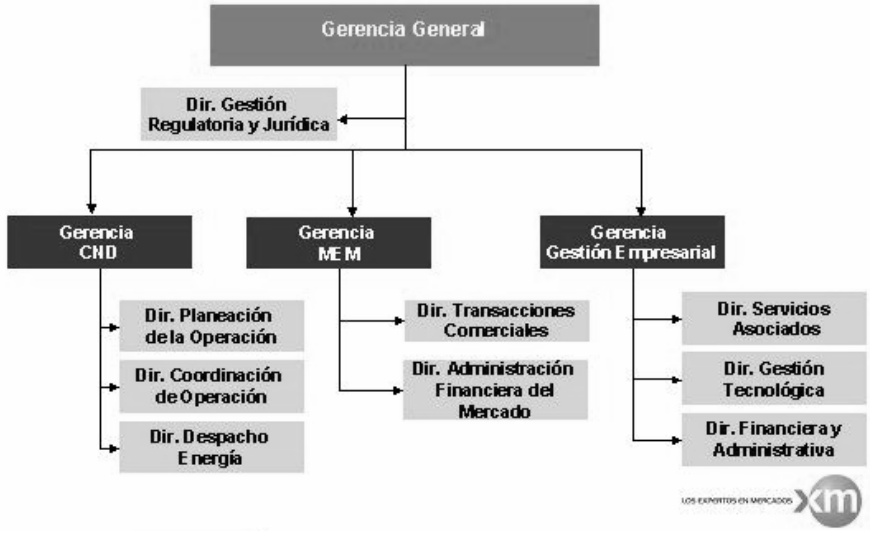

Figura 3.1 Estructura Organizacional de XM. Tomado de la página oficial de XM.

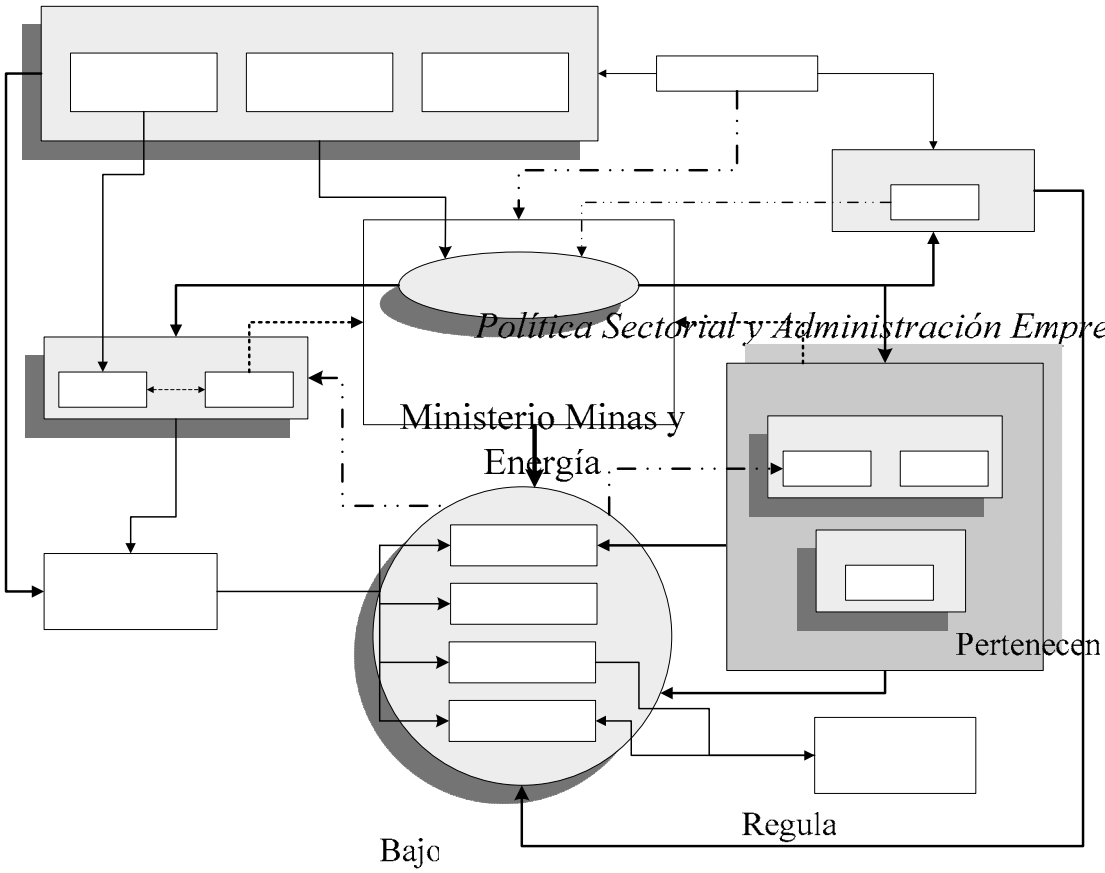

Figura 3.2. Estructura Institucional del Mercado Eléctrico de Colombia

3.1.2. Sistemas de información

En el área concerniente al mercado de energía se identifican cuatro niveles de información [4], el primero integrado por los agentes del mercado, el segundo nivel lo conforman las entidades vinculadas directamente al mercado de energía eléctrica (XM, CREG, UPME y SSPD), el tercer nivel son los sistemas públicos de información y por último la información dada por los usuarios.

La recopilación de información está a cargo de diferentes entidades que se dividen esta tarea de acuerdo con su área de interés y pueden verse a continuación algunos sistemas que posen información de mercado.

Sistema NEON: El XM recibe y almacena la información necesaria para la operación del sistema, la operación del despacho, la liquidación de los intercambios, y la liquidación de los cargos de transmisión. El sistema NEON está diseñado para realizar consultas puntuales de información y su acceso no es público. Pueden acceder los agentes del mercado (quienes pagan por la administración del mercado) y otras personas o entidades que realicen el pago correspondiente a la licencia de acceso.

ISA.COM: Es un boletín difundido vía Internet, en él se resume la evolución del mercado y de la operación del Sistema de Interconexión Nacional (SIN). Su actualización es semanal.

Sistema unificado de información: organizado por la SSPD, este sistema de información debe ser alimentado y mantenido por las empresas de servicios públicos y posee información relacionada con la actividad administrativa, financiera y operativa de las empresas prestadoras de servicios públicos.

Estadísticas del sector: La CREG ofrece en su página de Internet la información suministrada por las empresas a través del Instructivo Eléctrico; en esta página las empresas comercializadoras pueden acceder a información acerca de usuarios, consumo, valores facturados, etc. Esta información está totalmente ligada a la información entregada por las diferentes empresas, la cual no es del todo consistente ya que algunas veces no son actualizadas con la periodicidad necesaria de un mercado continuamente variable.

3.2. Variables propias del MEM

Para la determinación de las variables que se utilizarían para el análisis es esta investigación se tomó como referencia lo estipulado en el trabajo realizado por la UPME en el año 2004 titulado "Mercado de Energía Eléctrica en Colombia – Análisis comercial y de estrategias", acerca de la determinación de las variables estratégicas y no estratégicas en la fijación de los precios en bolsa.

De acuerdo al documento, las variables estratégicas están conformadas por los elementos que resultan explicativos de la formación del precio de oferta para los generadores dentro de los que se pueden mencionar los precios de los contratos, la generación del despacho ideal, el precio de bolsa, las ventas en bolsa, los ingresos por reconciliaciones, las inflexibilidades, los cambios de regulación, el embalse ofertable del sistema y los precios de otras plantas de la misma empresa.

Aunque el mencionado estudio abarca una buena porción de los elementos comprendidos en el MEM y es el punto de partida en la selección de las variables de la presente investigación, una revisión de las bases de datos con que cuenta el XM muestra que aún hay más componentes a tener en cuenta cuando se quiere estudiar el comportamiento de las centrales a la hora de fijar sus precios en bolsa.

En la sección 6.3. se detalla cuales fueron las variables que se emplearon en este trabajo y la forma como se llegó a esta selección; además, a manera de ampliación, el anexo A presenta una ampliación de las definiciones de las variables de acuerdo con el sistema de información NEON y con las resoluciones CREG relacionadas.

3.3. Bolsa de energía: Formación del precio en bolsa

Para realizar transacciones de energía eléctrica en Colombia uno de los principales medios utilizados es la "bolsa de energía" la cual funciona como medio para fijar el programa horario de generación y el precio de la energía, donde los oferentes son la centrales generadoras y quienes compran la energía son los comercializadores y distribuidores.

El precio se forma mediante un proceso que comienza diariamente cuando las centrales generadoras envían sus ofertas antes de las 8 a.m., compuestas por *un único precio y una disponibilidad* para todas las horas del día siguiente. Estas ofertas son ordenadas de menor a mayor en función del precio y determinan las plantas que atenderán la demanda. El ***precio de bolsa*** se determina el día posterior al despacho y es el precio del último recurso requerido para atender la demanda real para un despacho ideal de una hora para cada una de las 24 franjas horarias; esto es, se determina un precio de bolsa para cada hora, valor al cual se les paga a todos los agentes generadores presentes en el despacho.

Dado que el despacho ideal no tiene en cuenta las restricciones del sistema eléctrico ni eventualidades durante la operación, lo más probable es que sea diferente que el despacho real; a las diferencias entre el despacho ideal y el despacho real se les conocen como *reconciliaciones.* Una ampliación de estos términos se da en el Anexo A.

3.4. Curva de demanda residual

3.4.1.Conceptos

Las curvas de demanda residual se utilizan para modelar el efecto de las ofertas de la competencia dentro de cualquier mercado; para el caso del MEM, estas son útiles para comprender el efecto que tienen los precios que fijan las centrales generadoras diariamente, sobre el mercado y sobre el precio de bolsa.

Suponga que para el día siguiente, las empresas generadoras envían sus ofertas al operador del mercado especificando una disponibilidad y un precio, con lo que se construye para cada una de las horas de del día la curva de oferta agregada S(p), la cual permite definir el precio de bolsa para una demanda dada D como se ve en la figura 2.3.

Si en la construcción de dicha curva no se tiene en cuenta la oferta de la empresa M, se obtiene la curva de oferta agregada del resto de empresas llamada ($S_{-1}(p)$); la *curva de demanda residual* R(p) que enfrenta la empresa M para esa franja horaria está dada entonces por $R(p)=D- S_{-1}(p)$. Refiérase a las figuras 2.3 y 2.4.

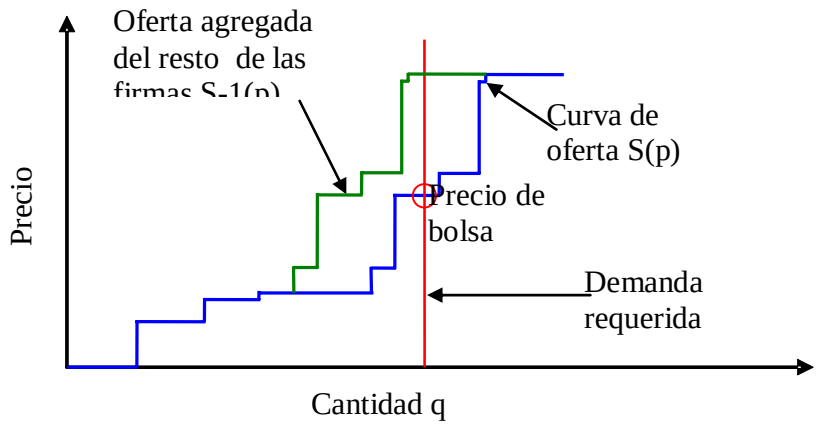

Figura 3.3. Demanda requerida en una franja horaria, Curva de oferta y oferta agregada de la competencia

En la figura 3.4 se grafica la función R(p), donde se muestra en el eje 'Y' el precio que puede lograr la empresa M y en el eje X, la cantidad máxima que el mercado estaría dispuesto a consumir al precio dado en Y.

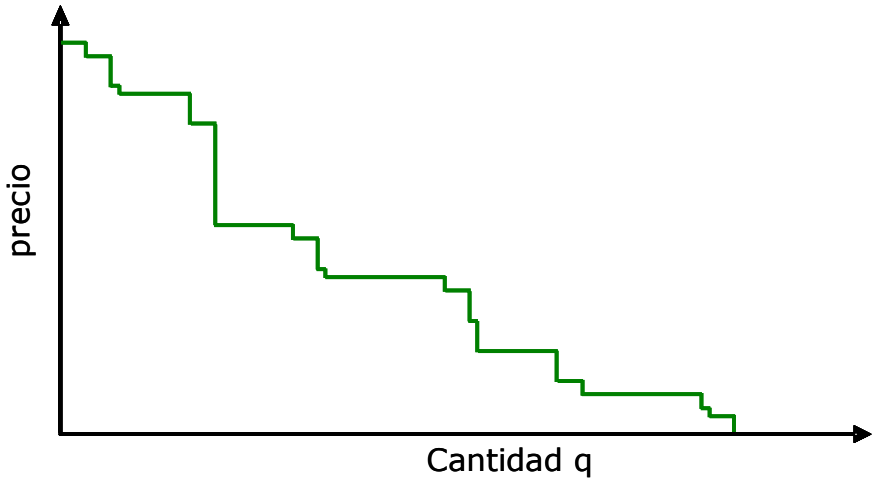

Figura 3.4. Curva de Demanda Residual

Es importante resaltar que si el mercado fuera de competencia perfecta, se obtendría una demanda residual cuya grafica sería una línea horizontal, ya que en un mercado de competencia perfecta, el precio es fijado por el mercado y no puede ser influenciado por la producción de una determinada empresa. Este hecho se puede expresar diciendo que una empresa en un mercado de competencia perfecta enfrenta una demanda residual infinitamente elástica.

El concepto de la demanda residual es algo semejante a la demanda que enfrenta el monopolio, en este caso la demanda sería la demanda residual y el monopolio, sería la empresa para la cual se calculo la demanda residual. Desde este punto de vista podemos obtener los ingresos que dicho monopolio puede tener, los cuales se pueden calcular con la siguiente formula: $Ingr(q)=R^{-1}(q)*q$.

3.4.2. Construcción de las curvas aplicadas a la formación del precio en la bolsa de energía

Para cada una de las plantas generadoras incluidas en el estudio debe realizarse el cálculo de la curva de demanda residual para una determinada franja horaria, esto debido a que es por cada hora que el operador del sistema hace el cálculo del despacho de energía.

El objetivo de usar la curva de demanda residual es brindar nuevos elementos para el análisis del mercado eléctrico mayorista, diferentes a las variables que clásicamente se tienen en cuenta. Se espera que los datos que se obtengan a partir de la curva favorezcan la descripción del comportamiento de las centrales al momento de hacer sus ofertas en la bolsa de energía.

Para la construcción de las curvas se requiere de las ***ofertas*** de la totalidad de las centrales que participan en la bolsa de energía, entendiéndose la oferta de una central como la declaración de una única disponibilidad en MW (ó KW) y un único precio para la energía en $/MWh (ó $/KWh), para cada uno de los días de estudio. Debido a que estos precios de oferta incluyen el costo equivalente de la energía (CEE[3]), a los precios de oferta se les sustrajo esta componente de acuerdo su valor diario; esta modificación aplica para todos los casos donde se utilicen estos precios durante toda la investigación.

Se necesita conocer las ***inflexibilidades*** que declararon las centrales para el espacio de tiempo seleccionado. Esta información se encuentra para cada hora de cada uno de los días, pero en la mayoría de los casos la inflexibilidad es igual para todas las horas en un mismo día, por lo que se hará uso del promedio diario de las inflexibilidades como forma de aproximación y simplificación del cálculo de la curva de demanda residual.

El otro insumo para la construcción de la curva es la ***proyección de demanda***[4] para el lapso de tiempo que se está estudiando. Los datos de dichas proyecciones están dados para cada hora de cada día; como la curva de demanda se construye para una hora específica, sería posible construir 24 curvas diarias por generador, lo que representaría una carga de cálculo muy alta siendo una opción poco práctica para la realización de los análisis; por esta razón se seleccionaron tres referencias en el día para la construcción de la curva: la demanda mínima, la demanda mediana y la demanda máxima de cada uno de los días, ya que se consideran puntos representativos del comportamiento de la demanda diaria.

Ahora, para calcular la curva de demanda residual de la central M para una franja horaria específica, primero se toman los datos correspondientes al precio y disponibilidad entregados por las restantes centrales generadoras para ese día, se ordenan los precios de menor a mayor y se construye la curva de oferta agregada de la competencia, $S_{-1}(p)$.

---

[3] Ver anexo A para conocer su definición.

[4] Se utilizará la proyección hecha por el CND.

Si se tienen N centrales con disponibilidad para el despacho, la curva de oferta agregada puede tener cuando más N escalones, en el caso de que ningún par de centrales hubiesen ofertado al mismo precio; en la práctica, la curva solo se calcula hasta que en un escalón se satisfaga la demanda correspondiente a la franja horaria, ya que allí es donde se determina el precio en bolsa. Esto puede verse en la figura 3.3.

Finalmente, la curva de demanda residual para una determinada central se obtiene de sustraer la curva de oferta agregada del resto de empresas ($S_{-1}(p)$) de la demanda de dicha hora, de modo que el punto de corte sobre el eje 'y' sería el precio de bolsa en el caso en el que la central M no hubiese participado en la bolsa durante esa franja horaria.

4. ANÁLISIS ESTADÍSTICO MULTIVARIADO: ANÁLISIS CLUSTER

En la mayoría de las ciencias empíricas los fenómenos observables son de naturaleza multivariada, es decir, no pueden ser descritos por un solo parámetro o variable. El análisis estadístico multivariado comprende todos los métodos utilizados para estudiar estos fenómenos en los que se analizan simultáneamente diferentes medidas de cada individuo sometido a investigación. En últimas, cualquier análisis simultáneo de dos o más variables, ya sea descriptivo, analítico o predictivo, se considera como análisis multivariado.

4.1. Conceptos básicos del análisis estadístico multivariado

En los siguientes numerales se enuncian algunos términos de la estadística multivarida debido a que serán empleados en las secciones siguientes, pero principalmente, porque estos conceptos son la base de las técnicas que se emplearon para el desarrollo de la investigación. Se busca que estos ítems sirvan de marco para el lector interesado en la investigación más no que sean un referente bibliográfico para quienes quieren profundizar en el análisis estadístico multivariado.

4.1.1. Vector de medias

La medida de tendencia central más común para describir datos multivariados es el vector de medias, el cual tiene una dimensión p, y cada elemento corresponded a la media de cada una de las p variables.

$$\overline{x_p} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \textbf{(4.1)}$$

$$\bar{x} = \begin{bmatrix} \overline{x_1} \\ \vdots \\ \overline{x_p} \end{bmatrix} \qquad \textbf{(4.2)}$$

Llevar la media al caso multivariado tiene ventaja sobre las medidas de tendencia central escalares que se basan en el orden de las observaciones, como el caso de la mediana, ya que estas no se pueden extrapolarse fácilmente al caso multivariado por la falta de un orden natural de los datos multivariados.

4.1.2. Covarianza

La covarianza es una medida de una dependencia entre variables aleatorias. Dadas dos variables X y Y la covarianza se define:

$$\sigma_{XY} = Cov(X,Y) = E(XY) - (EX)(EY) \qquad \textbf{(4.3)}$$

Si X y Y son independientes una respecto a la otra, la covarianza *Cov(X,Y)* es cero; la covarianza de una variable respecto a si misma corresponde a la varianza (σ_x).

Cuando se tienen dos o más variables, las diferentes covarianzas entre todas ellas pueden ser organizadas en una matriz, que se conoce como la matriz de covarianzas (S); esta es una matriz cuadrada simétrica cuya diagonal principal corresponde a las varianzas y los elementos fuera de esta corresponden a las covarianzas entre las variables.

$$S = \begin{bmatrix} s_1^{\,2} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s^{2}{}_{p} \end{bmatrix} \qquad (4.4)$$

4.1.3. Matriz de correlación

La interdependencia lineal entre las variables se mide con el coeficiente de correlación muestral r, el cual se debe analizar teniendo en cuenta su magnitud y signo así: valores cercanos a 1 o -1 indican que los puntos están a lo largo de una línea recta con pendiente positiva o negativa respectivamente, mientras que valores cercanos a cero indican que no hay relación lineal, sin descartar que exista otro tipo de relación [D.L, 2002, R.T1, 2005]. La correlación entre las variables x_i y la variable x_j se calcula como se muestra en la ecuación (4.5):

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix} \qquad \mathbf{(4.5)}$$

Para el caso en que las variables sean de magnitud distinta, estas se pueden transformar para obtener nuevas variables que si se puedan comparar, lo cual se consigue tipificando las variables, es decir restándoles su media y dividiéndolas por su desviación típica (normalizar).

4.2. El análisis de componentes principales (ACP)

4.2.1. Descripción

Este método permite establecer la interdependencia lineal entre las variables de análisis a partir de un coeficiente de correlación muestral (r). A partir de ACP se busca reconstruir el conjunto de datos multivariados (multidimensionales), reduciendo el número de variables al crear nuevas que agrupen, conservando la información, las variables originales que estén

altamente relacionadas. Estas nuevas variables se obtienen a partir de combinaciones lineales del conjunto de variables iniciales. La técnica de *Análisis de componentes principales* encuentra relaciones poco evidentes entre variables, agrupa los datos en nuevas variables denominadas *componentes principales* (CP).

Por lo general las dos primeras componentes principales contienen la mayor parte de la información, por lo tanto casi siempre se ubican los datos en estas nuevas componentes o ejes. El plano conformado por estos ejes recibe el nombre de plano factorial.

En síntesis, con el ACP se busca obtener información acerca de la interdependencia de las variables, reducir la multidimensionalidad de los datos y llevarlos a un plano de 2 o 3 dimensiones determinado por nuevas variables incorreladas.

4.2.2.Cálculo de las componentes principales:

Seguidamente se muestra un algoritmo para obtener las componentes principales de un conjunto de variables correlacionadas; Se sugiere trabajar con la matriz de correlaciones de las variables originales, en lugar de emplear la matriz de covarianzas con los datos normalizados.

1. Hallar la matriz de correlación de las variables que están describiendo un contexto.

2. Calcular los valores propios resolviendo la siguiente relación: $|(R - \lambda I)| = 0$

3. Ordenar de manera descendente los valores propios.

4. Los vectores propios asociados se calculan resolviendo: $(R - \lambda I)X = 0$

Donde:

R es la matriz de correlación de las variables originales.

I es la matriz identidad.

λ son los valores propios hallados al resolver la ecuación.

X es el vector propio generado por cada valor propio hallado.

Cada componente principal es la combinación lineal de las variables originales cuyos coeficientes son los valores del vector propio asociado a cada valor propio hallado.

4.2.3.Selección del número de componentes principales

Es difícil encontrar un criterio estrictamente formal para determinar el número de componentes principales (CP) a utilizar. El analista de los datos es quien define la cantidad de pérdida de variabilidad permitida.

Como se explicó anteriormente cada CP aporta una parte de variabilidad de los datos analizados, distribuyéndose en forma decreciente desde la primera hasta la última. La clave está en determinar que cantidad de variación explicada que

se considera satisfactoria, y se selecciona el número de componentes que cumplen con éste requisito.

La mejor forma de explicar el procedimiento anterior, es utilizar un diagrama de barras que muestra que cantidad de información contiene cada componente principal

4.3. Análisis de cluster o de conglomerados

Para agrupar datos, es posible utilizar un grupo de técnicas multivariantes denominado análisis "cluster", este análisis se basa en las características propias de cada uno de los elementos y se agrupan de acuerdo con su similitud, permitiendo que los objetos de cada conglomerado presenten un alto grado de homogeneidad y un alto grado de heterogeneidad con respecto a otros conglomerados; el grado de semejanza de los elementos a agrupar se puede medir de diferentes formas, pero lo principal es mantener la similitud "media" entre elementos de un conglomerado.

El análisis "cluster" es una herramienta útil en diferentes situaciones, incluso en las que se piense que los datos recopilados no tienen sentido por ser tan variables, precisamente este tipo de análisis clasifica los datos de manera que se facilite su manipulación, comprensión y permita una mejor visualización de los mismos; además el análisis "cluster" es útil cuando se desea separar grupos de datos de acuerdo con determinada característica y revelar resultados respecto a otras.

Este análisis es implementado como una técnica exploratoria o descriptiva y no como una técnica para realizar deducciones estadísticas para una población. Esa técnica es totalmente dependiente de las variables utilizadas como base y una adición o eliminación de variables afecta directamente a la solución dada para el grupo de datos de entrada.

4.3.1 Medidas de similitud

La similitud entre objetos es una medida del parecido existente entre los objetos que serán agrupados. En primer lugar se establecen las características que definen la similitud entre los datos y luego se hace combinaciones de las características para calcular una medida de similitud entre todos los pares de objetos, para por último agrupar objetos similares en conglomerados.

La similitud entre objetos puede medirse de varias formas pero el método predominante dentro de los utilizados en el análisis "cluster" se conoce como ***medidas de distancia.*** Estas representan la similitud como la proximidad de unas observaciones respecto a las otras, para las variables del conjunto que representa las características a comparar. Estas, son realmente medidas de diferencia, lo que implica que a valores elevados se tiene una menor similitud.

Existen varias medidas de distancia, pero las más utilizadas son: la distancia Euclídea que equivale a la longitud de la hipotenusa de un triángulo rectángulo; la distancia Euclídea al cuadrado o distancia Euclídea absoluta que es la suma de diferencias al cuadrado, sin tomar la raíz cuadrada lo que

acelera los cálculos y la distancia de Mahalanobis que es una medida de la distancia Euclídea que incorpora un procedimiento de estandarización de los datos.

Dentro de las medidas de distancia empleadas más a menudo existen otras que no se basan en la distancia Euclídea, para aclarar lo referente a este tema se resumen, en la tabla 2.1, las principales características de las distancias más usadas.

DISTANCIA	DEFINICIÓN
Euclídea	Longitud de la hipotenusa, de un triángulo rectángulo, en la que se ubican las coordenadas de los objetos que se desean medir. $D=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$ **(4.6)**
Euclídea cuadrada absoluta	Equivale a la distancia euclídea al cuadrado, es decir la suma de las diferencias al cuadrado sin tomar la raíz, lo que acelera los cálculos. $D=(x_2-x_1)^2+(y_2-y_1)^2$ **(4.7)**
Mahalanobis	Se emplea para medir la distancia entre una observación multivariada y el centro de la población a la cual pertenece. Se emplea cuando los datos están en diferentes escalas y se encuentran correlados.
Cityblock	Corresponde a la distancia calculada como la suma algebraica de la diferencia entre las coordenadas de los objetos que se desean medir. $D=\lvert x_1-x_2\rvert+\lvert y_1-y_2\rvert$ **(4.8)**
Correlación	Se calcula como uno menos la correlación entre los objetos que se están midiendo. Estos elementos son tomados como una secuencia de valores.
Chebychev	Se calcula la máxima diferencia entre coordenadas
Coseno	Equivale a uno menos el coseno del ángulo entre los puntos, tomándolos como vectores.

Tabla 4.1 Resumen medidas de similitud

4.3.2 Obtención de conglomerados

Los conglomerados pueden ser de dos tipos: jerárquicos o no jerárquicos. Dentro de los *procedimientos jerárquicos* (construcción de estructuras en forma de árbol) existen básicamente dos tipos, los de **aglomeración** y los **divisivos**; dentro de los de *aglomeración* cada objeto comienza dentro de su conglomerado y se comienza la combinación de los conglomerados por su proximidad con otros para formar nuevos conglomerados y de esta forma reducir el número total de conglomerados; dentro de los procedimientos *divisivos* se comienza con un gran conglomerado y en la medida en que se avanza, los datos más diferentes se dividen y forman nuevos conglomerados más pequeños y este proceso continúa hasta que cada dato es un conglomerado en sí mismo. Las estructuras en forma de árbol que se generan en los métodos jerárquicos son representadas en un esquema llamado *dendograma.*

Los más empleados en la práctica son los procedimientos de aglomeración, dentro de los que se pueden mencionar el *encadenamiento simple,* que se basa en la distancia mínima y lo que hace es encontrar el par de datos que se ubican a la distancia más corta y los agrupa en un solo conglomerado y de esta forma va adicionando elementos cercanos a los conglomerados para aumentar el tamaño de ellos; el proceso finaliza al obtener todos los datos dentro de un conglomerado. El problema de este proceso es que el primer y el último dato pueden estar separados por una distancia muy grande lo que implica que por lo menos dos elementos de un mismo conglomerado tendrán características muy diferentes.

También existe el *encadenamiento medio, el encadenamiento completo, el método de Ward y el método del centroide,* que se diferencian básicamente por sus criterios en la medición de las distancias (criterios de aglomeración) que se toman como referencia para agrupar los datos. Por ejemplo, el encadenamiento completo no toma la distancia mínima, si no la máxima entre los elementos de los conglomerados ya formados en una etapa anterior. Para clarificar la diferencia existente los métodos de agrupamiento se amplia, en la siguiente tabla, las características principales de los diferentes métodos de agrupamiento.

MÉTODOS DE AGRUPAMIENTO AGLOMERATIVOS	
NOMBRE	CARACTERÍSTICAS
ENCADENAMIENTO SIMPLE	Se agrupa en un mismo conglomerado los objetos separados por la distancia más corta. Los elementos de un conglomerado pueden formar largas cadenas y los individuos de los extremos podrían carecer de similitud.

ENCADENAMIENTO COMPLETO	El criterio de agrupamiento se basa en la máxima medida de la separación entre objetos, dicha distancia representa el diámetro de la esfera más reducida que incluye todos los elementos en ambos conglomerados. Elimina el problema generado con el encadenamiento simple.
ENCADENAMIENTO MEDIO (AVERAGE)	El criterio de aglomeración es la distancia media entre los individuos de un grupo con todos los individuos de otro. Combina los conglomerados con variaciones reducidas dentro del conglomerado
MÉTODO WARD	La distancia entre conglomerados es la suma de los cuadrados entre dos conglomerados sumados para todas las variables. Tiende a combinar los conglomerados con un número reducido de observaciones. Se obtienen mejores resultados si la medida de distancia empleada es la Euclídea.
MÉTODO CENTROIDE	La distancia entre dos conglomerados se determina por la distancia (Euclídea, por lo general) entre sus centroides o valores medios de las observaciones de las variables. Se ve menos afectado por los atípicos que otros métodos. Los centroides de los grupos cambian a medida que se fusionas los conglomerados.

Tabla 4.2 Métodos de agrupamiento aglomerativos

En los *procedimientos no jerárquicos* (no implica construcción de árboles) se asignan los objetos a los conglomerados después de establecer el número de conglomerados deseado; Los objetos, después de estar clasificados en un conglomerado, pueden ser reubicados en otro que se encuentre cercano. Estos procesos no jerarquizados generalmente parten de la selección de una o varias semillas como centros de los conglomerados y se denominan comúnmente, conglomeración de *K-medias*.

Dentro de los métodos utilizados para la obtención de este tipo de conglomerados se tienen los siguientes: *umbral secuencia, que* se inicia con la selección de una semilla y dentro de un radio especificado se ubican los demás elementos del conglomerado; después se toman nuevas semillas y se repite el proceso. *Umbral paralelo,* donde se seleccionan varias semillas al mismo tiempo y se asignan objetos que se ajusten a una distancia umbral asignada.

P*rocedimiento de optimización,* que funciona igual que los anteriores pero permite la reubicación de objetos a otro conglomerado más cercano que al que tiene asignado en ese momento.

5. MÁQUINAS DE SOPORTE VECTORIAL (MSV)

5.1. Generalidades

as máquinas de vectores soporte o MSV es una técnica enmarcada dentro de la inteligencia artificial, que hace parte de la minería de datos, por lo que busca la extracción de conocimiento a partir de información en bruto. Dentro de este ámbito las MSV cumplen tareas predictivas de clasificación al ser capaces de emitir una respuesta correcta (etiqueta de clasificación) ante una entrada similar a la usada en el entrenamiento. Una MSV cumple las condiciones de los métodos anticipativos o impacientes[5] ya que obtiene un modelo, optimizado globalmente, a partir de todos los ejemplos, es decir, se requiere un tiempo dedicado al entrenamiento (que suele ser grande), pero una vez entrenado el modelo, su aplicación en clasificación es instantánea. Una MSV posee ventajas sobre otros modelos complejos como las técnicas Bayesianas y las redes neuronales, también muy útiles en la solución de problemas reales, ya que no requiere ningún tipo de hipótesis sobre la densidad de probabilidad y son muy convenientes en problemas de alta dimensionalidad.

Las MSV son una consecuencia práctica de la teoría del aprendizaje, cuyo problema inicial fue planteado desde el siglo XVIII, pero sólo es hasta 1992 que se introduce la técnica y en 1995 se adiciona el margen débil, gracias al trabajo de Vapnik y sus colaboradores. Las MSV se caracterizan por el uso de las funciones Kernel y se basan en la idea del hiperplano de margen máximo.

5.2. Aprendizaje a través de ejemplos

Considere *n* datos de entrenamiento N dimensional (x_i) con su respectiva etiqueta (y_i).

$$x_i \in \Re^N \qquad y \qquad y_i \in \{+1,-1\} \qquad \textbf{(5.1)}$$

Con los cuales se busca obtener una función *f* tal que para una entrada en $\Re^N$ produzca una salida en {+1, -1}, para que así se pueda clasificar correctamente un nuevo dato (x, y).

5.3. Óptimo hiperplano clasificador

Los clasificadores de soporte vectorial están basados en hiperplanos que separan los datos de entrenamiento en dos subgrupos que posee cada uno una etiqueta propia, en medio de todos los posibles planos de separación entre las dos clases, etiquetadas $y_i \in \{+1,-1\}$, existe un único hiperplano óptimo de separación (OSH), de forma que la distancia entre el hiperplano óptimo y el patrón de entrenamiento más cercano sea máxima, con la intención de forzar

[5] Términos y definiciones tomadas de [Hernández, Ramírez, 2004]

la generalización de la máquina de aprendizaje [Burges, 1998]. En la figura 3 se aprecia el OSH así como el margen que es la distancia perpendicular entre los objetos más cercanos al hiperplano; dicho margen, es el que se busca maximizar.

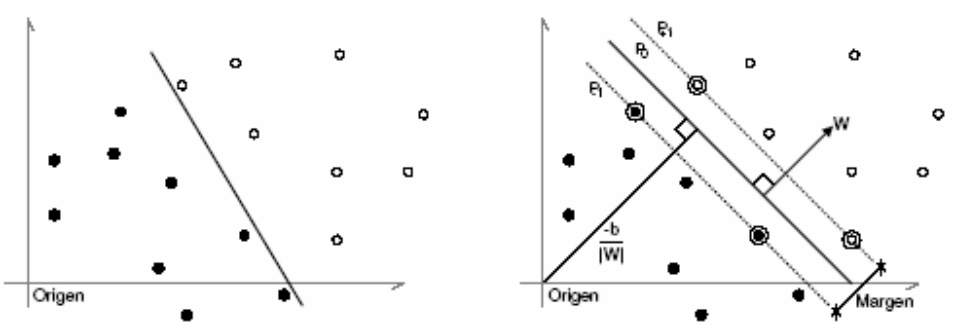

Figura 5.1 Hiperplanos que separan correctamente los datos. El OSH de la derecha tiene un mayor margen de separación entre clases, por lo tanto se espera una mejor generalización. Tomado de [Morales, Gómez, 2005]

El OSH se expresa de la forma:

$$g(\vec{x}) = (\vec{\omega} \cdot \vec{x}) + b = 0 \qquad \textbf{(5.2)}$$

Se define la función decisión como el signo que resulta de evaluar un dato en la función del OSH:

$$f_{\omega,b}(\vec{x}_i) = sign\left[g(\vec{x}_i)\right] = sign\left[(\vec{\omega} \cdot \vec{x}_i) + b\right] \qquad \textbf{(5.3)}$$

Para encontrar el OSH se debe maximizar el margen llegando a:

$$\min_{\omega,b} \frac{1}{2}(\vec{\omega} \cdot \vec{\omega})$$
$$sujeto \quad a \quad y_i(\vec{\omega} \cdot \vec{x}_i + b) \geq 1 \; \forall i \qquad \textbf{(5.4)}$$

Este es un *problema de optimización cuadrático* sujeto a restricciones, el cual es tratado mediante la introducción del método de *los multiplicadores de Lagrange*, $\alpha_i > 0$, uno por cada restricción.

Con esto, la ecuación del OSH y la función decisión pueden escribirse como:

$$g(\vec{x}) = \sum_{i=1}^{n}\left[\alpha_i y_i (\vec{x}_i \cdot \vec{x})\right] + b \qquad \textbf{(5.5)}$$

$$f(\vec{x}) = sign\left(\sum_{i=1}^{n}\left[\alpha_i y_i (\vec{x}_i \cdot \vec{x})\right] + b\right) \qquad \textbf{(5.6)}$$

5.4. Caso linealmente no separable: MSV con margen débil.

Lastimosamente, en la práctica no siempre es posible encontrar un hiperplano lineal separador, lo que puede deberse a datos erróneos, a ruido o solapamiento de clases en los datos de entrenamiento. Una solución sería buscar el hiperplano que conduzca al menor número de errores de entrenamiento, pero esto sería un problema combinatorial difícil de aproximar. [Cortes, Vapnik, 1995] para solucionar esto, se introducen unas *variables de relajación o variables slack*: $\varepsilon_i \geq 0,\ \forall i$

Entonces la forma de obtener el hiperplano clasificador óptimo con margen débil es minimizando la función:

$$\min_{\omega}\left[\frac{1}{2}(\vec{\omega}\cdot\vec{\omega})+C\sum_{i=1}^{n}\varepsilon_i\right] \qquad \textbf{(5.7)}$$

El parámetro *C* es elegido a priori por el usuario, de modo que un valor grande es una penalización alta a los errores.

Ahora, el principio de las MSV no lineales es el de mapear el espacio de entrada a un espacio de mayor dimensionalidad y con producto punto mediante una función no lineal ***(Φ)*** elegida a priori. Ver figura 5.2.

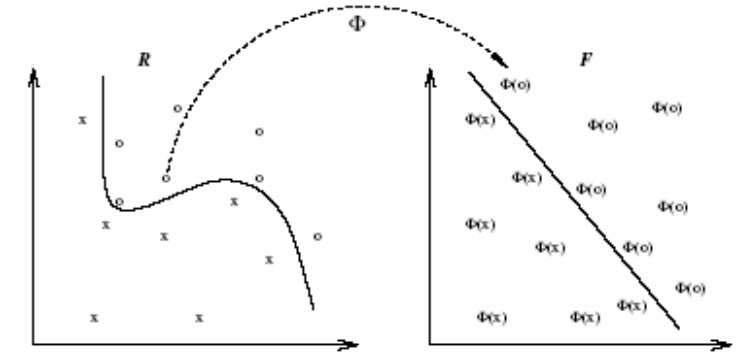

Figura 5.2 Transformación del espacio de entrada al espacio característico. Tomado de [Morales, Gómez, 2005].

La función del hiperplano clasificador se convierte en:

$$g(\vec{x})=\sum_{I=1}^{n}\left[\alpha_i y_i(\phi(\vec{x}_i)\cdot\phi(\vec{x}))\right]+b \qquad \textbf{(5.8)}$$

Se define entonces una función que sea el producto punto entre los vectores en el espacio característico:

$$k(\vec{x}_i,\vec{x})=\phi(\vec{x}_i)\cdot\phi(\vec{x}) \qquad \textbf{(5.9)}$$

Como el lado de la ecuación de alta dimensión, es costosa en términos computacionales, sin embargo, existe una ***función kernel (k)***, que puede evaluarse eficazmente, que ahorra la búsqueda explicita de la función λ y lleva directamente al resultado del producto punto, que es lo que realmente interesa. Los kernel más utilizados son:

Polinomial:

$$k(\vec{x},\vec{y}) = ((\vec{x})\cdot(\vec{y})+c)^d \;\; para\; c>0 \qquad \textbf{(5.10)}$$

Función de base radial (RBF):

$$k(\vec{x},\vec{y}) = e^{\left(-\frac{|\vec{x}-\vec{y}|^2}{2\sigma^2}\right)} \qquad \textbf{(5.11)}$$

Sigmoide:

$$k(\vec{x},\vec{y}) = \tanh(k(\vec{x},\vec{y})+\theta) \qquad \textbf{(5.12)}$$

Finalmente, la función objetivo y las restricciones se pueden escribir como:

$$\max_{\alpha}\left[\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k(\vec{x}_i,\vec{x}_j)\right] \qquad \textbf{(5.13)}$$

$$Sujeto\ a\ \ 0 \le \alpha_i \le C, \forall\ i\ y \sum_{i=1}^{n}\alpha_i y_i = 0$$

Y la ecuación del OSH y la función decisión como:

$$g(\vec{x}) = \sum_{i=1}^{n}\left[\alpha_i y_i k(\vec{x}_i,\vec{x})\right]+b \qquad \textbf{(5.14)}$$

$$f(\vec{x}) = sign\left(\sum_{i=1}^{n}\left[\alpha_i y_i k(\vec{x}_i,\vec{x})\right]+b\right) \qquad \textbf{(5.15)}$$

5.5. Multiclasificación con MSV

Todo lo tratado hasta ahora ha sido para el problema de biclasificación pero es sabido que en la vida real la mayoría de los problemas son de más de dos clases. Para solucionar este problema existen dos tipos de arquitecturas: *Máquinas multiclasificadoras SV* y *Máquinas biclasificadoras SV generalizadas.* Esta última ha probado tener una mayor simplicidad y un menor tiempo de respuesta.

Dentro de las Máquinas biclasificadoras existen dos principales arquitecturas de descomposición. Las *arquitecturas de descomposición estándar* (uno contra uno y uno contra el resto) y la *arquitectura de descomposición ECOC.* En este trabajo se utilizaran las arquitecturas estándar.

6. Mezclas Finitas

6.1. Mezclas Finitas

Las mezclas finitas o mezclas de distribuciones, es una técnica estadística que permite estimar la función de densidad de probabilidad de un conjunto de datos. Las mezclas representan dicha función como una suma ponderada finita de las componentes de densidad multivariables. [26]

La ecuación (6.1) expresa matemáticamente lo que hasta aquí se dijo:

$$f(x \mid \pi, \theta) = \sum_{k=1}^{c} \pi_k g_k(x \mid \theta_k) \qquad \textbf{(6.1)}$$

La mezcla sobre una población *X d-dimensional* compuesta por *'n'* observaciones contiene *'c'* componentes $g_k(x|\theta_k)$. θ_k es usado para denotar cualquier tipo o número de parámetros. Los pesos están dados por π_k, restringidos a que tienen que ser positivos y de suma uno; estos pesos son llamados comúnmente proporciones de mezcla o coeficientes de mezcla.

Como se ve debe conocerse el número de componentes de mezcla pero así mismo los estimadores, ya que son ellos quienes definen el comportamiento de los datos dentro del grupo y la forma de este. En el caso de la distribución normal los estimadores son el vector de medias y la matriz de covarianza. Esto nos lleva a reescribir la ecuación (6.1) de la siguiente manera:

$$f(x \mid \pi, \hat{\mu}_k, \hat{S}_k) = \sum_{k=1}^{c} \pi_k \phi(x \mid \hat{\mu}_k, \hat{S}_k) \qquad \textbf{(6.2)}$$

Donde Φ representa la función de densidad de probabilidad normal multivariada. Por esta razón la ecuación representa las mezclas finitas multivariadas Gausianas.

Con esto queda determinado cuales son los parámetros que deben estimarse a partir del grupo de datos a analizar: Las proporciones de mezcla, los vectores de medias para cada término o componente y la matriz de covarianzas.

El método más común para estimar los parámetros de la mezcla finita es el algoritmo EM (Expectation-Maximimization Algorithm), el cual se basa en la estimación de máxima similitud planteado por Dempster, Laird y Rubin en 1977. En el ítem a continuación se describen los conceptos principales del algoritmo. [26]

6.1.1. Algoritmo EA

Para aplicar la metodología deben tenerse en cuenta algunos aspectos:

Se deben especificar el número de componentes.

El algoritmo es iterativo así que se necesita tener una estimación inicial de los parámetros para empezar

Se debe asumir alguna forma para las densidades de las componentes.

Las técnicas de cluster proveen un marco apropiado para atacar estos problemas, de allí que es usual encontrar estas técnicas juntas como parte de una metodología. Sin embargo, la aplicación del algoritmo a partir de cluster dependerá de los resultados de esta última, esto es, lo estimadores iniciales se tomaran de cluster si con ella se logran obtener grupos definidos y precios característicos. De no ser así, los grupos iniciales serán aquellos formados por las diferentes etiquetas de precios conformadas.

Para el desarrollo del algoritmo, se sabe que se necesitan estimar los parámetros de proporción, de medias y de covarianzas de los componentes, esto es, se desea estimar: $\theta = \pi_1,...,\pi_{c,}\mu_1,...,\mu_c,S_1,...,S_c$

Usando el criterio de máxima similitud, se tiene que debe maximizarse:

$$L(\theta \mid \bar{x}_1,...,\bar{x}_n) = \sum_{i=1}^{n} \ln\left[\sum_{k=1}^{c} \pi_k \phi(\bar{x}_i \mid \mu_k, S_k)\right] \qquad \textbf{(6.3)}$$

Se asume que las componentes existen en una proporción fija dentro de la mezcla y está dada por π_k. Entonces tiene sentido calcular la probabilidad a posteriori que un caso x_i, pertenezca a alguna de las componentes de densidad. Es por esta pertenencia desconocida que debemos usar un método como el algoritmo EM para maximizar la ecuación (6.3). La probabilidad en mención esta dada por la ecuación (6.4):

$$\hat{\tau}_{ik}(x_i) = \frac{\hat{\pi}_k \phi(\bar{x}_i \mid \hat{\mu}_k, \hat{S}_k)}{\hat{f}(x_i \mid \hat{\pi}_k, \hat{\mu}_k, \hat{S}_k)}; \qquad \textbf{(6.4)}$$

$$k = 1,2,...,c; i = 1,2,..,n;$$

Donde:

$$\hat{f}(x_i \mid \hat{\pi}_k, \hat{\mu}_k, \hat{S}_k) = \sum_{k=1}^{c} \hat{\pi}_k \phi(x_i \mid \hat{\mu}_k, \hat{S}_k); \qquad \textbf{(6.5)}$$

Para maximizar la función dada en la ecuación (6.3) se debe calcular la primera derivada parcial respecto a los parámetros y después igualar a cero. La solución de este problema esta dada por las siguientes ecuaciones:

$$\hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n} \hat{\tau}_{ik} \qquad \textbf{(6.6)}$$

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{\tau}_{ik}\bar{x}_i}{\hat{\pi}_k} \qquad \textbf{(6.7)}$$

$$\hat{S}_k = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\tau}_{ik}(\bar{x}_i - \hat{\mu}_k)(\bar{x}_i - \hat{\mu}_k)^T}{\hat{\pi}_k} \qquad \textbf{(6.8)}$$

La ecuación para actualizar las matrices de covarianza dada en la ecuación 6.8 es para el caso más general que puede darse de acuerdo a las restricciones de la matriz de covarianzas, el cual es quien define la forma de las densidades; este caso general, es el caso sin restricciones, por eso es el que se presenta aquí. Para conocer con más detalle este tema, por favor refiérase a [25] en su sección 6.2.2.

Teniendo el soporte matemático para desarrollar el algoritmo, a continuación se enuncian los pasos que deben seguirse:

- Determinar el número de componentes '*c*' o términos de la mezcla
- Establecer una estimación inicial de los parámetros de mezcla: Las proporciones de mezcla, las medias y matrices de covarianza.
- Para cada uno de los casos x_i calcule la probabilidad a posteriori con la ecuación (6.4) usando los valores actuales de los parámetros.
- Actualizar los parámetros de mezcla para cada uno de los componentes usando las ecuaciones (6.6) a (6.8).
- Repetir los pasos 3 y 4 hasta que los estimados de los parámetros converjan.

El paso 5 es implementado iterativamente hasta que los cambios entre los estimados de los parámetros en cada iteración son menores que un valor de tolerancia preestablecido por el investigador. Note como el Algoritmo EM usa todo el grupo de datos para actualizar simultáneamente los parámetros en cada paso de iteración.

7. METODOLOGÍA PROPUESTA

Para lograr el cumplimiento de los objetivos planteados se plantearon una serie etapas para procesar la gran cantidad de datos con las que se contaba al inicio de la investigación. Se parte seleccionando un grupo de centrales importantes para el mercado bajo ciertos criterios expuestos adelante. Se determinaron las variables que entrarían a analizarse para cada una de esas centrales y se le realizó una adecuación de dichas variables con el fin de que favorecieran la implementación de las técnicas principales que fueron establecidas para llevar a cabo este trabajo. En las siguientes secciones se explica en detalle los pasos llevados a cabo.

7.1. Selección de centrales

Para acotar el número de centrales generadoras empleadas en esta investigación se tuvieron en cuenta dos criterios. Inicialmente se analizó uno de los indicadores del comportamiento del MEM establecido por la SSPD denominado número de coincidencias del precio de oferta con el precio en bolsa. Dicho indicador, como su nombre lo revela, determina el número de veces que coincide el precio ofertado por un agente generador con el precio de bolsa, estos resultados se tienen para cada hora y cada agente. Este indicador se calcula, para cada hora, como el número de coincidencias del agente sobre el número de coincidencias en el mes para esa hora.

El análisis de este indicador se realizó con datos de Julio de 2004 a Enero de 2006, estos son el total de datos suministrados por la Superintendencia de Servicios Públicos Domiciliarios y se encuentran disponibles en: http://www.superservicios.gov.co/energiagas/energia_ind_comp_mem.htm

Basados en el análisis de los resultados obtenidos para este indicador en el periodo de tiempo ya mencionado se tomo como segundo criterio, para la selección de las centrales, la capacidad de generación. De este modo luego de tener preseleccionadas las plantas generadoras con mayor cantidad de coincidencias del precio de oferta con el precio de bolsa se realizó una selección final y definitiva de 6 centrales hidráulicas y 6 térmicas que tienen mayor capacidad, los resultados del proceso anterior se resume en la siguiente tabla.

Empresa	Generador	Coincidencias	Capacidad [MW]	
ISAGEN	San Carlos	4362 (26,76%)	1240	✓
EMGESA	Guavio	3036 (18,62%)	1150	✓
CHIVOR	Chivor	1493 (9,16%)	1000	✓

EEPPM	Guatrón	1247 (7,65%)	512	✓
EEPPM	Porce II	854 (5.24%)	405	✓
EMGESA	Paraíso-Guaca	709 (4,35%)	276	
EPSA	Alto y bajo anchicayá	703 (4,31%)	439	✓
CEN. HID. BET.	Betania	644	540	
EEPPM	La tasajera	635	306	
CORELCA	Tebsa- total	313 (1,92%)	890	✓
EEPPM	Guatapé	308	560	
EEPPM	Playas	256	204	
EPSA	Calima	247	132	
EGETSA	Prado 4	195		
EGETSA	Prado	195		
GEST. ENERG.	Paipa 4	168 (1,03%)	150	✓
EEPPM	Riogrande I	145		
ISAGEN	Jaguas	136	170	
EEPPM	Miel I	127	396	
TERMOTASAJERO	Tasajero I	103 (0.63%)	163	✓
CHEC	Termodorada	93 (0,57%)	52	
CORELCA	Termoflores	74 (0,45%)	150	✓
URRA	Urrá	58	340	
TERMOFLORES	Termoflores 3	29 (0,18%)	150	✓
GEST. ENERG.	Paipa 2	28	68	
TERMOFLORES	Termoflores 2	24	99	

TERMOYOPAL	Termoyopal 2	23	30	
GEST. ENERG.	Paipa 3	23	68	
EMGESA	Zipa ISA 4	14		
EEPPM	La vuelta (planta menor)	13	19.8	
GEST. ENERG.	Paipa 1	13	28	
EPSA	Salvajina	11	285	
EMGESA	Zipa ISA 5	8		
CORELCA	Termoguajira 1	4	151	
ISAGEN	Termocentro 1	4 (0,02%)	300	✓
CORELCA	Termobarranquilla 4	3		
PROELÉCTRICA	Proeléctrica 1	2		
CORELCA	Termoguajira 2	1	151	

Tabla 7.1 Coincidencias y capacidad de la centrales generadoras colombianas

En la tabla 6.1 se observa que en la primera columna muestra la empresa generadora, la segunda indica la empresa generadora de la correspondiente empresa (amarillo para las plantas térmicas y azul para las plantas hidráulicas), la tercera columna muestra el número de coincidencias de la planta generadora, la cuarta columna indica la capacidad de la planta y en la quinta y última columna se señalan las plantas generadoras seleccionadas para el estudio en cuestión. Así pues si se tienen dos centrales que tengan aproximadamente igual número de coincidencias pero una es de mayor capacidad que la otra, se tuvo en cuenta la de mayor capacidad.

Finalmente las plantas generadoras seleccionadas definitivamente para el estudio se enlistan a continuación: Tebsa, Tasajero1, Paipa 4, Termoflores, Termoflores 3, Termocentro1, San Carlos, Guavio, Chivor, Guatrón, Porce II, Alto y bajo Anchicayá (Alban).

7.2. Cálculo de las curvas de demanda residual

Como se dijo en el capítulo 3 para cada una de las centrales de estudio se construyeron tres curvas para cada uno de los días del período analizado: las franjas de demanda mínima, mediana y máxima. Lo primero entonces es encontrar las demandas respectivas para estas franjas.

Como lo que se va a realizar posteriormente es un análisis de datos, de las curvas deben extraerse determinados puntos que puedan incluirse como variables o descriptores, que sirvan de entradas de las técnicas que se emplearan. Se buscó entonces cuales serían esos puntos que representen de mejor forma la esencia de las curvas de demanda residual.

Los puntos se seleccionados se llamaron P_x, P_m y Pdif; P_m corresponde al precio máximo alcanzado para cumplir con la demanda pronosticada si la central de estudio no hiciera casación, y se le denominó "Punto de mínima influencia"; P_x es el "punto de máxima influencia" e indica el precio al que debería ofertar la empresa generadora, para la cual se construyó la curva de demanda residual, para garantizar que toda su disponibilidad saliese despachada. Pdif, es la diferencia entre Pm y Px (Pdif = Pm – Px). En la figura 7.1 se puede apreciar esto claramente.

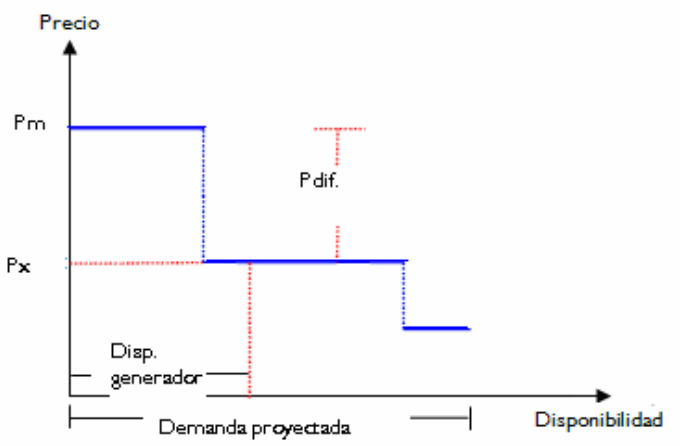

Figura 7.1 Representación gráfica variables que se extraen de la CDR

En la figura 7.2 se puede apreciar cómo de la curva de oferta agregada pueden obtenerse los mismos puntos; conocido esto y debido a que para un mismo día la única diferencia para la construcción de las tres curvas es la demanda de cada una de esas tres franjas, es mucho más práctico obtener los puntos a partir de la curva de oferta agregada (que de la curva de demanda residual), ya que estos pueden extraerse de los cortes de esta curva con tres verticales correspondientes a las demandas mínima, mediana y máxima del día. vea la figura 6.3

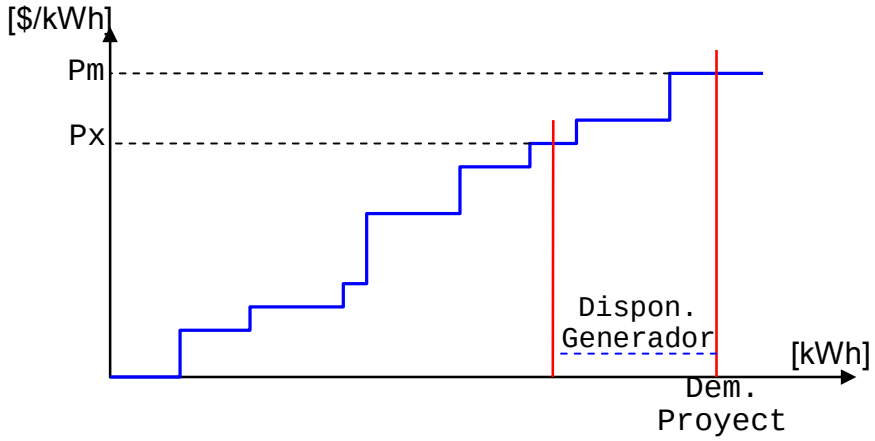

Figura 7.2 Curva de Oferta agregada donde se muestran los puntos representativos que se usaran como descriptores o variables, iguales a los obtenidos de la CDR.

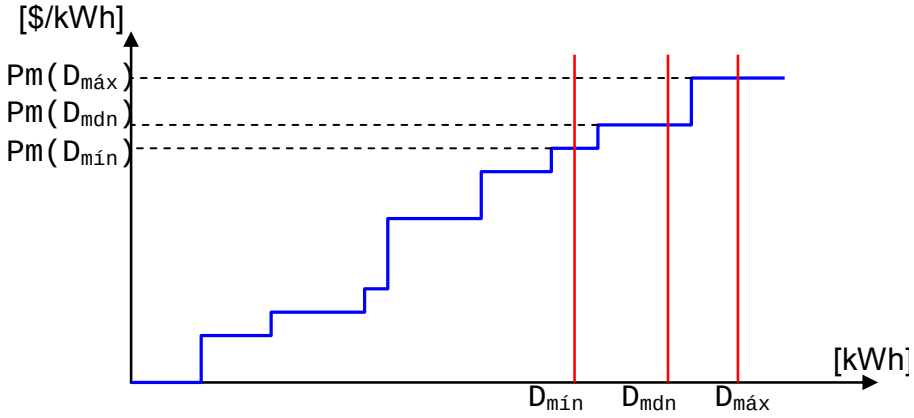

Figura 7.3 Curva de Oferta agregada. Se señalan los cortes de las diferentes demandas con la curva, para obtener las variables de la CDR

Para llevar a cabo estos cálculos se creó un programa en Matlab, el cual permite obtener todos los puntos para una central para todo el periodo de análisis. Las entradas son las explicadas en el capítulo 3, claro está, teniendo en cuenta que para cada día se tienen tres proyecciones de demanda.

Un detalle que debe aclararse ya que en ciertos casos puede afectar los resultados, es el de la forma como se obtiene el primer escalón de la curva de oferta agregada; este debe componerse por la suma de las inflexibilidades del día anterior[1] para las centrales diferentes a la de análisis más la inflexibilidad

[1] Se usa la del día anterior debido a que en un escenario real los datos actuales de las otras centrales serían desconocidos.

propia para el día[1] de la franja horaria que se está calculando. Ello conlleva a deducir que las disponibilidades (de cada central) usadas en los cálculos son las del día anterior[2] menos la inflexibilidad de ese día.

7.3. Selección de variables

La preselección inicial de las variables se realizó con base en el listado de variables estratégicas y no estratégicas que según el trabajo de 2004 realizado por la UPME, llamado "Una visión del mercado eléctrico colombiano", intervienen en la fijación de los precios en bolsa basados en modelos estadísticos, además se tuvieron en cuenta otras variables extraídas de la base de datos del XM, se incluyó el precio del petróleo y naturalmente se agregaron las seis (6) variables extraídas de la curva de demanda residual.

A continuación se resumen las variables que finalmente fueron a analizadas con correlación y ranking.

VARIABLE	IDENTIFICACIÓN
Embalse	Emb
Embalse agregado	EmbAgr.
Reconciliación negativa energía	RecNeg Energ
Reconciliación negativa pesos	RecNeg Pesos
Reconciliación positiva energía	RecPos Energ
Reconciliación positiva pesos	RecPos Pesos
Contratos energía	Cont.Energ
Contratos [$/kW]	Cont[$/kW]
Pm, Demanda máxima	Pm.Máx
Pm, Demanda mediana	Pm.Med
Pm, Demanda mínima	Pm.Mín

[1] 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 propia inflexibilidad para ese día.

[2] Se usa la del día anterior debido a que en un escenario real los datos actuales de las otras centrales serían desconocidos.

Px, Demanda máxima	Px.Máx
Px, Demanda mediana	Px.Med
Px, Demanda mínima	Px.Mín
WTI	WTI

Tabla 7.2 Variables preseleccionadas

Dentro de las variables que hacen parte de esta investigación hay algunas como: el precio de oferta, reconciliaciones positiva y negativa (pesos),ventas contratos, WTI y los insumos requeridos para el cálculo de la curva de demanda residual (ver sección 7.2), que tienen relación directa con el valor del dinero, por lo tanto ven modificado su valor con el paso del tiempo, por lo tanto fue necesario realizar una corrección de estos datos utilizando el índice de precios del productor (IPP) lo que permite llevarlos a un mismo horizonte en el tiempo.

Para comparar las variables, que tienen relación directa con las cifras monetarias, correspondientes a diferentes momentos del tiempo teniendo en cuenta el IPP se debe usar la expresión mostrada a continuación:

$$Vlr.hoy = Vlr.ayer * \frac{IPP\ hoy}{IPP\ ayer} \qquad \textbf{(7.1)}$$

Por otra parte considerando que algunas variables suministran información redundante, se utilizan dos técnicas (información mutua y correlación) que de forma complementaria permitan filtrar las variables, teniendo en cuenta aquellas que posean una mayor relación con el precio de oferta, variable que fue tomada como referencia.

A la variable correspondiente al precio del petróleo no le fue realizado el proceso denominado información mutua, únicamente se buscó la correlación que dicha variable tenía con el precio de oferta, por lo que no se encuentra ubicada en las tablas correspondientes al ranking y fue incluida en el conjunto de variables de aquellas centrales en las que presentó una alta correlación.

7.3.1.Correlación

Con este método se busca establecer la interdependencia lineal existente entre las variables seleccionadas. Este coeficiente puede tomar valores entre -1 y 1, teniendo en cuenta que los valores cercanos a los extremos de este rango indican que los puntos de la variable analizada se encuentran en una línea recta respecto a la variable de referencia, por lo tanto una mayor semejanza, y aquellos valores cercanos a cero indican la carencia de relación lineal entre las variables.

En el empleo de este método se tienen en cuenta los valores absolutos de los coeficientes, lo que permite considerar como indicador de alta correlación los resultados cercanos a 1; las correlaciones de las variables se organizaron en

las tablas 7.3 y 7.4 donde se aprecian las centrales hidráulicas y térmicas respectivamente.

Es necesario aclarar que la correlación, al igual que el ranking, se realiza entre cada una de las variables respecto al precio de oferta, ya que finalmente esta es la variable tomada como referencia en la selección de las variables de interés para esta investigación.

CORRELACIÓN						
POSICIÓN	San Carlos		Guavio		Chivor	
	Var	Coef.	Var	Coef.	Var	Coef.
1	Pm.Mín	0.7346	Pm.Mín	0.2548	Pm.Mín	0.2606
2	Px.Med	0.6483	Px.Med	0.2451	Px.Med	0.2118
3	Px.Mín	0.3774	RecNeg Energ	0.216	RecPos Energ	0.1934
4	RecNeg Energ	0.3524	RecPos Pesos	0.199	RecPos Pesos	0.1766
5	EmbAgr	0.349	RecNeg Pesos	0.1668	Px.Mín	0.1751
6	Emb	0.3452	RecPos Energ	0.1608	Emb	0.1389
7	Cont.Energ	0.2879	EmbAgr	0.1591	Px.Máx	01356
8	RecPos Pesos	0.2661	Emb	0.1467	Cont.Energ	0.1244
9	RecNeg Pesos	0.2591	Px.Mín	0.1345	Pm.Med	0.1143
10	Pm.Med	0.2339	Pm.Med	0.1107	EmbAgr	0.0853
11	RecPos Energ	0.2305	Cont[$/kW]	0.0294	Pm.Máx	0.0781
12	Cont[$/kW]	0.1681	Px.Máx	0.0133	RecNeg Energ	0.0388
13	Pm.Máx	0.0838	Cont.Energ	0.0064	RecNeg Pesos	0.0282

14	Px.Máx	0.0062	Pm.Máx	0.0016	Cont[$/kW]	0.0063
POSICIÓN	Guatrón		Porce II		Alban	
	Var	Coef.	Var	Coef.	Var	Coef.
1	Pm.Mín	0.4477	Emb	0.3225	Emb	0.3636
2	Px.Med	0.4218	Cont[$/kW]	0.1955	Cont[$/kW]	0.1738
3	Px.Mín	0.3901	RecNeg Energ	0.177	RecNeg Energ	0.1372
4	Pm.Med	0.3798	RecNeg Pesos	0.1518	RecNeg Pesos	0.1337
5	Emb	0.2504	Px.Máx	0.1315	EmbAgr	0.1243
6	EmbAgr	0.2052	Pm.Máx	0.0835	Px.Med	0.1098
7	RecPos Pesos	0.1759	Cont.Energ	0.071	Cont.Energ	0.1075
8	RecPos Energ	0.1448	Pm.Med	0.0706	Pm.Med	0.1032
9	RecNeg Energ	0.1348	Px.Med	0.0594	Pm.Mín	0.0971
10	RecNeg Pesos	0.0806	RecPos Pesos	0.0358	Px.Máx	0.0384
11	Cont.Energ	0.0744	Pm.Mín	0.0255	Pm.Máx	0.0346
12	Px.Máx	0.0528	Px.Mín	0.0226	Px.Mín	0.0305
13	Pm.Máx	0.0184	RecPos Energ	0.0107	RecPos Energ	0.0214
14	Cont[$/kW]	0.0124	EmbAgr	0.002	RecPos Pesos	0.0002

Tabla 7.3 Correlación centrales hidráulicas

CORRELACIÓN

POSICIÓN	Paipa IV		Tasajero 1		Centro 1	
	Var	Coef.	Var	Coef.	Var	Coef.
1	Px.Mín	0.3268	Pm.Mín	0.4275	Px.Mín	0.374
2	WTI	0.2921	WTI	0.3432	Pm.Med	0.3265
3	Cont.Energ	0.1776	Cont[$/kW]	0.2121	Cont.Energ	0.3234
4	Pm.Med	0.1682	Px.Med	0.2025	WTI	0.2995
5	RecPos Pesos	0.1179	EmbAgr.	0.1812	EmbAgr	0.2093
6	Cont[$/kW]	0.097	Pm.Máx	0.1593	Px.Máx	0.1366
7	EmbAgr.	0.0933	RecPos Energ	0.1552	Pm.Máx	0.121
8	Pm.Mín	0.0874	Px.Mín	0.1485	Px.Med	0.1178
9	Pm.Máx	0.0653	Cont.Energ	0.1322	Pm.Mín	0.0588
10	Px.Máx	0.0633	Px.Máx	0.1228	RecNeg Pesos	0.0585
11	RecNeg Pesos	0.0378	RecPos Pesos	0.1009	RecNeg Energ	0.0439
12	RecNeg Energ	0.0238	RecNeg Pesos	0.0737	Cont[$/kW]	0.0353
13	Px.Med	0.0174	Pm.Med	0.0691	RecPos Pesos	0.0185
14	RecPos Energ	0.0037	RecNeg Energ	0.0472	RecPos Energ	0.0107
POSICIÓN	Tebsa		Flores		Flores 3	
	Var	Coef.	Var	Coef.	Var	Coef.
1	RecPos Pesos	0.4922	RecPos Energ	0.1832	RecPos Pesos	0.7353
2	RecPos Energ	0.4481	RecPos Pesos	0.1788	RecPos Energ	0.7343

3	Pm.Mín	0.3909	Pm.Med	0.1346	EmbAgr	0.1822
4	Pm.Máx	0.3796	Cont.Energ	0.1318	WTI	0.1737
5	Cont[$/kW]	0.3786	Cont[$/kW]	0.1263	RecNeg Energ	0.1368
6	RecNeg Energ	0.3647	Pm.Mín	0.1208	RecNeg Pesos	0.1263
7	RecNeg Pesos	0.3581	RecNeg Energ	0.1164	Cont.Energ	0.0868
8	Px.Mín	0.3225	RecNeg Pesos	0.1121	Pm.Med	0.0827
9	WTI	-0.3033	WTI	-0.0738	Cont[$/kW]	0.0074
10	Px.Máx	0.2509	Pm.Máx	0.0691	Pm.Mín	0.006
11	EmbAgr.	0.1772	Px.Máx	0.058	Px.Máx	0.0033
12	Pm.Med	0.1503	EmbAgr	0.0112	Pm.Máx	0.0008
13	Px.Med	0.1007				
14	Cont.Energ	0.0489				

Tabla 7.4 Correlación centrales térmicas

Con este criterio se eliminaron las variables cuyo valor de correlación respecto al precio de oferta fuera cercano a cero. Para decidir finalmente cuales variables se emplearían se realizó un ranking variables usando la técnica "información mutua".

La variable llamada WTI fue adicionada al conjunto inicial de variables posterior a la realización del ranking y por ende para analizar su efecto para cada una de las centrales, únicamente se midió su correlación con el precio y con base en este resultados se incluyó o descartó del conjunto final de variables, el cual se muestra en la siguiente sección.

7.3.2. Ranking de variables (información mutua)

Para establecer las variables, dentro del grupo preseleccionado, que describen mejor el precio de oferta se realizó ***ranking univariado,*** éste sólo tiene en cuenta la relación entre la variable analizada y la variable de referencia, para este caso el precio de oferta.

El método asigna a cada variable un coeficiente que permite determinar la cantidad de incertidumbre que el conocimiento de dicha variable es capaz de despejar con respecto al estado en el que se encuentre la de referencia.

Las variables se ordenan en función de estos coeficientes y se seleccionan aquellas con más altos valores. En la tabla 7.5 se muestra el ranking de las variables para las centrales hidráulicas y en la tabla 7.6 está el ranking de las variables para las centrales térmicas.

RANKING VARIABLES						
POSICIÓN	San Carlos		Guavio		Chivor	
	Var	Coef.	Var	Coef.	Var	Coef.
1	Pm.Mín	0.4531	EmbAgr	0.4469	Cont[$/kW]	0.3902
2	Px.Med	0.3896	Pm.Mín	0.3794	EmbAgr	0.3418
3	EmbAgr	0.3543	Emb	0.3713	Emb	0.3130
4	Px.Mín	0.3460	Px.Med	0.3522	Pm.Mín	0.2914
5	Pm.Med	0.2903	Px.Mín	0.3314	Cont.Energ	0.2896
6	Px.Máx	0.2410	Cont[$/kW]	0.3098	Px.Med	0.2863
7	Cont.Energ	0.2409	Pm.Med	0.2867	Px.Mín	0.2855
8	Cont[$/kW]	0.2282	Cont.Energ	0.2605	Pm.Med	0.2414
9	Emb	0.2191	Px.Máx	0.2438	Px.Máx	0.2199
10	Pm.Máx	0.1973	Pm.Máx	0.2141	Pm.Máx	0.1944
11	RecNeg Energ	0.1890	RecNeg Energ	0.1674	RecPos Energ	0.1562
12	RecNeg Pesos	0.1706	RecNeg Pesos	0.1545	RecPos Pesos	0.1547
13	RecPos Pesos	0.1424	RecPos Energ	0.1237	RecNeg Pesos	0.1380
14	RecPos Energ	0.1366	RecPos Pesos	0.1229	RecNeg Energ	0.1345

POS	Guatrón		Porce II		Alban	
	Var	Coef.	Var	Coef.	Var	Coef.
1	Pm.Mín	0.4302	Pm.Mín	0.2538	EmbAgr	0.2420
2	Px.Mín	0.3996	Px.Med	0.2501	Emb	0.2212
3	EmbAgr	0.3933	Px.Mín	0.2428	Pm.Med	0.1919
4	Px.Med	0.3886	EmbAgr	0.2421	Px.Med	0.1895
5	Cont[$/kW]	0.3560	Pm.Med	0.2338	Pm.Mín	0.1870
6	Pm.Med	0.3362	Cont[$/kW]	0.2258	Px.Mín	0.1808
7	Px.Máx	0.2538	Px.Máx	0.1639	Cont[$/kW]	0.1801
8	Cont.Energ	0.2235	Emb	0.1557	Cont.Energ	0.1714
9	Emb	0.2166	Pm.Máx	0.1429	Px.Máx	0.1577
10	Pm.Máx	0.2140	Cont.Energ	0.1294	Pm.Máx	0.1359
11	RecNeg Energ	0.1318	RecNeg Pesos	0.1120	RecNeg Pesos	0.1018
12	RecNeg Pesos	0.1283	RecNeg Energ	0.1095	RecNeg Energ	0.0955
13	RecPos Energ	0.1155	RecPos Pesos	0.0701	RecPos Pesos	0.0759
14	RecPos Pesos	0.1111	RecPos Energ	0.0606	RecPos Energ	0.0692

Tabla 7.5 Ranking centrales hidráulicas

RANKING VARIABLES						
POS	Paipa IV		Tasajero 1		Centro 1	
	Var	Coef.	Var	Coef.	Var	Coef.
1	EmbAgr	0.342	Pm.Mín	0.2945	Px.Mín	0.4048
2	Px.Mín	0.3384	EmbAgr	0.2741	EmbAgr	0.3311

3	Pm.Máx	0.2516	Pm.Med	0.1965	Cont.Energ	0.2008
4	Pm.Mín	0.2476	Px.Máx	0.1853	Pm.Med	0.1925
5	Px.Máx	0.2467	Cont.Energ	0.1814	Pm.Mín	0.1886
6	Cont[$/kW]	0.2399	Px.Mín	0.1664	Px.Med	0.1827
7	Pm.Med	0.2202	Pm.Máx	0.1652	Px.Máx	0.1804
8	Cont.Energ	0.2197	Cont[$/kW]	0.1587	Cont[$/kW]	0.176
9	Px.Med	0.1979	Px.Med	0.1578	Pm.Máx	0.1663
10	RecPos Pesos	0.1515	RecPos Pesos	0.1509	RecNeg Pesos	0.0522
11	RecPos Energ	0.135	RecPos Energ	0.1264	RecNeg Pesos	0.0516
12	RecNeg Energ	0.0761	RecNeg Pesos	0.0777	RecPos Energ	0.0398
13	RecNeg Pesos	0.0725	RecNeg Energ	0.0768	RecPos Pesos	0.0388
14						
POSICIÓN	Tebsa		Flores		Flores 3	
	Var	Coef.	Var	Coef.	Var	Coef.
1	EmbAgr	0.3251	EmbAgr	0.3325	EmbAgr	0.3946
2	Px.Med	0.2976	Px.Máx	0.2118	RecPos Pesos	0.2269
3	Px.Mín	0.2396	Pm.Med	0.2057	Px.Máx	0.2236
4	RecPos Energ	0.2071	Pm.Máx	0.1996	Cont.Energ	0.2235
5	Pm.Máx	0.2029	Cont.Energ	0.1989	Cont[$/kW]	0.2191
6	Px.Máx	0.2	Pm.Mín	0.195	Pm.Máx	0.2158
7	RecPos Pesos	0.1847	RecPos Energ	0.1918	Pm.Mín	0.2144

8	Pm.Mín	0.1795	Cont[$/kW]	0.1878	RecPos Energ	0.208
9	Cont[$/kW]	0.1792	RecPos Pesos	0.1869	Pm.Med	0.1958
10	Cont.Energ	0.167	RecNegEnerg	0.0444	RecNeg Pesos	0.0458
11	Pm.Med	0.1534	RecNeg Pesos	0.044	RecNegEnerg	0.0458
12	RecNegEnerg	0.105				
13	RecNeg Pesos	0.0997				
14						

Tabla 7.6 Ranking centrales térmicas

Finalmente, se seleccionaron las variables que mejor quedaron clasificadas de acuerdo al ranking anteriormente mostrado y son mostradas en las tablas 7.7 y 7.8.

VARIABLES SELECCIONADAS CENTRALES HIDRÁULICAS			
POS.	San Carlos	Guavio	Chivor
1	Pm. Demanda mínima	Embalse agregado	Contratos [$/kW]
2	Px. Demanda mediana	Pm. Demanda mínima	Embalse agregado
3	Embalse agregado	Embalse	Embalse
4	Px. Demanda mínima	Px. Demanda mediana	Pm. Demanda mínima
5	Pm. Demanda mediana	Px. Demanda mínima	Contratos energía
6	Px. Demanda máxima	Contratos [$/kW]	Px. Demanda mediana
7	Contratos energía	Pm. Demanda mediana	
POS.	Guatrón	Porce II	Alban
1	Pm. Demanda mínima	Pm. Demanda mínima	Embalse agregado

2	Px. Demanda mínima	Px. Demanda mediana	Embalse
3	Embalse agregado	Px. Demanda mínima	Pm. Demanda mediana
4	Px. Demanda mediana	Embalse agregado	Px. Demanda mediana
5	Contratos [$/kW]	Pm. Demanda mediana	Pm. Demanda mínima
6	Pm. Demanda mediana	Contratos [$/kW]	Px. Demanda mínima
7	Px. Demanda máxima		

Tabla 7.7 Variables seleccionadas centrales hidráulicas

VARIABLES SELECCIONADAS CENTRALES TÉRMICAS			
POS.	Paipa IV	Tasajero I	Centro I
1	Embalse agregado	Px. Demanda mediana	Contratos [$/kW]
2	Contratos [$/kW]	Embalse agregado	Embalse agregado
3	Pm. Demanda mínima	Px. Demanda máxima	Pm. Demanda máxima
4	Px. Demanda mediana	Px. Demanda mínima	Px. Demanda máxima
5	Px. Demanda mínima	Pm. Demanda máxima	Px. Demanda mediana
6	WTI	Contratos [$/kW]	Contratos energía
7		WTI	WTI
POS.	Tebsa	Flores	Flores 3
1	Embalse agregado	Embalse agregado	Embalse agregado
2	Contratos energía	Px. Demanda mínima	Px. Demanda mínima
3	Contratos [$/kW]	Px. Demanda máxima	Pm. Demanda máxima

4	Reconciliación positiva energía	Pm. Demanda mínima	Pm. Demanda mediana
5	Pm. Demanda mínima	Pm. Demanda máxima	Pm. Demanda mínima
6	Px. Demanda mínima	Px. Demanda mediana	Px. Demanda mediana
7			WTI

Tabla 7.8 Variables seleccionadas centrales térmicas

7.3.3. Análisis de componentes principales (ACP)

Partiendo de las tablas anteriores donde se muestran las variables seleccionadas, se quiso reducir la dimensionalidad mediante el uso de las "componentes principales", descritas en el capítulo 3, principalmente para facilitar la posterior aplicación del análisis cluster.

El ACP se realizó para las centrales de estudio sin obtener resultados satisfactorios, ya que la reducción de variables lograda NO fue eficiente; esto es, las dos o tres primeras componentes no contienen la cantidad de información esperada, y eliminar las últimas componentes implicaría obviar una cantidad de información no insignificante que no se ve redituada en dimensionalidad baja (2 o 3 dimensiones). En la tabla 7.9 y la figura 7.4 se muestran los resultados de San Carlos, como muestra de lo anteriormente expuesto.

CP	Valores propios	% info. CPs [%]
1	5.2908	32.398
2	3.7253	22.812
3	2.6806	16.415
4	2.1499	13.165
5	1.5435	9.452
6	0.7514	4.601
7	0.1889	1.157

Tabla 7.9 Resultados ACP para San Carlos

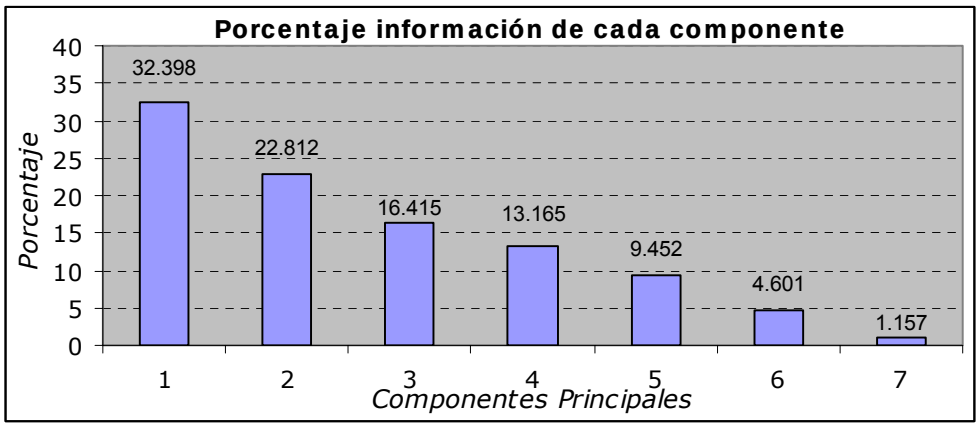

Figura 7.4 Porcentaje de información de cada CP para San Carlos

Se exponen los resultados del ACP en esta etapa del documento para hacer la claridad que en el segmento de resultados no se mencionará esta técnica por no ofrecer aportes importantes a los objetivos finales de esta investigación, esto en razón a que en ningún caso se consiguió la reducción de dimensiones esperada, por lo que en los posteriores análisis se usan la totalidad de las variables que se seleccionaron y que están listadas en las tablas 7.7 y 7.8.

7.4. Estandarización

Como se mencionó en el capítulo de análisis de conglomerados, en la aplicación de esta técnica, la escala en que se encuentren expresadas las variables puede tener un efecto nocivo sobre la exploración de datos, por ello se buscaron alternativas para llevar a cabo la estandarización de modo que sirviera como aporte a la presente investigación.

La estandarización es una transformación matemática mediante el cual se cambia la escala de lo parámetros de modo que todos los datos de ellos pertenezcan a un mismo dominio. El método más común es la normalización ó Z-score que se basa en el supuesto que la variable a estandarizar sigue una distribución normal. Sin embargo, existen métodos diferentes a este que pueden ser usados para lograr el mismo objetivo, y que se seleccionan de acuerdo a la aplicación o las necesidades del analista.

Para la presente investigación se quiso evaluar el funcionamiento de más de un método de estandarización, por lo que se consultaron diferentes trabajos en los que se estudiara este tema y se seleccionaron aquellos con los que se contará más información y que su complejidad no fuera tal, que desviará el objetivo final de este trabajo. Además, se tuvo en cuenta las fortalezas que se han encontrado, en estudios anteriores, para los diferentes métodos de estandarización.

En la siguiente tabla resumen, se muestran los métodos de estandarización que se consultaron; se trabajó solo con los tres primeros que allí se enuncian, en concordancia con los criterios expuestos en el párrafo anterior.

TÉCNICA DE NORMALIZACIÓN	ROBUSTEZ	EFICIENCIA
Z-Score	No	Alta (Ideal para datos con distribución normal)
Media y desviación absoluta de la mediana (Media y MAD)	No	Moderada
Mediana y desviación absoluta de la mediana	Si	Moderada
Mínimo – Máximo (min-max)	No	N/A
Escalamiento decimal	No	N/A
Estimadores Tanh	Si	Alta

Tabla 7.10 Resumen técnicas de normalización

A continuación se formulan los métodos probados en esta investigación. Entre paréntesis aparece el nombre interno que se les dio en este trabajo para facilitar su identificación.

Z- Score: (Normalización 1)

$$x'_k = \frac{x_k - \bar{x}}{\sigma} \qquad \textbf{(7.2)}$$

Media y desviación absoluta de la media: (Normalización 2)

$$x'_k = \frac{x_k - \bar{x}}{MAD_1} \qquad \textbf{(7.3)} \qquad MAD_1 = \frac{1}{n}\sum |x_i - \bar{x}| \qquad \textbf{(7.4)}$$

Mediana y desviación absoluta de la mediana: (Normalización 3)

$$x'_k = \frac{x_k - mediana}{MAD_2} \qquad \textbf{(7.5)} \qquad MAD_2 = mediana\left(|x_i - mediana|\right) \qquad \textbf{(7.6)}$$

Escalamiento decimal

$$x'_k = \frac{x_k - \min}{\max - \min} \qquad \textbf{(7.7)}$$

En todos los casos x_k representa el valor del dato y x'_k representa el dato estandarizado.

Con los tres primero métodos se hicieron diferentes pruebas de análisis cluster y de acuerdo a los resultados se concluyó al respecto y se seleccionó uno de

ellos, para finalizar el estudio; esta selección se coteja con las fortalezas que del método escogido se esperaban y se deja una referencia de consulta para futuros trabajos que necesiten llevar a cabo normalización como parte de la etapa de la adecuación de sus datos.

El método de escalamiento decimal fue usado en la aplicación de las MSV y de las mezclas finitas de acuerdo a las recomendaciones de los trabajos realizados previamente donde se aplicaron esta técnicas ([5] y [27]); la simpleza del cálculo con este método la hacen muy práctica para disminuir la carga computacional a la hora de realizar la estandarización de grandes cantidades de datos.

7.5. Formación de conglomerados

Al aplicar el análisis cluster se busca encontrar conglomerados en los que se aprecie un alto grado de semejanza entre sus elementos, y una gran disimilitud entre elementos de diferentes conglomerados; para realizar este análisis se requiere: ingresar datos estandarizados, además determinar la cantidad de conglomerados, la medida de similitud y método de agrupamiento como insumos..

Matlab® presenta, como herramientas útiles para esta etapa de la investigación, dos coeficientes cophenet y silohuette, los cuales fueron empleados como guía para determinar, con el primero, de una manera más precisa la mejor combinación de métodos de agrupamiento y distancias y con el segundo la cantidad de conglomerados. Es necesario aclarar que los altos resultados en este par de coeficientes no excluyen las opciones con bajo valor, por lo tanto no fueron un determinante exclusivo de las decisiones tomadas en las etapas posteriores.

El procedimiento que se siguió consiste en calcular el coeficiente de correlación Cophenetic (Cophenet) para cada central generadora, teniendo en cuenta los tres tipos de estandarización y todas las combinaciones posibles entre medidas de distancia y métodos de agrupamiento.

Dichos valores se resumen, en una tabla como la que se anexa a continuación (tabla 7.12); luego se seleccionaron las mejores combinaciones para proceder a realizar el agrupamiento correspondiente; es necesario aclarar que este coeficiente se hace mejor cuando su valor se acerca a uno (1). Cuando su valor es cercano a cero (0) quiere decir que la combinación de métrica y método de aglomeración no es buena.

	Single	Complete	Average	Ward	Centroid
Euclidean	0.5841	0.6056	0.7336	0.4992	0.7591
Mahalanobis	0.7813	0.6612	0.8366	0.4334	0.8289
Cityblock	0.4369	0.5528	0.6690	0.5305	0.6742

Chebychev	0.7344	0.6995	0.8349	0.5961	0.8350
Cosine	0.1533	0.5357	0.7122	0.6283	0.6978
Correlation	0.1533	0.5200	0.6668	0.6197	0.6473

Tabla 7.11 Cophenet para San Carlos

Por otra parte se desea determinar el óptimo número de conglomerados con la ayuda del valor Silhouette; este coeficiente establece la medida de similitud de un dato con los demás miembros del conglomerado al que pertenece y lo compara con la similitud a datos de otros grupos. Este comando en Matlab 6.5 se aplicó a varias combinaciones de métodos y distancias y para números de cluster desde 2 hasta 30, los resultados se graficaron para cada central generadora tal como se pueden apreciar en la figura mostrada a continuación, en ella se demuestra que los más altas similitudes se dan con 2 conglomerados, en cualquiera de las combinaciones de método y distancia; estos resultados fueron similares en todas las centrales generadoras analizadas.

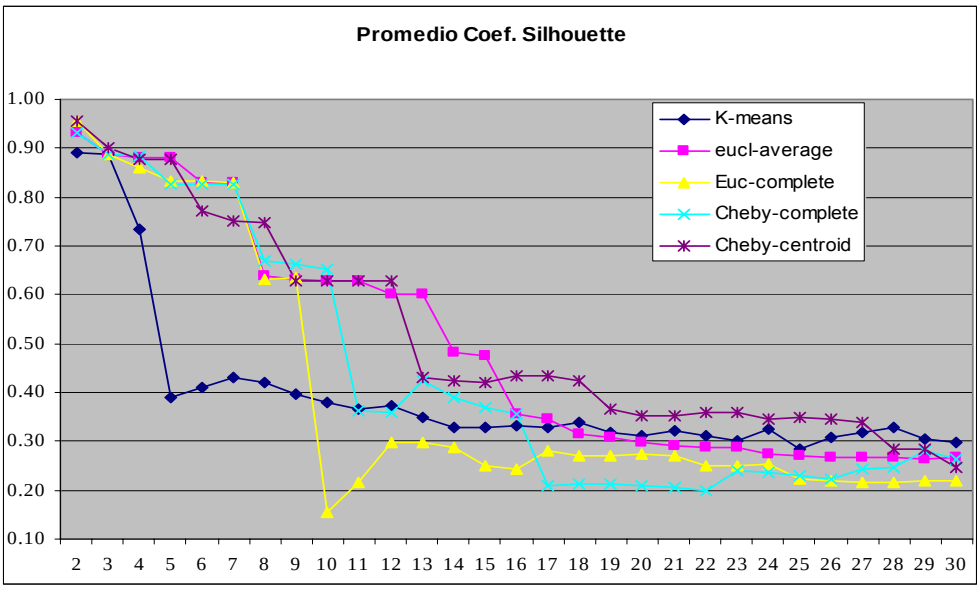

Figura 7.5 Coeficiente Silhouette

Con la ayuda de los coeficientes ya mencionados se escogieron las mejores combinaciones de método y distancia, y se dio inicio al análisis con 2 conglomerados, dicho valor se incrementa en la medida que los datos permitan una división sin presentar traslape en las franjas de precio de oferta, pues lo que finalmente se desea encontrar en la ejecución de esta etapa es conjuntos de datos agrupados bajo una misma "etiqueta" dada por el precio de oferta. Los resultados de la aglomeración se grafican de modo que se permita apreciar si existe o no precio característico, esta gráfica se realiza para cada prueba realizada y es similar a la mostrada en la siguiente figura.

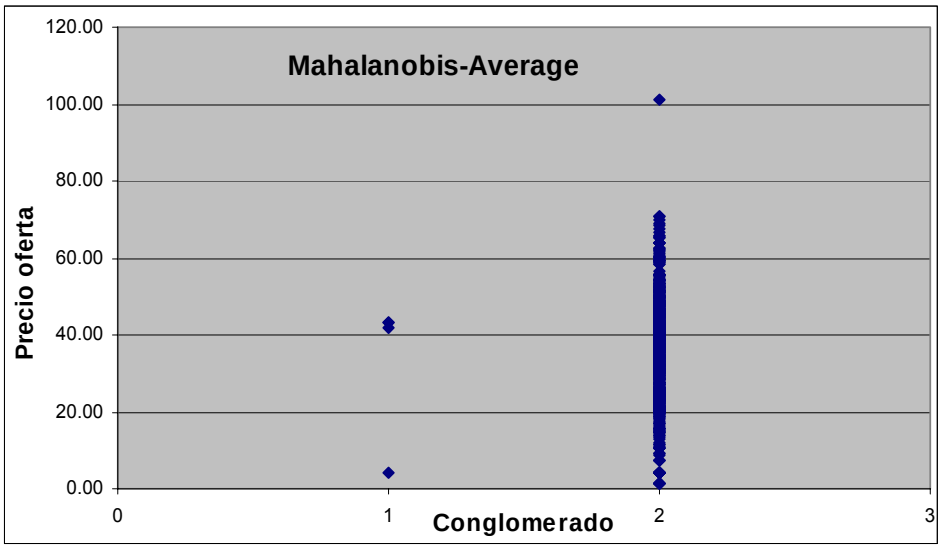

Figura 7.6 Precio Vs. Grupo. San Carlos Mahalanobis-Average

Luego de desarrollar este análisis para cada una de las centrales se consiguieron como máximo tres agrupamientos naturales para algunas de las centrales analizadas, en otros casos sólo se consiguieron dos conglomerados sin que hubiese traslape en las diferentes franjas de precios, a pesar de probar varias combinaciones de método y distancia de agrupamiento, por lo cual es posible afirmar que no existe precio característico para ninguna de las centrales.

Los resultados del análisis cluster no se mostrarán en este libro ya que no realizan un aporte considerable a los resultados finales de la investigación y ya que esta etapa se llevaba a cabo con el fin de establecer las etiquetas requeridas como insumo para los análisis consecuentes, se optó por usar 'Cluster unidimensional' para obtener las etiquetas de los precios, implementando la técnica en Matlab usando la distancia *Euclidiana* y utilizando el método de encadenamiento *medio* ó *average*. Esto es, utilizando únicamente los precios de oferta, corregidos para cada central, se formaron conglomerados de los precios que representan cada uno de los rangos de precios o etiquetas, necesarias en las MSV y en el clasificador Bayesiano de Naives.

7.6. Obtención de patrones mediante el uso de las MSV

Esta etapa corresponde al entrenamiento de las máquinas de soporte vectorial y posterior validación de estas, para lo cual se hará uso de la Toolbox en Matlab desarrollada por [Gómez, Morales, 2005]. Como se mencionó previamente, los datos de entrenamiento corresponden a los extraídos del periodo comprendido entre el 1 de enero de 2003 hasta el 31 de mayo del 2005; Los datos de validación corresponden a los datos comprendidos entres el 1 de junio de 2005 hasta el 31 de mayo del 2006.

Como se sabe, esta técnica requiere que los casos de entrenamiento estén etiquetados, etiquetas que se determinaron a partir de la formación de conglomerados univariada con el precio de oferta en bolsa, debido a la no obtención de precios característicos cuando se implementó Clustering con las variables de estudio, tal y como se expone en la sección 7.5. En la sección 8.1. se explica la forma como se obtuvieron las etiquetas y se muestran los grupos formados.

El entrenamiento de las MSV se llevó a cabo con todas las centrales seleccionadas para el estudio, 12 de acuerdo a lo enunciado en el apartado 7.1 Esto implica que para cada una de ellas se seleccionó previamente una función Kernel, los parámetros de dicho Kernel y el parámetro de penalización.

Esto implica un altísimo esfuerzo computacional, sustentados en pruebas de la herramienta realizadas con anterioridad, lo que se vería reflejado en un tiempo muy alto para la culminación de esta etapa. Debido a esto se optó por utilizar el *Kernel RBF*, basados en los buenos resultados obtenidos en [Gómez, Morales, 2005] y además posee una mayor simplicidad comparada con los Kernel Polinomial y Sigmoide.

Para obtener los parámetros para llevar a cabo el entrenamiento se utilizó la validación cruzada y búsqueda en malla, la cual está implementada en la Toolbox desarrollada por [Gómez, Morales, 2005]. La validación cruzada se implementó dividiendo el grupo de datos de entrenamiento en cinco (5) partes iguales. La búsqueda en malla se simplificó con el uso el Kernel RBF (2 dimensiones) y se implementó empezando con pasos de potencias de dos pero a medida que se obtenía mejores resultados en la validación cruzada la rejilla se hizo más pequeña. Ver anexo B.3. en [5] para ampliación de este tema.

Por otra parte, la validación consiste primero, en clasificar los casos del período de validación con la MSV entrenada obteniendo una etiqueta para cada uno de esos casos, etiquetas que corresponden a un rango de los precios de oferta; en segunda medida, se contrastan estos resultados con la etiqueta que deberían tener los precios de los casos de validación de acuerdo a las etiquetas y rangos de precios que se determinaron con los datos de entrenamiento. Veamos un ejemplo para entender esto de manera más clara.

Suponga que con los datos de entrenamiento se obtuvieron las siguientes etiquetas:

Etiquetas entrenamiento	Rango precios de oferta
1	[20 – 30)
2	[30-40)
3	[40-50]

Tabla 7.12 Etiquetas datos de entrenamiento

Ahora suponga que se va a validar un caso 'k'[1] cuyo precio de oferta es 35; según esto, el caso debería pertenecer a la etiqueta 2. Si la clasificación con la MSV entrenada arroja que este caso debe pertenecer a la etiqueta 2, esto se consideraría como un acierto; si la clasificación con la MSV arroja que el caso 'k' debe pertenecer, por ejemplo, a la etiqueta 3, entonces no hubo un acierto para este caso.

7.7. Clasificador Bayesiano de Naives a partir de Mezclas Finitas

En adición a lo que se planeó previo al desarrollo de la investigación, se agregó un componente adicional aprovechando que se tendría el conocimiento sobre el análisis cluster y apoyado en una investigación paralela a esta, enmarcada también en la monitorización del MEM de Colombia.

En esta etapa se pretende aplicar mezclas finitas y en particular el "Algoritmo EM [2]" sobre los conglomerados conformados en las fases predecesoras, con lo que será posible determinar la función de densidad de probabilidad de ellos y con esto tener el insumo para aplicar el "Clasificador Bayesiano de Naives[3]". Con la implementación de dicho clasificador será posible llegar a resultados similares en cuanto al tipo de respuesta, que los que se obtienen después de realizar la validación con las MSV.

El Algoritmo EM es un método usado para estimar los parámetros de las mezclas, pero para empezar necesita que se le especifiquen el número de componentes y valores iniciales de dichos parámetros, por lo que en la literatura se propone aplicar técnicas de clúster jerárquico antes de usar el algoritmo, para obtener esos valores necesarios para su inicialización; dichas técnicas se utilizan en esta investigación explicando así el motivo de adicionar este nuevo elemento al trabajo.

Con la implementación del algoritmo EM para obtener los parámetros de la mezcla se logra modelar la función de densidad de probabilidad de los conglomerados, con lo que se puede utilizar el clasificador Bayesiano de Naives. El clasificador utiliza las funciones de densidad de que arrojan las mezclas a partir de los datos de entrenamiento, para estimar la probabilidad de que un nuevo caso pertenezca a cada uno de los conglomerados, siendo la mayor de estas probabilidades el criterio para decidir a cual grupo debe asignarse dicho caso.

---

[1] Si el caso 'k' hace parte de los datos de validación quiere decir que este no estuvo incluido dentro del entrenamiento; sin embargo, cualquier caso, incluyendo los de entrenamiento, podrían ser clasificados con la máquina entrenada para verificar los aciertos de esta.

[2] Siglas del Inglés Expectation-Maximization Algorithm.

[3] Ver [27] y [28] para una explicación a profundidad del clasificador Bayesiano de Naives.

Con esto realizado, debe hacerse una comparación de la asignación obtenida del clasificador con las etiquetas del precio de oferta de los conglomerados conformados con los datos de entrenamiento, tal y como se llevará a cabo con los datos clasificados por las MSV entrenadas; para entender mejor este proceso por favor vea el ejemplo explicado al final de la sección 7.6.

8. PRUEBAS Y RESULTADOS

8.1. Análisis cluster

Recordando lo mencionado en la sección 7.5 podemos asegurar que debido a que el análisis cluster no permitió encontrar más de tres agrupamientos naturales, en ninguna de las centrales generadoras analizadas, los resultados para los datos correspondientes al período comprendido entre 1 de enero de 2003 hasta el 31de mayo de 2005 de las variables seleccionadas por cada central no serán mostrados de manera específica.

Sin embargo el análisis cluster, realizado de forma univariada sobre los precios de oferta, permitió obtener las etiquetas que fueron empleadas para las técnicas usadas posteriormente (ver 8.1.1); además, a partir de esto, se identificaron algunos casos de acuerdo a un criterio especificado en 8.1.2, lo cual permitió la entrega final del listado de datos que por su mínimo aporte a los resultados finales de la investigación se consideraron como atípicos.

8.1.1. Obtención de las etiquetas

Se realizaron varias pruebas para establecer cuantos grupos de precios, y por ende de etiquetas, se deben conformar. Se tomó la decisión de conformar 10 etiquetas de precios para cada central ya que con esta cantidad se obtuvieron rangos de precios lo suficientemente grandes como para acumular varios casos en cada grupo para poder aplicar las MSV, pero no tan grandes que los posteriores resultados no permitieran realizar inferencias precisas.

El uso de esta técnica permitió identificar atípicos en los precios de oferta de cada central. Esto se consiguió aplicando cluster con todos los precios de los casos del periodo de entrenamiento; cuando se formaba un conglomerado con muy pocos datos (menos del 0.56% de los casos), se entendía que estos casos representaban precio atípicos. Estos casos se excluían y nuevamente se realizaba la aplicación de cluster unidimensional con los restantes casos, para formar 10 nuevos grupos válidos, con los que se realizaría el entrenamiento de las máquinas de soporte vectorial.

La razón para extraer como atípicos los grupos menores a 0.56% es que por debajo de este porcentaje, la cantidad de datos del grupo es insuficiente para llevar a cabo la validación cruzada en las MSV, tal cual como se llevó a cabo. La extracción de los atípicos se hizo basado en este criterio ya que debe estar acorde con la forma como se formaron las etiquetas; si con la aplicación original del cluster multidimensional se hubiesen obtenidos precio típicos, a partir de este criterio se hubieran obtenido los atípicos.

A continuación se mostrará para cada central una tabla donde se resume los rangos de precios finales que se conformaron para llevar a cabo el entrenamiento de las MSV, después de descartar los atípicos.

Alban:

Mínimo	Máximo	% de casos en cada rango
1.2665	8.1149	39.35 %
13.324	33.614	10.13 %
34.244	67.692	38.07 %
84.876	97.912	1.40 %
105.45	139.02	2.56 %
144.43	155.71	1.40 %
472.82	486.94	1.28 %
492.79	504.64	0.70 %
571.39	582.04	0.81 %
588.85	611.01	4.13 %

Tabla 8.1 Etiquetas obtenidas para Alban

Chivor:

Mínimo	Máximo	% de casos en cada rango
1.2542	5.3335	43.41%
7.4506	11.798	1.75%
12.828	18.489	2.57%
19.357	27.259	9.68%
28.305	32.608	2.80%
34.093	39.025	3.73%
39.857	48.794	14.70%
49.158	55.961	10.27%
56.79	63.257	5.83%
64.124	73.808	5.25%

Tabla 8.2 Etiquetas obtenidas para Chivor

Guatrón:

Mínimo	Máximo	% de casos en cada rango
6.527	14.662	5.47%
15.372	20.131	11.29%
20.771	27.343	11.52%
28.269	36.924	23.03%
37.209	43.07	19.38%
43.415	49.407	13.91%
49.819	53.918	7.53%
54.414	58.853	4.22%
60.556	64.633	2.05%
66.594	73.111	1.60%

Tabla 8.3 Etiquetas obtenidas para Gutrón

Guavio:

Mínimo	Máximo	% de casos encada rango
1.2665	4.2817	19.66%
7.5533	12.685	2.74%
14.225	21.011	8.34%
21.41	28.945	15.09%
29.387	38.563	19.43%
38.952	46.327	22.06%
46.82	53.369	9.49%
54.937	59.839	1.83%
62.552	63.92	0.80%
67.692	68.253	0.57%

Tabla 8.4 Etiquetas obtenidas para Guavio

Porce 2:

Mínimo	Máximo	% de casos en cada rango
1.2583	4.2689	19.88%
9.0926	14.532	1.18%
16.641	19.903	1.06%
20.548	28.342	9.06%
29.212	37.937	18.71%
38.245	44.267	16.47%
44.531	52.834	19.88%
53.369	58.853	6.24%
59.788	67.692	5.76%
69.129	76.281	1.76%

Tabla 8.5 Etiquetas obtenidas para Porce 2

San Carlos:

Mínimo	Máximo	% de casos en cada rango
1.2695	4.2462	2.50%
7.44	11.761	1.59%
12.685	17.304	2.50%
18.083	26.251	19.91%
26.747	34.844	17.63%
35.135	40.596	17.18%
40.867	46.45	18.77%
46.794	56.478	15.70%
58.517	62.444	2.50%
63.776	70.75	1.71%

Tabla 8.6 Etiquetas obtenidas para San Carlos

Termocentro:

Mínimo	Máximo	% de casos en cada rango
7.2543	19.187	1.73%
21.771	30.991	3.93%
33.482	38.234	1.96%
122.45	128.78	9.35%
131.66	139.7	26.56%
166.46	166.46	2.66%
224.76	227.43	2.89%
448.72	451.04	11.09%
452.21	460.11	34.76%
463.24	465.83	5.08%

Tabla 8.7 Etiquetas obtenidas para Termocentro

Flores:

Mínimo	Máximo	% de casos en cada rango
31.274	46.374	1.49%
54.929	63.313	1.15%
71.17	83.413	10.00%
90.911	113.46	22.07%
119.37	128.59	2.99%
133.5	155.37	10.34%
169.94	170.71	8.51%
270.59	270.59	3.22%
439.02	440.29	20.57%
457.08	457.08	19.66%

Tabla 8.8 Etiquetas obtenidas para Flores

Flores 3:

Mínimo	Máximo	% de casos en cada rango
43.341	45.264	2.79%
87.901	91.222	1.28%
93.414	97.321	4.99%
98.767	102.8	26.36%
114.64	122.79	4.76%
435.04	436.59	3.37%
439.02	442.54	7.67%
447.15	455.37	23.34%
456.76	460.11	20.09%
463.24	465.83	5.34%

Tabla 8.9 Etiquetas obtenidas para Flores 3

Paipa 4:

Mínimo	Máximo	% de casos en cada rango
3.5639	12.252	12.08%
13.809	28.867	36.36%
29.299	46.422	32.68%
47.201	63.614	5.29%
133.97	140.97	0.69%
146.46	151.23	1.73%
161.49	169.72	3.22%
331.93	339.71	4.26%
442.15	448.72	0.58%
452.47	457.93	3.11%

Tabla 8.10 Etiquetas obtenidas para Paipa 4

Tebsa:

Mínimo	Máximo	% de casos en cada rango
31.275	38.855	11.25%
39.439	43.266	3.60%
45.332	48.775	1.04%
70.809	73.817	1.28%
75.699	81.668	7.19%
85.84	93.295	6.96%
94.542	101.41	25.06%
101.88	107.33	31.67%
111.24	114.62	2.32%
118.01	124.99	9.63%

Tabla 8.11 Etiquetas obtenidas para Tebsa

Tasajero:

Mínimo	Máximo	% de casos en cada rango
30.191	34.012	1.25%
35.605	50.808	54.27%
51.196	59.804	10.01%
60.216	74.585	20.25%
139.77	146.45	4.78%
147.43	157.57	3.41%
160.68	167.18	1.71%
172.81	175.91	0.57%
179.53	187.07	2.73%

189.51	195.11	1.02%

Tabla 8.12 Etiquetas obtenidas para Tasajero

8.1.2. Extracción de atípicos

La extracción de estos valores se hizo a partir de la formación de las etiquetas obtenidas de la aplicación de la técnica de cluster univariado con los precios de oferta en bolsa de cada una de los centrales para el periodo de datos de entrenamiento.

El criterio utilizado, como se menciona en la sección 8.1.1., fue sacar los grupos que estuvieran conformados por una cantidad de casos menor al 0.56% del total de los casos de entrenamiento. La razón para usar este criterio es que grupos con tan pocos elementos representan casos con cierta homogeneidad que no aportan de manera significativa a la búsqueda de patrones en los precios de oferta. La razón de escoger exactamente un porcentaje menor de 0.56%, es que por debajo de este valor, no es posible llevar a cabo la validación cruzada para las MSV utilizando los 5 subgrupos con los que se implementó.

A excepción de Alban, para la Centrales hidráulicas se encontró como característica común, que los casos extraídos como atípicos corresponden a precios altos, por encima de los precios que usualmente fijaron en ese período. Se planteó la hipótesis de que estos precios podían coincidir con niveles muy bajos de sus embalses propios, pero no se encontró evidencia que dichos precios coincidieran con esta condición.

Se observó que estos altos precios, para cada central, se fijaron en días consecutivos; esto puede indicar que la central estaba previendo una disminución en el agua que "llegaría a su embalse" ó que la central presentaba alguna restricción técnica en ese momento, como una falla en un equipo o un mantenimiento. Sin embargo la principal razón debe ser estratégica, y aunque la razón no es posible determinarse con exactitud si se aprecia que estas centrales se basan en lo que ocurre en el mercado y no solo en los eventos propios. Se descartó además la posibilidad, que los atípicos para las diferentes centrales se dieran en las mismas fechas.

El caso de los atípicos de Alban se diferencia del resto de las hidráulicas analizadas ya que esta central manejó una franja con unos precios "bajos" (90.5% de los casos) y una franja con precios "altos" (8.5% de los casos); los casos seleccionados (<1%) corresponden a precios fijados entre las dos franjas, razón por la cual no se buscó establecer una posible causalidad de estos.

El caso de las térmicas presentó la dificultad de que los atípicos seleccionados correspondieron en un 91% de las veces, a precios que no eran ni muy bajos ni muy altos, esto es, la mayoría de los casos extraídos como atípicos corresponden a precios en el intermedio del rango de los precios casados por cada central.

Ocurre que dichos rangos de precios en las centrales térmicas son más grandes comparados con los de las centrales hidráulicas, por lo que es fácil identificar varios grupos de precios con pocos casos, distribuidos a lo largo de el intervalo, algunos de los cuales clasificaron como atípicos. Por ello, explicar las causas de estos atípicos tendría poca validez

8.2. Resultados con MSV

Las Máquinas de Soporte Vectorial necesitan de tener definidos etiquetas para poder llevar a cabo el entrenamiento y poder realizar la posterior validación, las cuales se presentan en la sección 8.1.1

8.2.1.Entrenamiento de las MSV

Se implementó la validación cruzada y la búsqueda en malla para cada una de las centrales con lo que se obtuvo el valor del parámetro del Kernel y el parámetro de penalización; con estos valores se entrenaron las máquinas y posteriormente se clasificaron los datos destinados para la validación.

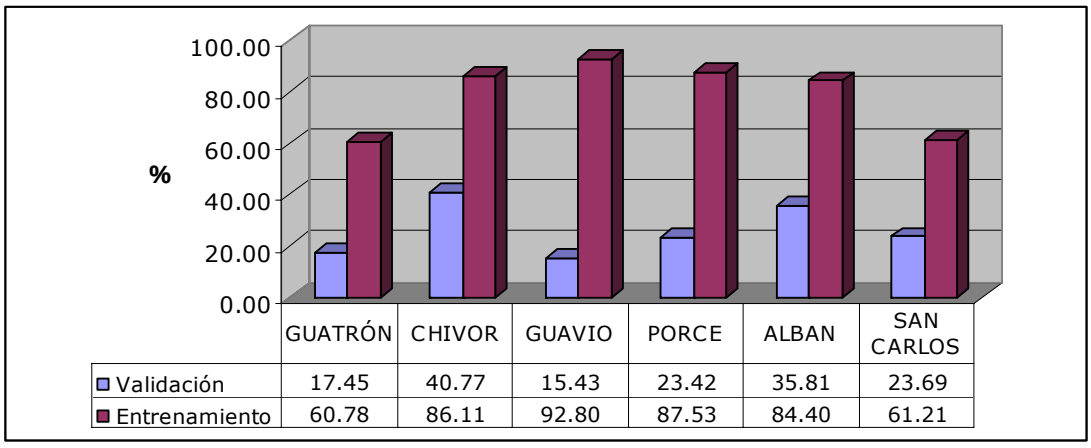

Figura 8.1 Comparación del porcentaje de aciertos con datos de entrenamiento y datos de validación de las MSV entrenadas de las centrales hidráulicas.

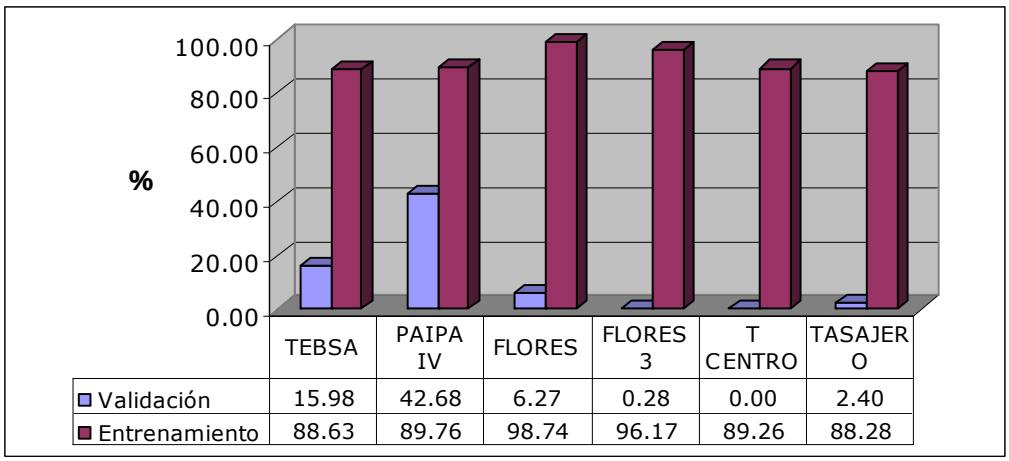

Figura 8.2 Comparación del porcentaje de aciertos con datos de entrenamiento y datos de validación de las MSV entrenadas de las centrales térmicas.

A partir de las gráficas 7.33 y 7.34 se puede ver que los valores de los porcentajes de acierto con los datos de validación son bajos, siendo Chivor la que presenta el valor más alto con un 40.77% de entre las centrales hidráulicas y Paipa IV con un 42.68% la que presentó mayor cantidad de aciertos con los datos de validación de entre las centrales térmicas.

Esto creó la necesidad de buscar nuevas alternativas para medir los resultados del entrenamiento y sobre todo de identificar patrones en los precios de oferta fijados por las empresas generadoras; debe recordarse que cada uno de los precios que diariamente fijan las centrales no solo está afectada por las variables del mercado si no también por la estrategia que plantee cada una.

Una alternativa fue el diseño de indicadores basados, primero en el hecho que los rangos son consecutivos y segundo, en los aciertos y cercanía ó lejanía de los casos de validación clasificados respecto a los precios de oferta correspondientes. Se nombraron estos indicadores de manera tal que facilitaran recordar su significado. Estos se explican a continuación:

Eficiencia 0: Corresponde al porcentaje de casos clasificados con la MSV que caen en el mismo rango al cual pertenecen de acuerdo a sus precios de oferta. En otras palabras es el porcentaje de aciertos de los datos clasificados por la máquina.

Eficiencia 1: Corresponde al porcentaje de casos clasificados con la MSV que están 1 ó menos rangos desviados de aquel al cual pertenecen de acuerdo a sus precios de oferta.

Por ejemplo, suponga que la máquina entrenada clasifica un caso *x* como perteneciente a un rango *k*; Si el precio fijado por la central corresponde al rango arrojado por la máquina ó al rango inmediatamente anterior o al rango inmediatamente siguiente, entonces esto es considerado como un 'acierto' que suma para el índice "Eficiencia 1". El porcentaje de los aciertos obtenidos de esta manera se le denominó Eficiencia 1

La razón de este índice es darle un poco de flexibilidad a la evaluación de la máquina clasificadora debido a que hubo rangos de precios que son muy pequeños, hasta del orden de los $2/kWh.

Desviados 3 ó más clases (Desv3clase): Corresponde al porcentaje de casos clasificados con la MSV que están desviados 3 ó más rangos de aquel al cual pertenecen de acuerdo a sus precios de oferta.

Por ejemplo, suponga que la máquina entrenada clasifica un caso *x* como perteneciente a un rango *k*; Si el precio fijado por la central está por lo menos tres (3) rangos distante del que arroja la máquina entrenada, entonces esto suma p. El porcentaje de casos clasificados que cumplan esta condición se denominó Desv3clase.

La idea de este índice es resaltar una gran diferencia entre los precios que está fijando la central y los que arroja la MSV de acuerdo a su entrenamiento.

Out Superior: Porcentaje de casos de validación cuyos precios de oferta son más altos que el precio de oferta máximo de los casos de entrenamiento.

Para efectos del cálculo de los índices de eficiencia y desv3clase, estos casos se asignan a un ficticio rango once (11).

Out Inferior: Porcentaje de casos de validación cuyos precios de oferta son más bajos que el precio de oferta mínimo de los casos de entrenamiento.

Para efectos del cálculo de los índices de eficiencia y desv3clase, estos casos se asignan a un ficticio rango cero (0).

Con el cálculo de estos indicadores se busca dar una herramienta a los entes interesados en la monitorización que les permita notar cambios en el comportamiento de las centrales generadoras, ya que un valor alto en el índice de desv3clase o valores bajos en los de eficiencia funcionan como señales de alarma de estos cambios.

La otra alternativa que se diseñó para buscar los patrones fue establecer los porcentajes de aciertos de la clasificación dentro de cada uno de los rangos que se calcularon para cada central; esto permite establecer el entrenamiento a cual franja de precios, para cada central, es a la que más logra describir.

Como elemento adicional se llevó a cabo el entrenamiento de la MSV y la validación, no solo con el juego de parámetros que resultó en el valor más alto en la validación cruzada si no con los 4 o 5 mejores en este aspecto. Con esto se buscó ratificar si para esta aplicación ceñirse estrictamente al resultado de la validación cruzada es la mejor opción.

8.2.2. Resultados centrales hidráulicas

Guatrón:

Kernel	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*
C	*4*	**C**	*1.4142*	**C**	*2*	**C**	*4*	**C**	*16*
Parám. Kernel	*0.25*	**Parám. Kernel**	*0.125*	**Parám. Kernel**	*0.25*	**Parám. Kernel**	*0.125*	**Parám. Kernel**	*0.25*

	% Val. Cruzada 48.246	% Val. Cruzada 50.62	% Val. Cruzada 51.66	% Val. Cruzada 50.55	% Val. Cruzada 48.467
	ENTRENAMIENTO	**ENTRENAMIENTO**	**ENTRENAMIENTO**	**ENTRENAMIENTO**	**ENTRENAMIENTO**
INDICES	**Valor [%]**	**Valor [%]**	**Valor [%]**	**Valor [%]**	**Valor [%]**
Eficiencia 0	64.99	73.32	60.78	81.64	74.00
Eficiencia 1	88.37	91.33	86.66	94.53	91.11
Desv3clase	4.79	3.53	5.59	2.17	3.31
% casos por rango	**% aciertos en cada rango**	**% aciertos en cada rango**	**% aciertos en cada rango**	**% aciertos en cada rango**	**% aciertos en cada rango**
5.47	91.67	97.92	89.58	97.92	100.00
11.29	69.70	74.75	60.61	80.81	77.78
11.52	69.31	79.21	69.31	87.13	78.22
23.03	74.75	80.20	73.76	85.64	80.69
19.38	78.24	84.12	73.53	87.65	81.18
13.91	45.90	61.48	44.26	72.95	60.66
7.53	36.36	45.46	25.76	68.18	48.49
4.22	27.03	48.65	13.51	67.57	59.46
2.05	44.44	44.44	27.78	55.56	50.00
1.60	35.71	42.86	35.71	71.43	50.00
	VALIDACIÓN	**VALIDACIÓN**	**VALIDACIÓN**	**VALIDACIÓN**	**VALIDACIÓN**
INDICES	**Valor [%]**	**Valor [%]**	**Valor [%]**	**Valor [%]**	**Valor [%]**
Eficiencia 0	25.76	20.78	25.76	18.56	24.65
Eficiencia 1	51.25	40.44	49.86	38.50	51.80
Desv3clase	33.80	44.04	34.07	44.04	33.24
Out Infer.	5.26	5.26	5.26	5.26	5.26
Out Super.	7.76	7.76	7.76	7.76	7.76
% casos por rango	**% aciertos en cada rango**	**% aciertos en cada rango**	**% aciertos en cada rango**	**% aciertos en cada rango**	**% aciertos en cada rango**
9.70	42.86	28.57	45.71	28.57	51.43
6.93	36.00	8.00	28.00	8.00	36.00
8.03	0.00	0.00	0.00	0.00	0.00
10.25	70.27	81.08	67.57	64.87	62.16
5.82	19.05	23.81	33.33	9.52	9.52
11.36	60.98	43.90	65.85	43.90	48.78
12.47	24.44	17.78	22.22	17.78	24.44
4.43	6.25	6.25	0.00	12.50	31.25
10.25	2.70	2.70	2.70	2.70	2.70
7.76	3.57	0.00	0.00	0.00	0.00

Tabla 8.13 Resultados del entrenamiento y validación con MSV para Guatrón

Los parámetros con los que se obtuvo el valor mayor en la validación cruzada son los que ofrecen los valores de eficiencia más bajos para los datos de entrenamiento, al contrario de los que se esperaba antes de empezar este análisis. Sin embargo, el valor más alto de *eficiencia 0* para los datos de entrenamiento es apenas del 81,64% y de *eficiencia 1*, de 94,53% para el mismo conjunto de parámetros. Esto indica que ni siquiera para los datos de entrenamiento los aciertos son muy altos, con los parámetros obtenidos con la validación cruzada y la búsqueda en malla.

Observando los resultados con los datos de validación y en particular los correspondientes a los parámetros que ofrecieron las mejores eficiencias con los datos de entrenamiento, se observa que los valores de las eficiencias 0 y 1, son 11,91% y 26,04%, respectivamente; Ahora, estos mismos indicadores

con los parámetros que ofrecieron el mejor resultado en la validación cruzada, son los que ofrecen los valores más altos de todos las combinaciones probadas: 17,45% y 40,44%. Estos valores indican que ni a una quinta parte de los casos clasificados se les acertó el rango del precio de oferta y ni dando un (1) rango de precios de holgura se consiguió acertar al menos en la mitad de los casos. En concordancia a lo expuesto el valor de desviación es superior al 40%, es decir, a cada 4 de 10 casos clasificados en la validación, el rango obtenido a partir de la MSV estuvo alejado al menos tres rangos de precios del fijado por la central, una diferencia de aproximadamente $14 en el mejor de los casos, de acuerdo a la tabla de rangos de precios de Guatrón.

Lo que muestra esto, es que Guatrón ha cambiado de manera significativa la forma como hace la casación de los precios de oferta en la bolsa de energía. Véase también como un 13,02% de los precios de los casos de validación están por encima o por debajo de los límites de los precios fijados en los casos de entrenamiento; asi mismo, la concentración de los precios en los diferentes rangos cambió mucho entre los casos de entrenamiento y los de validación lo que lleva a que en cuatro de estos grupos no hubiera ni un solo acierto de parte de la clasificación echa por la MSV.

CHIVOR:

	Kernel	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*
	C	*8*	**C**	*22.63*	**C**	*45.25*	**C**	*181.02*	**C**	*256*
	Parám. Kernel	*0.125*	**Parám. Kernel**	*0.25*	**Parám. Kernel**	*0.25*	**Parám. Kernel**	*0.50*	**Parám. Kernel**	*0.25*
	% Val. Cruzada	*56.47*	**% Val. Cruzada**	*57.30*	**% Val. Cruzada**	*56.83*	**% Val. Cruzada**	*57.05*	**% Val. Cruzada**	*57.52*
	ENTRENAMIENTO		**ENTRENAMIENTO**		**ENTRENAMIENTO**		**ENTRENAMIENTO**		**ENTRENAMIENTO**	
INDICES	**Valor [%]**		**Valor [%]**		**Valor [%]**		**Valor [%]**		**Valor [%]**	
Eficiencia 0	80.98		73.28		77.13		69.78		86.11	
Eficiencia 1	85.30		78.41		81.91		74.80		89.15	
Desv3clase	14.24		20.65		17.27		24.27		10.50	
% casos por rango	**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**	
43.41	88.98		88.98		89.52		88.44		90.59	
1.75	80.00		80.00		80.00		80.00		100.00	
2.57	68.18		63.64		68.18		63.64		72.73	
9.68	89.16		78.31		84.34		72.29		92.77	
2.80	79.17		50.00		70.83		54.17		83.33	
3.73	65.63		43.75		59.38		40.63		78.13	
14.70	72.22		58.73		61.11		49.21		78.57	
10.27	70.46		55.68		62.50		48.86		80.68	
5.83	70.00		52.00		62.00		38.00		80.00	
5.25	75.56		68.89		71.11		73.33		84.44	
	VALIDACIÓN		**VALIDACIÓN**		**VALIDACIÓN**		**VALIDACIÓN**		**VALIDACIÓN**	
INDICES	**Valor [%]**		**Valor [%]**		**Valor [%]**		**Valor [%]**		**Valor [%]**	
Eficiencia 0	44.63		36.64		36.36		29.20		36.36	
Eficiencia 1	49.04		41.87		41.05		36.09		40.77	
Desv3clase	45.73		49.04		50.41		56.47		50.41	
Out Infer.	0.00		0.00		0.00		0.00		0.00	
Out Super.	7.99		7.99		7.99		7.99		7.99	

% casos por rango	% aciertos en cada rango	% aciertos en cada rango	% aciertos en cada rango	% aciertos en cada rango	% aciertos en cada rango
47.66	91.91	71.10	72.25	57.23	72.83
1.93	0.00	0.00	0.00	0.00	0.00
3.31	0.00	0.00	0.00	0.00	0.00
6.61	0.00	0.00	0.00	0.00	0.00
3.31	0.00	0.00	0.00	0.00	0.00
4.68	0.00	5.88	5.88	17.65	5.88
4.96	5.56	11.11	11.11	0.00	11.11
3.31	16.67	33.33	16.67	16.67	16.67
7.71	0.00	7.14	7.14	3.57	3.57
8.54	0.00	3.23	0.00	3.23	0.00

Tabla 8.14 Resultados del entrenamiento y validación con MSV para Chivor

Para Chivor, los parámetros que ofrecieron el mejor resultado de la validación cruzada (57,521%) sí ofrecieron los mejores resultados en cuanto a eficiencias para los datos de entrenamiento: 86,11% y 89,15%. Con los casos de validación, estos mismos parámetros, no fueron los que brindaron los valores más altos de eficiencia.

Se resalta que el primer rango de precios de Chivor, comprendido entre $1,2542 y $5.3335, es en el que se concentran la mayor cantidad de datos, teniéndose un porcentaje de aciertos superior al 88% en los datos de entrenamiento y para los datos de validación, por encima del 71% para 4 de los 5 conjuntos de datos probados.

Analizando los resultados con lo que se obtuvo el mayor valor de eficiencia en la validación (44,63%), se aprecia que para el mencionado primer rango de precios, el porcentaje de aciertos es superior al 91%.

A pesar de la concentración de precios en el primer rango, se ve que la concentración de precios en los rangos más altos aumentó un poco, lo concuerda con el 8% de los casos de validación cuyos precios fueron mayores al precio más alto en los datos de entrenamiento.

GUAVIO:

	Kernel	*RBF*	Kernel	*RBF*	Kernel	*RBF*	Kernel	*RBF*	Kernel	*RBF*
	C	*8*	C	*22.63*	C	*16*	C	*1.4142*	C	*64*
	Parám. Kernel	*0.125*	Parám. Kernel	*0.177*	Parám. Kernel	*0.25*	Parám. Kernel	*0.125*	Parám. Kernel	*0.0883*
	% Val. Cruzada	*60.02*	% Val. Cruzada	*60.08*	% Val. Cruzada	*58.95*	% Val. Cruzada	*59.35*	% Val. Cruzada	*58.24*
	ENTRENAMIENTO		ENTRENAMIENTO		ENTRENAMIENTO		ENTRENAMIENTO		ENTRENAMIENTO	
INDICES	Valor [%]		Valor [%]		Valor [%]		Valor [%]		Valor [%]	
Eficiencia 0	93.60		92.80		81.60		80.80		99.54	
Eficiencia 1	97.94		97.37		91.66		91.54		99.89	
Desv3clase	1.49		2.17		7.20		6.74		0.11	
% casos por rango	% aciertos en cada rango		% aciertos en cada rango		% aciertos en cada rango		% aciertos en cada rango		% aciertos en cada rango	
19.66	94.77		93.61		77.33		80.81		99.42	
2.74	87.50		87.50		66.67		54.17		100.00	
8.34	80.82		79.45		72.60		71.23		97.26	
15.09	96.21		95.46		87.88		87.12		99.24	

19.43	94.12	92.35	80.00	81.77	100.00
22.06	95.34	94.82	88.08	88.60	100.00
9.49	95.18	95.18	78.31	77.11	100.00
1.83	93.75	93.75	87.50	50.00	100.00
0.80	100.00	100.00	100.00	57.14	100.00
0.57	80.00	100.00	80.00	40.00	100.00
	VALIDACIÓN	**VALIDACIÓN**	**VALIDACIÓN**	**VALIDACIÓN**	**VALIDACIÓN**
INDICES	**Valor [%]**	**Valor [%]**	**Valor [%]**	**Valor [%]**	**Valor [%]**
Eficiencia 0	18.73	17.91	17.63	20.94	24.24
Eficiencia 1	30.30	33.61	37.74	32.23	34.16
Desv3clase	59.78	54.55	47.38	56.20	59.23
Out Infer.	0.00	0.00	0.00	0.00	0.00
Out Super.	2.75	2.75	2.75	2.75	2.75
% casos por rango	**% aciertos en cada rango**	**% aciertos en cada rango**	**% aciertos en cada rango**	**% aciertos en cada rango**	**% aciertos en cada rango**
31.41	48.25	41.23	36.84	53.51	66.67
2.75	0.00	0.00	0.00	0.00	0.00
1.65	0.00	16.67	16.67	0.00	0.00
4.96	5.56	5.56	11.11	5.56	5.56
9.37	11.77	23.53	26.47	14.71	11.77
8.26	20.00	16.67	13.33	30.00	20.00
11.02	5.00	7.50	12.50	0.00	2.50
12.67	0.00	0.00	0.00	0.00	0.00
11.02	0.00	0.00	2.50	0.00	0.00
4.13	0.00	0.00	0.00	0.00	0.00

Tabla 8.15 Resultados del entrenamiento y validación con MSV para Guavio

Obsérvese que hay tres conjuntos de parámetros con los que se obtuvo eficiencias mayores a 93% en los datos de entrenamiento, 2 de ellos corresponden a los que ofrecieron el mejor resultado en la validación cruzada, pero el otro, con el que se mostraron las mejores eficiencias, corresponde al que entregó el menor valor de la validación cruzada de entre los resultados aquí mostrados, recordando que todos estos resultados son los mejores de acuerdo a dicho criterio.

Los resultados de la clasificación con los datos da validación muestran eficiencias que no sobrepasan el 25% en ninguno de los casos presentados y en cambio si presenta valores de "desv3clase" del orden de 59%. Para justificar esto, se observa como hubo un cambio sustancial en la concentración de la fijación de los precios, ya que en rangos en los que en la etapa de clasificación el porcentaje de casos era de alrededor del 20%, subieran ó bajaran más de 10 puntos porcentuales en la etapa de validación; esto hace que los aciertos en cada uno de los rangos también se bajen de manera notoria lo que se refleja en los valores bajos de eficiencia.

El primer rango de precios para la etapa de validación concentra el 31,41% de los casos y es donde se logran los mejores porcentajes de aciertos. En los restantes rangos los aciertos son muy pocos y por ello los índices de eficiencia son tan bajos.

PORCE:

% Val. Cruzada	*45.16*	**% Val. Cruzada**	*44.21*	**% Val. Cruzada**	*43.01*	**% Val. Cruzada**	*43.65*	**% Val. Cruzada**	*43.40*

	ENTRENAMIENTO	ENTRENAMIENTO	ENTRENAMIENTO	ENTRENAMIENTO	ENTRENAMIENTO
INDICES	**Valor [%]**	**Valor [%]**	**Valor [%]**	**Valor [%]**	**Valor [%]**
Eficiencia 0	87.53	91.06	79.88	88.24	96.94
Eficiencia 1	92.35	94.71	87.41	92.71	98.47
Desv3clase	5.76	3.88	10.12	5.53	1.41
% casos por rango	**% aciertos en cada rango**	**% aciertos en cada rango**	**% aciertos en cada rango**	**% aciertos en cada rango**	**% aciertos en cada rango**
19.88	85.21	92.90	68.64	87.57	97.04
1.18	80.00	80.00	70.00	80.00	100.00
1.06	66.67	66.67	77.78	66.67	88.89
9.06	85.71	89.61	84.42	88.31	94.81
18.71	88.05	88.68	83.02	87.42	96.86
16.47	89.29	91.43	80.71	88.57	96.43
19.88	88.76	93.49	82.25	89.94	97.63
6.24	90.57	92.45	86.79	90.57	98.11
5.76	89.80	91.84	83.67	89.80	100.00
1.76	86.67	86.67	86.67	86.67	93.33
	VALIDACIÓN	**VALIDACIÓN**	**VALIDACIÓN**	**VALIDACIÓN**	**VALIDACIÓN**
INDICES	**Valor [%]**	**Valor [%]**	**Valor [%]**	**Valor [%]**	**Valor [%]**
Eficiencia 0	19.84	29.48	18.73	23.69	26.72
Eficiencia 1	31.96	35.81	32.78	34.16	36.09
Desv3clase	57.85	60.61	57.30	58.68	56.47
Out Infer.	0.00	0.00	0.00	0.00	0.00
Out Super.	19.28	19.28	19.28	19.28	19.28
% casos por rango	**% aciertos en cada rango**	**% aciertos en cada rango**	**% aciertos en cada rango**	**% aciertos en cada rango**	**% aciertos en cada rango**
36.92	44.03	73.88	39.55	55.22	62.69
1.93	0.00	0.00	0.00	0.00	0.00
0.55	0.00	0.00	0.00	0.00	0.00
1.65	33.33	16.67	33.33	16.67	16.67
5.79	9.52	4.76	4.76	4.76	9.52
4.41	12.50	6.25	25.00	6.25	6.25
9.37	5.88	8.82	11.77	11.77	14.71
8.26	13.33	6.67	10.00	13.33	13.33
6.06	4.55	0.00	4.55	4.55	0.00
5.79	0.00	0.00	0.00	0.00	0.00

Tabla 8.16 Resultados del entrenamiento y validación con MSV para Porce

En Porce, los parámetros con la mejor validación cruzada (45,16%) presentan eficiencias por encima al 87%, pero este valor no es de los más altos comprando con las otras pruebas mostradas. Pasando a la etapa de validación se encuentra que el valor más alto de eficiencia 1 es apenas de 31,96%, y el porcentaje de casos desviados tres o más rangos es de 57,85% indicando un cambio marcado en la casación de los precios en este período de tiempo.

Véase que con los datos de validación, el primer de rango es donde más se concentraron los precios y para este, con tres conjuntos de parámetros se obtuvo un porcentaje de aciertos superior al 55%; para esos mismos parámetros los aciertos en los otros rangos son muy bajos. Por otro lado, los dos restantes conjuntos de parámetros no fueron tan buenos clasificadores del primer rango de precios, pero con un porcentaje de aciertos superior para el resto de los rangos comparado con los otros parámetros.

No se puede dejar pasar por alto que casi un 20% de los precios fijados por la central en la etapa de validación fueron más altos que el precio más alto fijado en la etapa de entrenamiento, confirmando un cambio en el comportamiento de la Central a la hora de fijar los precios.

ALBAN:

	Kernel	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*
	C	*4*	**C**	*2*	**C**	*2*	**C**	*4*
	Parám. Kernel	*0.25*	**Parám. Kernel**	*0.088*	**Parám. Kernel**	*0.125*	**Parám. Kernel**	*0.125*
	% Val. Cruzada	*66.35*	**% Val. Cruzada**	*67.38*	**% Val. Cruzada**	*66.31*	**% Val. Cruzada**	*67.26*
	ENTRENAMIENTO		**ENTRENAMIENTO**		**ENTRENAMIENTO**		**ENTRENAMIENTO**	
INDICES	**Valor [%]**		**Valor [%]**		**Valor [%]**		**Valor [%]**	
Eficiencia 0	71.60		84.40		79.51		82.42	
Eficiencia 1	80.33		88.71		85.68		87.54	
Desv3clase	7.10		5.01		5.24		4.89	
% casos por rango	**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**	
39.35	83.73		90.53		87.28		89.05	
10.13	57.47		78.16		73.56		73.56	
38.07	80.12		91.13		87.16		89.30	
1.40	0.00		58.33		0.00		41.67	
2.56	0.00		36.36		18.18		36.36	
1.40	16.67		91.67		75.00		91.67	
1.28	0.00		9.09		9.09		9.09	
0.70	0.00		50.00		33.33		50.00	
0.81	28.57		28.57		28.57		28.57	
4.31	43.24		56.76		56.76		56.76	
	VALIDACIÓN		**VALIDACIÓN**		**VALIDACIÓN**		**VALIDACIÓN**	
INDICES	**Valor [%]**		**Valor [%]**		**Valor [%]**		**Valor [%]**	
Eficiencia 0	29.75		31.41		31.96		31.96	
Eficiencia 1	47.66		47.93		48.21		50.14	
Desv3clase	28.10		28.38		28.65		27.82	
Out Infer.	0.00		0.00		0.00		0.00	
Out Super.	1.38		1.38		1.38		1.38	
% casos por rango	**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**	
22.31	58.03		79.01		71.61		65.43	
14.33	15.39		5.77		11.54		9.62	
36.09	38.93		33.59		38.93		41.99	
2.20	0.00		0.00		0.00		0.00	
0.00	NaN		NaN		NaN		NaN	
0.00	NaN		NaN		NaN		NaN	
6.34	0.00		8.70		0.00		8.70	
14.60	0.00		0.00		0.00		0.00	
1.93	0.00		0.00		0.00		0.00	
0.83	66.67		33.33		33.33		33.33	

Tabla 8.17 Resultados del entrenamiento y validación con MSV para Alban

Para Alban se aprecia que los mejores resultados de la validación cruzada, los mostrados aquí, son valores muy similares. A pesar de esto, hay unas diferencias un poco más marcadas en las eficiencias de los datos de

entrenamiento, sin embargo, las eficiencias en la etapa de validación son muy similares, del alrededor de 31% para la eficiencia 0 y de 48% para la eficiencia 1.

Esta central tuvo pocos casos en la etapa de validación cuyos precios se salieron del rango que se estableció en la etapa de entrenamiento, sin embargo si hubo un descenso significativo en la concentración de casos en el rango de precios más bajo, pasó de 39,35% a 22,31%, y un incremento de 0,7% a 14,60% de casos en el octavo rango de precios más lato. Esto ocasiona que la máquina entrenada baje su eficiencia.

El valor del índice "desv3clase" es de 28%, valor que comparado con los resultados de las restantes centrales parece no ser muy alto pero si es indicador de una significativa variación en la fijación de los precios.

SANCARLOS:

	Kernel	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*
	C	*16*	**C**	*8*	**C**	*2*	**C**	*2*	**C**	*4*
	Parám. Kernel	*1*	**Parám. Kernel**	*0.500*	**Parám. Kernel**	*0.1768*	**Parám. Kernel**	*0.3535*	**Parám. Kernel**	*0.25*
	% Val. Cruzada	*55.65*	**% Val. Cruzada**	*55.42*	**% Val. Cruzada**	*54.62*	**% Val. Cruzada**	*55.42*	**% Val. Cruzada**	*55.08*
	ENTRENAMIENTO		**ENTRENAMIENTO**		**ENTRENAMIENTO**		**ENTRENAMIENTO**		**ENTRENAMIENTO**	
INDICES	**Valor [%]**		**Valor [%]**		**Valor [%]**		**Valor [%]**		**Valor [%]**	
Eficiencia 0	61.21		64.85		77.93		64.39		73.83	
Eficiencia 1	87.37		89.08		93.06		88.74		92.15	
Desv3clase	5.01		3.87		3.19		4.32		3.41	
% casos por rango	**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**	
2.50	4.55		9.09		27.27		9.09		27.27	
1.59	50.00		71.43		71.43		64.29		71.43	
2.50	9.09		18.18		31.82		9.09		27.27	
19.91	85.71		86.86		96.00		89.14		93.14	
17.63	59.36		65.81		76.77		62.58		74.19	
17.18	56.29		62.25		78.15		62.91		69.54	
18.77	69.70		71.52		80.00		70.30		76.36	
15.70	58.70		59.42		79.71		60.15		75.36	
2.50	0.00		0.00		31.82		0.00		27.27	
1.71	33.33		40.00		53.33		40.00		53.33	
	VALIDACIÓN		**VALIDACIÓN**		**VALIDACIÓN**		**VALIDACIÓN**		**VALIDACIÓN**	
INDICES	**Valor [%]**		**Valor [%]**		**Valor [%]**		**Valor [%]**		**Valor [%]**	
Eficiencia 0	28.93		27.82		25.62		27.55		27.82	
Eficiencia 1	63.36		60.61		53.17		60.06		57.03	
Desv3clase	21.76		23.42		29.20		26.72		28.10	
Out Infer.	0.00		0.00		0.00		0.00		0.00	
Out Super.	8.54		8.54		8.54		8.54		8.54	
% casos por rango	**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**	
9.92	0.00		2.78		0.00		0.00		0.00	
1.65	0.00		0.00		0.00		0.00		16.67	
5.51	5.00		15.00		15.00		5.00		25.00	
7.16	69.23		30.77		19.23		34.62		26.92	
11.30	36.59		46.34		53.66		51.22		51.22	

6.06	18.18	13.64	13.64	27.27	13.64
9.09	15.15	6.06	9.09	9.09	9.09
23.42	62.35	65.88	65.88	65.88	64.71
10.47	0.00	2.63	2.63	0.00	7.89
6.89	36.00	32.00	0.00	16.00	12.00

Tabla 8.18 Resultados del entrenamiento y validación con MSV para San Carlos.

En San Carlos se aprecia gran paridad en los valores de la validación cruzada, sin embargo, el que tiene el menor de estos valores fue con el que la MSV entregó los valores más altos de eficiencia para la etapa de clasificación.

Pasando a analizar los resultados de la validación, la situación se invierte y los mejores valores de eficiencias corresponden al caso donde los parámetros ofrecieron la mejor validación cruzada: 60,33% de eficiencia 1. Se resalta que aproximadamente una cuarta parte de los casos de validación clasificados, estuvieron alejados 3 ó más rangos de aquel al que pertenece de acuerdo a sus respectivos precios de oferta.

Los porcentajes de casos por rangos variaron un poco la concentración en el período de tiempo correspondiente a la validación; durante la clasificación, los precios caían en su mayoría en los rangos intermedios, del 4to al 8vo, pero en la etapa de validación, el porcentaje de casos concentrados en los rangos de los extremos se incrementó.

Así mismo, el 8,54% de los casos de validación cuyos precios estuvieron por encima del precio más alto de los casos usados para el entrenamiento, influencian en la disminución de las eficiencias. Tampoco se aprecia, ni para el caso de los datos de entrenamiento ni de validación, un rango de precios en el que la MSV ofrezca porcentajes de acierto significativamente altos, esto es, valores cercanos al 100%.

8.2.3. Resultados centrales térmicas

TEBSA:

	Kernel	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*
	C	*4*	**C**	*4*	**C**	*1*	**C**	*2*	**C**	*8*
	Parám. Kernel	*0.25*	**Parám. Kernel**	*0.125*	**Parám. Kernel**	*0.125*	**Parám. Kernel**	*0.1768*	**Parám. Kernel**	*0.125*
	% Val. Cruzada	*84.82*	**% Val. Cruzada**	*82.03*	**% Val. Cruzada**	*81.68*	**% Val. Cruzada**	*81.68*	**% Val. Cruzada**	*81.80*
	ENTRENAMIENTO		**ENTRENAMIENTO**		**ENTRENAMIENTO**		**ENTRENAMIENTO**		**ENTRENAMIENTO**	
INDICES	**Valor [%]**		**Valor [%]**		**Valor [%]**		**Valor [%]**		**Valor [%]**	
Eficiencia 0	88.63		94.90		89.68		89.33		96.52	
Eficiencia 1	93.39		97.45		93.97		93.62		98.26	
Desv3clase	5.45		2.44		4.99		5.10		1.39	
% casos por rango	**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**	
11.25	84.54		91.75		84.54		84.54		94.85	
3.60	54.84		77.42		54.84		54.84		87.10	
1.04	77.78		77.78		66.67		77.78		88.89	
1.28	81.82		90.91		81.82		90.91		90.91	

7.19	95.16	98.39	93.55	95.16	98.39
6.96	73.33	90.00	71.67	71.67	93.33
25.06	97.22	98.61	97.22	96.76	99.07
31.67	91.21	95.24	94.14	93.04	96.34
2.32	60.00	90.00	65.00	65.00	95.00
9.63	90.36	98.80	93.98	91.57	98.80
	VALIDACIÓN	**VALIDACIÓN**	**VALIDACIÓN**	**VALIDACIÓN**	**VALIDACIÓN**
INDICES	**Valor [%]**	**Valor [%]**	**Valor [%]**	**Valor [%]**	**Valor [%]**
Eficiencia 0	11.02	6.06	6.06	8.26	6.06
Eficiencia 1	42.98	59.23	57.58	57.58	58.95
Desv3clase	25.07	29.48	29.75	27.00	30.03
Out Infer.	0.00	0.00	0.00	0.00	0.00
Out Super.	0.00	0.00	0.00	0.00	0.00
% casos por rango	**% aciertos en cada rango**	**% aciertos en cada rango**	**% aciertos en cada rango**	**% aciertos en cada rango**	**% aciertos en cada rango**
1.6529	0.00	0.00	0.00	0.00	0.00
0	NaN	NaN	NaN	NaN	NaN
14.325	1.92	0.00	0.00	0.00	0.00
13.499	0.00	0.00	0.00	0.00	0.00
9.6419	22.86	11.43	11.43	37.14	11.43
41.047	13.42	0.00	0.00	4.03	0.00
7.7135	21.43	64.29	64.29	35.71	64.29
12.121	11.36	0.00	0.00	2.27	0.00
0	NaN	NaN	NaN	NaN	NaN
0	NaN	NaN	NaN	NaN	NaN

Tabla 8.19 Resultados del entrenamiento y validación con MSV para Tebsa.

Observando los resultados del conjunto de parámetros con los que dio el valor más alto la validación cruzada, se ve una eficiencia 0 de 88,63% y una eficiencia 1 de 93,39%, valores claramente más bajos que el 96,52% y 98,26% para los mismos indicadores con un conjunto de parámetros que dieron una validación cruzada menor.

La eficiencia 0 del conjunto de parámetros con que se obtuvo la mayor validación cruzada, es apenas de 11%, y la eficiencia 1 es de 42.98%. Este último, es el resultado más bajo de todos los resultados mostrados, ya que para los restantes conjuntos de parámetros este índice tiene un valor alrededor de un 58%. Además, el parámetro de 'desv3clase' es superior en todos los casos al 25%.

Es fácilmente apreciable que para Tebsa el porcentaje de casos por rango cambió en la etapa de validación comparado con la etapa de clasificación, ya que los valores en los rangos de los extremos re redujo drásticamente aumentándose a su vez los de los rangos intermedios, en particular el del 6to rango, de 6,96% a 41,05%. Siendo consistente con esto, no hubo ningún caso en la etapa de validación con precio de oferta por fuera del rengo establecido por los precios de los datos de entrenamiento.

PAIPA IV:

Kernel	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*
C	*2*	**C**	*16*	**C**	*2*	**C**	*128*	**C**	*8*

	Parám. Kernel	0.125	Parám. Kernel	0.0883	Parám. Kernel	0.0883	Parám. Kernel	0.25	Parám. Kernel	0.1768
	% Val. Cruzada	82.06	% Val. Cruzada	81.36	% Val. Cruzada	81.50	% Val. Cruzada	81.14	% Val. Cruzada	81.94
	ENTRENAMIENTO		ENTRENAMIENTO		ENTRENAMIENTO		ENTRENAMIENTO		ENTRENAMIENTO	
INDICES	Valor [%]		Valor [%]		Valor [%]		Valor [%]		Valor [%]	
Eficiencia 0	89.76		98.50		93.79		94.82		91.14	
Eficiencia 1	94.25		99.31		96.55		97.70		95.40	
Desv3clase	4.37		0.58		2.65		1.84		3.34	
% casos por rango	% aciertos en cada rango		% aciertos en cada rango		% aciertos en cada rango		% aciertos en cada rango		% aciertos en cada rango	
12.08	88.57		99.05		92.38		92.38		90.48	
36.36	98.73		99.37		99.05		98.10		98.42	
32.68	91.20		97.89		95.42		93.31		90.85	
5.29	82.61		100.00		89.13		93.48		86.96	
0.69	33.33		100.00		66.67		100.00		66.67	
1.73	80.00		100.00		93.33		93.33		86.67	
3.22	64.29		92.86		75.00		92.86		75.00	
4.26	72.97		97.30		83.78		91.89		75.68	
0.58	0.00		80.00		20.00		80.00		20.00	
3.11	70.37		100.00		81.48		92.59		77.78	
	VALIDACIÓN		VALIDACIÓN		VALIDACIÓN		VALIDACIÓN		VALIDACIÓN	
INDICES	Valor [%]		Valor [%]		Valor [%]		Valor [%]		Valor [%]	
Eficiencia 0	42.37		44.86		43.30		28.35		39.88	
Eficiencia 1	76.32		78.82		78.19		53.89		69.47	
Desv3clase	14.33		14.33		14.33		28.66		17.76	
Out Infer.	0.00		0.00		0.00		0.00		0.00	
Out Super.	0.00		0.00		0.00		0.00		0.00	
% casos por rango	% aciertos en cada rango		% aciertos en cada rango		% aciertos en cada rango		% aciertos en cada rango		% aciertos en cada rango	
7.48	0.00		0.00		0.00		16.67		0.00	
14.64	12.77		12.77		0.00		10.64		21.28	
43.30	93.53		99.28		100.00		58.99		84.89	
20.25	0.00		0.00		0.00		0.00		0.00	
0.00	NaN		NaN		NaN		NaN		NaN	
0.31	0.00		0.00		0.00		0.00		0.00	
10.28	0.00		0.00		0.00		0.00		0.00	
0.00	NaN		NaN		NaN		NaN		NaN	
3.74	0.00		0.00		0.00		0.00		0.00	
0.00	NaN		NaN		NaN		NaN		NaN	

Tabla 8.20 Resultados del entrenamiento y validación con MSV para Paipa IV

Se resalta en la etapa de clasificación la obtención de valores de eficiencia superior al 98% para un conjunto de parámetros, el cual no corresponde al que ofreció el mejor resultado en la validación cruzada; para este, la eficiencia 0 es 89,76%, el valor más bajo de entre los resultados mostrados.

De manera opuesta, en la etapa de validación las eficiencias más altas (Aproximadamente 78%) se dieron con las máquinas cuyas eficiencias en la etapa de clasificación fueron las más bajas. Con dichas máquinas el valor de "desv3clase" fue 14,33%, lo cual es un valor relativamente bajo comparado con los resultados encontrados con las restantes centrales de estudio.

De los rangos de precios se observa como en el tercero de estos acumula un 43,3% de los casos de validación, y las MSV entrenadas tienen un alto porcentaje de aciertos para dicho rango, 100% para una de ellas. Para los restantes rangos, los porcentajes de acierto en su mayoría fueron cero.

Los cambios de concentración de los porcentajes de datos en cada rango no variaron significativamente así como no hubo ningún caso de validación con un precio por encima o por debajo de los precios de los casos de entrenamiento.

FLORES:

	Kernel	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*
	C	*256*	**C**	*64*	**C**	*32*	**C**	*8*	**C**	*16*
	Parám. Kernel	*0.0625*	**Parám. Kernel**	*0.125*	**Parám. Kernel**	*0.0883*	**Parám. Kernel**	*0.0625*	**Parám. Kernel**	*0.0883*
	% Val. Cruzada	*69.66*	**% Val. Cruzada**	*71.16*	**% Val. Cruzada**	*70.80*	**% Val. Cruzada**	*69.76*	**% Val. Cruzada**	*70.11*
	ENTRENAMIENTO		**ENTRENAMIENTO**		**ENTRENAMIENTO**		**ENTRENAMIENTO**		**ENTRENAMIENTO**	
INDICES	**Valor [%]**		**Valor [%]**		**Valor [%]**		**Valor [%]**		**Valor [%]**	
Eficiencia 0	100.00		98.74		99.43		98.97		98.39	
Eficiencia 1	100.00		98.85		99.54		99.08		98.62	
Desv3clase	0.00		1.03		0.34		0.80		1.26	
% casos por rango	**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**	
1.49	100.00		100.00		100.00		100.00		100.00	
1.15	100.00		100.00		100.00		100.00		100.00	
10.00	100.00		96.55		98.85		95.40		95.40	
22.07	100.00		99.48		100.00		100.00		98.96	
2.99	100.00		96.15		96.15		96.15		96.15	
10.35	100.00		97.78		98.89		98.89		97.78	
8.51	100.00		98.65		100.00		100.00		100.00	
3.22	100.00		100.00		100.00		100.00		100.00	
20.58	100.00		98.88		98.88		98.88		97.77	
19.66	100.00		99.42		100.00		99.42		99.42	
	VALIDACIÓN		**VALIDACIÓN**		**VALIDACIÓN**		**VALIDACIÓN**		**VALIDACIÓN**	
INDICES	**Valor [%]**		**Valor [%]**		**Valor [%]**		**Valor [%]**		**Valor [%]**	
Eficiencia 0	0.00		19.10		0.00		0.30		0.90	
Eficiencia 1	23.88		44.78		23.88		24.18		25.08	
Desv3clase	76.12		50.15		74.33		75.52		72.54	
Out Infer.	0.00		0.00		0.00		0.00		0.00	
Out Super.	25.08		25.08		25.08		25.08		25.08	
% casos por rango	**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**	
6.2687	0.00		0.00		0.00		0.00		0.00	
2.6866	0.00		0.00		0.00		0.00		0.00	
25.97	0.00		6.90		0.00		0.00		0.00	
36.716	0.00		47.15		0.00		0.81		2.44	
2.6866	0.00		0.00		0.00		0.00		0.00	
0.59701	0.00		0.00		0.00		0.00		0.00	
0	NaN		NaN		NaN		NaN		NaN	
0	NaN		NaN		NaN		NaN		NaN	
0	NaN		NaN		NaN		NaN		NaN	
0	NaN		NaN		NaN		NaN		NaN	

Tabla 8.21 Resultados del entrenamiento y validación con MSV para Termo Flores.

Los valores de validación cruzada mostrados se encuentran alrededor del 70 % y en todos los casos los resultados de eficiencias de los datos de entrenamiento están por encima del 98%, inclusive se obtuvo un valor de 100% de eficiencia 0 con el conjunto de parámetros que brindó el menor resultado en la validación cruzada.

Observando las eficiencias de los datos de validación, la situación se revierte completamente ya que la eficiencia 1 más alta es apenas 44,78 % y la eficiencia 0 más alta apenas sobrepasó el 19% en uno de los casos mostrados. Los restantes casos mostraron Eficiencia 0 cercanas al 0%. En 4 de los 5 casos presentados, la "desv3clase" está por encima del 72% indicando una vasta diferencia entre los casos clasificados y los precios fijados por la central para los casos de validación.

Un factor importante para que se de lo expuesto en el párrafo anterior, es que un cuarto de los datos de validación tuvieron precios superiores al más alto de los casos de entrenamiento y se pasó de tener en la etapa de entrenamiento un poco más de un 50% de los datos concentrados en los 4 rangos de precios más altos a no tener ningún caso en estos rangos en los casos de validación.

FLORES 3:

	Kernel	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*
	C	*64*	**C**	*16*	**C**	*32*	**C**	*16*	**C**	*32*
	Parám. Kernel	*0.25*	**Parám. Kernel**	*0.125*	**Parám. Kernel**	*0.1768*	**Parám. Kernel**	*0.0883*	**Parám. Kernel**	*0.125*
	% Val. Cruzada	*79.07*	**% Val. Cruzada**	*89.70*	**% Val. Cruzada**	*79.41*	**% Val. Cruzada**	*79.98*	**% Val. Cruzada**	*80.00*
	ENTRENAMIENTO		**ENTRENAMIENTO**		**ENTRENAMIENTO**		**ENTRENAMIENTO**		**ENTRENAMIENTO**	
INDICES	**Valor [%]**		**Valor [%]**		**Valor [%]**		**Valor [%]**		**Valor [%]**	
Eficiencia 0	93.61		96.17		95.47		98.84		97.79	
Eficiencia 1	94.54		96.98		96.52		99.42		98.37	
Desv3clase	4.88		2.79		3.25		0.35		1.39	
% casos por rango	**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**	
2.79	100.00		100.00		100.00		100.00		100.00	
1.28	63.64		81.82		81.82		90.91		90.91	
4.99	88.37		95.35		93.02		97.67		97.67	
26.37	93.83		95.60		95.15		98.68		97.80	
4.76	92.68		95.12		95.12		97.56		97.56	
3.37	93.10		100.00		100.00		100.00		100.00	
7.67	96.97		96.97		96.97		98.49		96.97	
23.35	88.06		93.53		91.54		99.01		96.02	
20.09	99.42		100.00		99.42		100.00		100.00	
5.34	100.00		95.65		97.83		97.83		97.83	
	VALIDACIÓN		**VALIDACIÓN**		**VALIDACIÓN**		**VALIDACIÓN**		**VALIDACIÓN**	
INDICES	**Valor [%]**		**Valor [%]**		**Valor [%]**		**Valor [%]**		**Valor [%]**	
Eficiencia 0	9.64		2.20		9.64		0.00		1.93	
Eficiencia 1	11.85		3.03		11.57		0.00		2.75	
Desv3clase	60.61		73.00		60.61		93.66		73.83	

Out Infer.	0.00	0.00	0.00	0.00	0.00
Out Super.	37.19	37.19	37.19	37.19	37.19
% casos por rango	**% aciertos en cada rango**	**% aciertos en cada rango**	**% aciertos en cada rango**	**% aciertos en cada rango**	**% aciertos en cada rango**
0.28	0.00	0.00	0.00	0.00	0.00
11.85	0.00	0.00	0.00	0.00	0.00
1.65	0.00	0.00	0.00	0.00	0.00
3.31	0.00	0.00	0.00	0.00	0.00
44.35	21.74	4.97	21.74	0.00	4.35
1.38	0.00	0.00	0.00	0.00	0.00
0.00	NaN	NaN	NaN	NaN	NaN
0.00	NaN	NaN	NaN	NaN	NaN
0.00	NaN	NaN	NaN	NaN	NaN
0.00	NaN	NaN	NaN	NaN	NaN

Tabla 8.22 Resultados del entrenamiento y validación con MSV para Flores 3.

El valor de validación cruzada más alto en el caso de Flores 3 es claramente superior que los valores que le siguieron, los cuales se presentan aquí. Sin embargo, las eficiencias con los casos de entrenamiento son similares, superiores al 93%.

Muy al contrario de esto, los valores de eficiencias para los casos de validación son verdaderamente bajos, no alcanzan el 12% ni para el índice "Eficiencia 1". En concordancia, el valor de "desv3clase" fue siempre superior al 60%.

Es de resaltar el 37,19% del índice "outsuperior" y el notorio cambio de la distribución de la concentración de los precios en los casos de la fase de validación, en comparación con los de la etapa de entrenamiento.

TCENTRO:

	Kernel	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*
	C	*256*	**C**	*64*	**C**	*4*	**C**	*4*	**C**	*16*
	Parám. Kernel	*1*	**Parám. Kernel**	*0.50*	**Parám. Kernel**	*0.1768*	**Parám. Kernel**	*0.25*	**Parám. Kernel**	*0.3535*
	% Val. Cruzada	*83.50*	**% Val. Cruzada**	*84.77*	**% Val. Cruzada**	*83.26*	**% Val. Cruzada**	*83.50*	**% Val. Cruzada**	*84.09*
	ENTRENAMIENTO		**ENTRENAMIENTO**		**ENTRENAMIENTO**		**ENTRENAMIENTO**		**ENTRENAMIENTO**	
INDICES	**Valor [%]**		**Valor [%]**		**Valor [%]**		**Valor [%]**		**Valor [%]**	
Eficiencia 0	89.26		92.15		94.00		90.65		91.69	
Eficiencia 1	91.92		94.00		95.15		93.07		93.76	
Desv3clase	6.93		5.08		4.04		5.66		5.08	
% casos por rango	**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**	
1.73	33.33		40.00		46.67		26.67		40.00	
3.93	11.77		38.24		50.00		32.35		41.18	
1.96	52.94		64.71		76.47		41.18		52.94	
9.35	97.53		98.77		98.77		97.53		98.77	
26.56	97.83		98.70		98.26		98.26		98.26	
2.66	73.91		78.26		91.30		78.26		78.26	
2.89	96.00		96.00		92.00		92.00		92.00	
11.09	84.38		86.46		90.63		86.46		86.46	
34.76	96.35		97.67		98.34		97.01		97.34	
5.08	88.64		95.46		100.00		95.46		95.46	
	VALIDACIÓN		**VALIDACIÓN**		**VALIDACIÓN**		**VALIDACIÓN**		**VALIDACIÓN**	

INDICES	Valor [%]	Valor [%]	Valor [%]	Valor [%]	Valor [%]
Eficiencia 0	0.55	0.55	0.00	0.00	0.83
Eficiencia 1	1.38	1.66	0.55	0.83	1.93
Desv3clase	12.71	15.19	11.60	18.51	22.65
Out Infer.	0.00	0.00	0.00	0.00	0.00
Out Super.	88.12	88.12	88.12	88.12	88.12
% casos por rango	**% aciertos en cada rango**	**% aciertos en cada rango**	**% aciertos en cada rango**	**% aciertos en cada rango**	**% aciertos en cada rango**
0.00	NaN	NaN	NaN	NaN	NaN
5.80	9.52	9.52	0.00	0.00	14.29
5.52	0.00	0.00	0.00	0.00	0.00
0.00	NaN	NaN	NaN	NaN	NaN
0.00	NaN	NaN	NaN	NaN	NaN
0.00	NaN	NaN	NaN	NaN	NaN
0.00	NaN	NaN	NaN	NaN	NaN
0.00	NaN	NaN	NaN	NaN	NaN
0.00	NaN	NaN	NaN	NaN	NaN
0.55	0.00	0.00	0.00	0.00	0.00

Tabla 8.23 Resultados del entrenamiento y validación con MSV para Tcentro

Los resultados de Termocentro muestran valores de eficiencias que sobrepasan el 90% con los datos de entrenamiento, gracias a la gran cantidad de aciertos en los rangos de precios donde más se concentraron los datos. Cuando se pasa a explorar las eficiencias de los datos de validación se encuentran los valores más bajos de entre las centrales estudiadas: "eficiencia 0" cercana al 0% y "eficiencia 1" menor al 2%.

Aparte de esto, lo más resaltable de estos resultados, es el que el 88% de los datos de validación presentaron precios de oferta mayores al precio más alto en los datos de entrenamiento, haciendo que la clasificación de la MSV fuera totalmente ineficiente.

TASAJERO:

	Kernel	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*	**Kernel**	*RBF*
	C	*64*	**C**	*16*	**C**	*2*	**C**	*2*	**C**	*8*
	Parám. Kernel	*0.7071*	**Parám. Kernel**	*0.3535*	**Parám. Kernel**	*0.1768*	**Parám. Kernel**	*0.5*	**Parám. Kernel**	*0.25*
	% Val. Cruzada	*82.82*	**% Val. Cruzada**	*82.81*	**% Val. Cruzada**	*82.36*	**% Val. Cruzada**	*81.47*	**% Val. Cruzada**	*82.71*
	ENTRENAMIENTO		**ENTRENAMIENTO**		**ENTRENAMIENTO**		**ENTRENAMIENTO**		**ENTRENAMIENTO**	
INDICES	**Valor [%]**		**Valor [%]**		**Valor [%]**		**Valor [%]**		**Valor [%]**	
Eficiencia 0	86.46		87.71		87.03		82.71		88.28	
Eficiencia 1	89.53		90.56		90.44		86.58		90.90	
Desv3clase	6.94		5.92		6.03		7.85		5.80	
% casos por rango	**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**		**% aciertos en cada rango**	
1.25	0.00		9.09		0.00		0.00		9.09	
54.27	98.32		98.74		98.32		98.11		98.74	
10.01	76.14		85.23		86.36		79.55		86.36	
20.25	94.38		96.63		97.19		95.51		97.19	
4.78	59.52		50.00		45.24		0.00		54.76	
3.41	63.33		63.33		66.67		50.00		66.67	
1.71	13.33		20.00		13.33		0.00		20.00	

0.57	20.00	20.00	20.00	0.00	20.00
2.73	16.67	16.67	16.67	16.67	16.67
1.02	55.56	44.44	11.11	0.00	44.44
	VALIDACIÓN	**VALIDACIÓN**	**VALIDACIÓN**	**VALIDACIÓN**	**VALIDACIÓN**
INDICES	**Valor [%]**	**Valor [%]**	**Valor [%]**	**Valor [%]**	**Valor [%]**
Eficiencia 0	11.11	14.41	36.04	52.25	14.72
Eficiencia 1	42.94	35.14	67.87	84.38	35.74
Desv3clase	24.32	25.53	11.71	8.11	23.42
Out Super.	0.00	0.00	0.00	0.00	0.00
Out Infer.	0.00	0.00	0.00	0.00	0.00
% casos por rango	**% aciertos en cada rango**	**% aciertos en cada rango**	**% aciertos en cada rango**	**% aciertos en cada rango**	**% aciertos en cada rango**
0.00	NaN	NaN	NaN	NaN	NaN
0.90	0.00	0.00	0.00	0.00	0.00
86.19	7.67	10.80	37.98	58.89	10.11
2.10	42.86	28.57	28.57	71.43	42.86
10.51	34.29	42.86	25.71	0.00	48.57
0.00	NaN	NaN	NaN	NaN	NaN
0.30	0.00	0.00	0.00	0.00	0.00
0.00	NaN	NaN	NaN	NaN	NaN
0.00	NaN	NaN	NaN	NaN	NaN
0.00	NaN	NaN	NaN	NaN	NaN

Tabla 8.24 Resultados del entrenamiento y validación con MSV para Tasajero

En estos resultados, analizando los datos de la etapa de entrenamiento, se encuentran valores de eficiencia 0 y eficiencia 1de aproximadamente 88% y 90%, respectivamente, para los mejores casos. La concentración de precios en los diferentes rangos está marcada por el alto porcentaje en el segundo de estos, para le cual la MSV logra un alto porcentaje de aciertos.

A pesar de la paridad en los porcentajes de validación cruzada y las eficiencias de los casos entrenamiento, si se apreció una marcada diferencia en los resultados para un conjunto de parámetros con los casos de validación; con este, las eficiencias del entrenamiento son 52,25% y 84,38% y "desv3clase" no alcanza a ser 9%. Estos son los valores más altos de eficiencia en validación para alguna de las térmicas estudiadas en esta invesigación.

Es necesario resaltar, con los datos de validación, que el 86,19% de los casos corresponden al tercer rango de precios y que no hubo ningún caso de la fase de validación cuyo precio se saliera del rango demarcado por los precios de la etapa de entrenamiento.

8.3. Resultados del clasificador Bayesiano de Naives a partir de Mezclas Finitas.

8.3.1. Aplicación del clasificador de Naives

Para la aplicación de esta etapa de la investigación se llevo a cabo una combinación de dos técnicas: primero se implementó el algoritmo EA de mezclas finitas para obtener la función de densidad de probabilidad de los datos de estudio. Posteriormente, con este resultado, se implementó el

clasificador bayesiano de Naives para llevar a cabo la clasificación de los casos de validación.

El procesamiento de los datos en esta etapa se basó inicialmente en el código presentado en [26] en el cual se implementa el algoritmo EM, con el cual se realizaron unas pruebas iniciales; Posteriormente se desarrolló un conjunto de programas y funciones en Matlab adaptada a las características de los datos de entrada y a lo que se esperaba obtener como salida. Sobre está misma herramienta, se agregaron las funciones que permitieron realizar la clasificación con Naives.

En suma, se creó una herramienta en Matlab que permite entrar los datos con los que se contó para esta investigación y que como salida final, entrega una tabla similar a la que se diseño para los resultadotes de las MSV, es decir, arroja los valores de los indicadores que se explican en el capítulo 6 tanto para los datos de entrenamiento como para los datos de validación.

8.3.2. Resultados centrales hidráulicas

En las tablas 8.25 a 8.29 se muestran para las centrales hidráulicas estudiadas los resultados de la clasificación realizada con el clasificador Bayesiano de Naives a partir de las funciones de densidad de probabilidad de los grupos etiquetados obtenidos con la técnica de mezclas finitas implementada con el algoritmo EM. No hay presentes resultados de Alban debido a que la implementación del Algoritmo EM para esta central no convergió, por ende, no se pudo realzar la clasificación con Naives-Bayes.

Los resultados obtenidos aquí, comparados con los que dejaron las MSV muestran que este clasificador, implementado en la forma en que se planteó y llevo a cabo, no entrega buenos porcentajes de acierto con los datos de entrenamiento; solo San Carlos muestra un valor de "Eficiencia 0" cercano al 60% y las restantes centrales dieron valores claramente más bajos.

Lo que se busca es encontrar patrones de comportamiento en la fijación de los precios y usar los índices como señales de cambio de parte de las centrales; lo expuesto en el párrafo anterior permite cuestionarse si los índices que se calcularon para los datos de validación pueden efectivamente cumplir ese propósito.

Por otra parte tampoco se encuentra para alguna de las centrales, que alguno de sus rangos de precios hubiese sido bien explicado por esta técnica lo que reafirma que ésta no favoreció el hallazgo de patrones y, alguna conclusión específica basada en estos resultados sería fácilmente debatible.

GUATRÓN:

MODELO 9			
GUATRÓN			
VALIDACIÓN		**ENTRENAMIENTO**	
Eficiencia 0	13.02	***Eficiencia 0***	48.35
Eficiencia 1	37.95	***Eficiencia 1***	78.45
Desv3clase	45.15	***Desv3clase***	9.58
Out Inferior	5.26	***Out Inferior***	0.00

Out superior	7.76	***Out superior***	0.00
% casos por rango	% aciertos c/rango	% casos por rango	% aciertos c/rango
9.70	25.71	5.47	91.67
6.93	0.00	11.29	68.69
8.03	0.00	11.52	53.47
10.25	51.35	23.03	43.07
5.82	4.76	19.38	35.29
11.36	4.88	13.91	54.92
12.47	0.00	7.53	12.12
4.43	0.00	4.22	48.65
10.25	0.00	2.05	61.11
7.76	57.14	1.60	50.00

Tabla 8.25 Resultados clasificador Naives-Bayes para Guatrón.

CHIVOR:

MODELO 9			
CHIVOR			
VALIDACIÓN		**ENTRENAMIENTO**	
Eficiencia 0	40.50	***Eficiencia 0***	44.34
Eficiencia 1	46.28	***Eficiencia 1***	57.18
Desv3clase	46.01	***Desv3clase***	39.32
Out Inferior	0.00	***Out Inferior***	0.00
Out superior	7.99	***Out superior***	0.00
% casos por rango	% aciertos c/rango	% casos por rango	% aciertos c/rango
47.66	79.19	43.41	15.59
1.93	0.00	1.75	100.00
3.31	0.00	2.57	95.46
6.61	0.00	9.68	85.54
3.31	0.00	2.80	95.83
4.68	11.77	3.73	50.00
4.96	0.00	14.70	44.44
3.31	25.00	10.27	61.36
7.71	10.71	5.83	54.00
8.54	6.45	5.25	86.67

Tabla 8.26 Resultados clasificador Naives-Bayes para Chivor.

GUAVIO:

MODELO 9			
GUAVIO			
VALIDACIÓN		**ENTRENAMIENTO**	
Eficiencia 0	5.79	***Eficiencia 0***	6.63
Eficiencia 1	21.21	***Eficiencia 1***	9.26
Desv3clase	67.49	***Desv3clase***	88.69
Out Inferior	0.00	***Out Inferior***	0.00
Out superior	2.75	***Out superior***	0.00
% casos por rango	% aciertos c/rango	% casos por rango	% aciertos c/rango
31.41	0.00	19.66	0.58

2.75	0.00	2.74	4.17
1.65	0.00	8.34	0.00
4.96	0.00	15.09	13.64
9.37	0.00	19.43	1.76
8.26	10.00	22.06	8.29
11.02	5.00	9.49	1.20
12.67	2.17	1.83	50.00
11.02	0.00	0.80	71.43
4.13	100.00	0.57	100.00

Tabla 8.27 Resultados clasificador Naives-Bayes para Guavio.

PORCE 2:

MODELO 9			
PORCE II			
VALIDACIÓN		**ENTRENAMIENTO**	
Eficiencia 0	10.74	***Eficiencia 0***	37.88
Eficiencia 1	37.19	***Eficiencia 1***	60.35
Desv3clase	49.59	***Desv3clase***	27.77
Out Inferior	0.00	***Out Inferior***	0.00
Out superior	19.28	***Out superior***	0.00
% casos por rango	% aciertos c/rango	% casos por rango	% aciertos c/rango
36.92	6.72	19.88	11.83
1.93	0.00	1.18	60.00
0.55	0.00	1.06	55.56
1.65	0.00	9.06	70.13
5.79	4.76	18.71	33.33
4.41	37.50	16.47	35.71
9.37	23.53	19.88	40.83
8.26	0.00	6.24	66.04
6.06	0.00	5.76	55.10
5.79	71.43	1.76	20.00

Tabla 8.28 Resultados clasificador Naives-Bayes para Porce 2.

SAN CARLOS:

MODELO 9			
SAN CARLOS			
VALIDACIÓN		**ENTRENAMIENTO**	
Eficiencia 0	26.45	***Eficiencia 0***	59.50
Eficiencia 1	60.06	***Eficiencia 1***	86.35
Desv3clase	20.11	***Desv3clase***	6.37
Out Inferior	0.00	***Out Inferior***	0.00
Out superior	8.54	***Out superior***	0.00
% casos por rango	% aciertos c/rango	% casos por rango	% aciertos c/rango
9.92	13.89	2.50	63.64
1.65	0.00	1.59	85.71
5.51	15.00	2.50	90.91
7.16	*38.46*	19.91	73.71
11.30	*39.02*	17.63	54.19
6.06	*9.09*	17.18	49.01

9.09	*30.30*	18.77	55.15
23.42	*29.41*	15.70	52.17
10.47	*18.42*	2.50	90.91
6.89	*72.00*	1.71	46.67

Tabla 8.29 Resultados clasificador Naives-Bayes para San Carlos.

8.3.3. Resultados centrales térmicas

En las tablas 8.30 a 8.34 se muestran para las centrales térmicas estudiadas los resultados de la clasificación realizada con el clasificador Bayesiano de Naives a partir de las funciones de densidad de probabilidad de los grupos etiquetados obtenidos con la técnica de mezclas finitas implementada con el algoritmo EM. No hay presentes resultados de Paipa 4 debido a que la implementación del Algoritmo EM para esta central no convergió, por ende, no se pudo realzar la clasificación con Naives-Bayes.

Al igual que para las MSV, los porcentajes de acierto con los datos de clasificación para las centrales térmicas son un poco más altos que para las hidráulicas; sin embargo los valores obtenidos con esta técnica en ningún caso alcanzan ni el 85% y solo para Tebsa y Termocentro sobrepasan apenas el 80%. Los valores de "Eficiencia 1" no son mucho mejores tampoco

De manera similar a lo que se comentó con las centrales hidráulicas, los resultados de eficiencias aquí encontrados son significativamente menores que los obtenidos con las MSV y por ello no resultó ser un elemento de aporte para encontrar los patrones de comportamiento de la oferta.

TEBSA:

TEBSA			
VALIDACIÓN		**ENTRENAMIENTO**	
Eficiencia 0	7.44	***Eficiencia 0***	82.83
Eficiencia 1	39.39	***Eficiencia 1***	90.84
Desv3clase	47.11	***Desv3clase***	7.31
Out Inferior	0.00	***Out Inferior***	0.00
Out superior	0.00	***Out superior***	0.00
%Casos/Rang	%Acierto/Rang	%Casos/Rang	%Acierto/Rang
1.65	0.00	11.25	76.29
0.00	NaN	3.60	83.87
14.33	0.00	1.04	88.89
13.50	0.00	1.28	100.00
9.64	0.00	7.19	88.71
41.05	0.00	6.96	73.33
7.71	96.43	25.06	87.04
12.12	0.00	31.67	76.19
0.00	NaN	2.32	100.00
0.00	NaN	9.63	96.39

Tabla 8.30 Resultados clasificador Naives-Bayes para Tebsa.

FLORES:

FLORES	
VALIDACIÓN	**ENTRENAMIENTO**

Eficiencia 0	2.09	***Eficiencia 0***	41.03
Eficiencia 1	7.46	***Eficiencia 1***	45.06
Desv3clase	75.22	***Desv3clase***	48.85
Out Inferior	0.00	***Out Inferior***	0.00
Out superior	25.08	***Out superior***	0.00
%Casos/Rang	%Acierto/Rang	%Casos/Rang	%Acierto/Rang
6.27	9.52	1.49	100.00
2.69	0.00	1.15	100.00
25.97	0.00	10.00	11.49
36.72	4.07	22.07	22.40
2.69	0.00	2.99	76.92
0.60	0.00	10.35	51.11
0.00	NaN	8.51	67.57
0.00	NaN	3.22	78.57
0.00	NaN	20.58	35.75
0.00	NaN	19.66	46.20

Tabla 8.31 Resultados clasificador Naives-Bayes para Flores.

FLORES 3:

FLORES 3			
VALIDACIÓN		**ENTRENAMIENTO**	
Eficiencia 0	7.44	***Eficiencia 0***	69.92
Eficiencia 1	9.37	***Eficiencia 1***	78.17
Desv3clase	87.05	***Desv3clase***	19.40
Out Inferior	0.00	***Out Inferior***	0.00
Out superior	37.19	***Out superior***	0.00
%Casos/Rang	%Acierto/Rang	%Casos/Rang	%Acierto/Rang
0.28	0.00	2.79	100.00
11.85	0.00	1.28	100.00
1.65	0.00	4.99	79.07
3.31	0.00	26.37	70.04
44.35	16.77	4.76	68.29
1.38	0.00	3.37	100.00
0.00	NaN	7.67	75.76
0.00	NaN	23.35	53.23
0.00	NaN	20.09	67.63
0.00	NaN	5.34	93.48

Tabla 8.32 Resultados clasificador Naives-Bayes para Flores 3.

TERMOCENTRO:

T. CENTRO			
VALIDACIÓN		**ENTRENAMIENTO**	
Eficiencia 0	1.38	***Eficiencia 0***	80.25
Eficiencia 1	2.49	***Eficiencia 1***	85.57
Desv3clase	32.60	***Desv3clase***	11.66
Out Inferior	0.00	***Out Inferior***	0.00
Out superior	88.12	***Out superior***	0.00
%Casos/Rang	%Acierto/Rang	%Casos/Rang	%Acierto/Rang
0.00	NaN	1.73	66.67
5.80	0.00	3.93	64.71
5.52	25.00	1.96	58.82

0.00	NaN	9.35	100.00
0.00	NaN	26.56	83.04
0.00	NaN	2.66	100.00
0.00	NaN	2.89	100.00
0.00	NaN	11.09	90.63
0.00	NaN	34.76	67.77
0.55	0.00	5.08	95.46

Tabla 8.33 Resultados clasificador Naives-Bayes para Termocentro.

TERMOTASAJERO:

TASAJERO			
VALIDACIÓN		**ENTRENAMIENTO**	
Eficiencia 0	7.21	***Eficiencia 0***	40.61
Eficiencia 1	12.01	***Eficiencia 1***	45.39
Desv3clase	85.89	***Desv3clase***	49.15
Out Inferior	0.00	***Out Inferior***	0.00
Out superior	0.00	***Out superior***	0.00
%Casos/Rang	%Acierto/Rang	%Casos/Rang	%Acierto/Rang
0.00	NaN	1.25	9.09
0.90	33.33	54.27	44.86
86.19	8.01	10.01	42.05
2.10	0.00	20.25	21.35
10.51	0.00	4.78	83.33
0.00	NaN	3.41	53.33
0.30	0.00	1.71	26.67
0.00	NaN	0.57	100.00
0.00	NaN	2.73	12.50
0.00	NaN	1.02	44.44

Tabla 8.34 Resultados clasificador Naives-Bayes para Tasajero.

9.Conclusiones, aportes y trabajos futuros

9.1. Conclusiones

- Las variables seleccionadas para los análisis mostraron ser linealmente independientes ya que el análisis de componentes principales ACP no contribuyó a la reducción de estas, con lo que se favorece la implementación del análisis cluster.

- La curva de demanda residual es una herramienta capaz de sintetizar para un determinada central, los elementos de la competencia que esta tiene en cuenta para realizar su casación de precios en el mercado 'spot' de energía, ya que los descriptores obtenidos de la construcción de estas curvas estuvieron siempre bien posicionados en el ranking de variables respecto al precio de oferta para cada una de las centrales estudiadas.

- Las curvas de demanda residual son útiles para modelar y estudiar el efecto de las ofertas de la competencia sobre el mercado y el precio de bolsa, facilitando el trabajo de los entes interesados en la regulación y de las mismas centrales ya que pueden resumir en unas pocas variables toda esa gran cantidad de información que representan las ofertas de los agentes participantes en la bolsa de energía.

- La formación de conglomerados de datos conformados por los diferentes estados de las variables externas a una central en el MEM y agrupados por su similitud de acuerdo a criterios de distancia, no son útiles para describir patrones de comportamiento en la casación de los precios de oferta por parte de dichas centrales, ya que no se logró estar cerca a encontrar algún precio característico para dichos conglomerados

- La obtención de rangos de precios de oferta de las centrales generadoras para utilizarse en herramientas de análisis tales como MSV y el clasificador Bayesiano de Naives, debe llevarse a cabo usando una técnica tal como el análisis cluster univariado que permite encontrar asociaciones naturales de los datos no perceptibles a "simple vista".

- Los casos atípicos extraídos para las centrales hidráulicas correspondieron mayormente a precios altos yen ningún caso correspondieron con niveles bajos de los embalses como se presumía por parte de los autores, con lo que se confirma que estas centrales basan su estrategia de fijación de precios basadas principalmente en la situación del MEM. No se excluye sin embargo, que alguno de dichos casos correspondiese a alguna restricción técnica muy puntual de cada planta generadora.

- La no identificación de precios altos o bajos representando casos atípicos para las centrales térmicas se debió al amplio rango de precios que manejan estas respecto de su contraparte, las hidráulicas; esto se ratifica

en el hecho que las plantas térmicas son más ineficientes y que suelen ser seguidoras de precio, a diferencia de las plantas hidráulicas.

- Una MSV entrenada con los parámetros que ofrecen la validación cruzada más alta no necesariamente implica, que esta máquina es la mejor clasificadora de los datos con los que se entrenó; sin embargo, la validación cruzada junto la búsqueda en malla siguen siendo el mejor aliado para acercarse a los parámetros óptimos de entrenamiento de una MSV y se recomienda por ende, utilizar los mejores conjuntos de parámetros de acuerdo a este criterio en lugar de solo emplear el mejor.

- Los resultados de eficiencia de clasificación de los datos de entrenamiento con las MSV fueron buenos, siempre superior al 80%, lo que ratifica el poder de esta herramienta para llevar a cabo minería de datos de grandes cantidades de datos sin importar su dimensionalidad y sin ser muy costosa, computacionalmente hablando.

- Los resultados de eficiencias más bajas y de desviación más altas de parte de las centrales térmicas respecto de las hidráulicas, al realizar la clasificación con MSV de los datos correspondientes al periodo de validación, indican que las plantas térmicas cambiaron de manera más significativa sus estrategias de fijación de precios de oferta en la bolsa; esto se explica por el incremento de la demanda de energía en el período de validación respecto del período de entrenamiento (incremento promedio de 5,92% y 5,20% para las demandas medianas y máximas diarias) lo que implicó la entrada de recursos más costosos para satisfacer dicha demanda. Ello está ratificado en el incremento de la volatilidad del precio en bolsa de energía que se calculó para los periodos comparados (Incremento de 110,23 % para los precios medianos diarios y de 93,82% para los precios mínimos diarios).

- Los índices diseñados en este trabajo para sintetizar los patrones identificados en los precios de oferta de las centrales generadoras por parte de las máquinas clasificadoras, son de utilidad para los entes responsables de la monitorización del MEM ya que funcionan como alarmas que les permiten detectar cambios de comportamiento en las estrategias de casación de los precios en el mercado 'spot'.

- Dentro de las centrales térmicas destacó Paipa IV por los altos valores de eficiencia obtenidos en la clasificación de los datos del período de validación, respecto de las restantes cinco plantas térmicas estudiadas. Se explica esto por el hecho que esta planta funciona a base de carbón el cual es subsidiado lo que la hace más competitiva, facilitándole entrar a "jugar" más veces al mercado 'spot'; esto la obliga a estar más conectada con las ofertas en bolsa de la competencia y por ello esta central tiene un comportamiento que tiende hacia la labor que realizan las centrales hidráulicas.

- La identificación de algunos rangos de precios de oferta específicos para las centrales, en los que la MSV fue capaz de realizar una buena clasificación sus casos, es también de utilidad para los interesados en el seguimiento de las ofertas ya que dichos rangos corresponden en su mayoría a los grupos donde la central concentró la mayor cantidad de casos y además muestra un espectro de precios para el cual la central tiene en teoría un comportamiento regido por el estado de las variables externas identificadas; una disminución en los aciertos de dicho rango en determinado periodo de tiempo apuntan a algún cambio en el patrón de comportamiento por parte de la central estudiada.

- El clasificador Bayesiano de Naives basado en las distribuciones de probabilidad calculadas con el algoritmo EM resultó ser mucho menos efectivo que las MSV para la identificación de los patrones en los precios de oferta. Esto se debe a que el algoritmo original de Naives-Bayes hace una aproximación, de que las variables utilizadas tienen independencia lineal, lo cual se da con las variables utilizadas en el trabajo; al aplicar el algoritmo EM sobre los grupos etiquetados, algunos con muy pocos, puede darse una condición que afecta la aproximación y redunda en bajos porcentajes de aciertos. Sin embargo, se confirmó que las centrales térmicas fueron quienes modificaron en mayor medida su estrategia de fijación de precios.

9.2. Aportes

- Se diseñaron y calcularon índices basados en los patrones encontrados por las MSV que pueden ser útiles para los entes interesados en la monitorización ya que les sirven como señales de cambios en el comportamiento de una central para la fijación de los precios de oferta en bolsa.

- Se demostró que la curva de demanda residual facilita en gran manera el análisis del efecto de las ofertas de la competencia sobre el mercado 'spot'. La herramienta permite sintetizar una gran cantidad de variables necesarias para los estudios de la competencia, en unas pocas variables, de forma rápida y efectiva. Esta técnica además puede extrapolarse para ser usada en otros mercados en los que el precio se rija por la competencia, de una manera similar al de la bolsa de energía de Colombia.

9.3. Proyectos futuros

- Se deja abierta la puerta abierta para llevar a cabo trabajos que apunten a la predicción de los predicción de los precios de oferta en bolsa de energía de las centrales generadoras del país.

- Encontrar relaciones entre los rangos dados por las curvas de demanda residual calculados para las centrales generadoras y la volatilidad de los precios de bolsa.

- Los buenos resultados de las Máquinas de Soporte Vectorial abren un abanico de posibilidades de implementación de la herramienta en aplicaciones en las que se cuente con bases de datos de gran tamaño y que requieran de la minería de datos para explotar información contenida.

- La utilización de las curvas de demanda residual en otros mercados con similitudes al mercado 'spot' de energía, en los que se necesite estudiar las ofertas de la competencia y el efecto de estas.

10. BIBLIOGRAFÍA

[1] [Addepalli 2004] Addepalli. Introductory Primer on the Monitoring and Surveillance of Electric Power Markets. Philippine Energy Regulatory Commission. 2004

[2] [Fernández Pérez, 2002] Fernández José Carlos. "Análisis y evaluación de mercados eléctricos liberizados a escala internacional". 2002. Universidad Pontificia Comillas. Tesis de Maestría en Gestión Técnica y Económica en el sector eléctrico.

[3] [López Hernández, 2005] López Juan Felipe. "Predicción de Medio Plazo de las Pendientes de Curvas de Demanda Residual de un Agente". 2005. Universidad Pontificia Comillas. Tesis de Maestría en Gestión Técnica y Económica en el sector eléctrico.

[4] [UPME, 2005] Aguilar Argemiro, Díaz Javier. "Una Visión del Mercado Eléctrico Colombiano". 2004. Estudio realizado por la firma Sistemas Digitales de Control Ltda. Bajo la dirección de la UPME, parte del estudio "Mercado de Energía Eléctrica en Colombia – Análisis Comercial y de estrategias"

[5] [Morales, Gómez, 2005] Morales Peña Germán Andrés, Gómez Ruiz Álvaro. "Estudio e Implementación de una Herramienta Basada en Máquinas de Soporte vectorial aplicada a la Localización de Fallas en Sistemas de Distribución". Universidad Industrial de Santander, Tesis de Pregrado de Ingeniería Eléctrica, Dirigida por Hermann Vargas y Codirigida por Juan Carlos Rodríguez.

[6] [Hernández, Ramírez, 2004] Hernández Orallo; Ramírez Quintana; Ferri Ramírez, "Introducción a la minería de datos", Pearson Educación S.A., Madrid, 2004.

[7] [Hair, Anderson y otros, 2001] Hair Joseph; Anderson Rolph; Tatham Ronald; Black William, "Análisis Multivariante", Quinta edición, Prentice Hall, Madrid, 2001.

[8] [Brugman, Wolak, 2004] Alberto Brugman Miramón, Frank Wolak y otros, "Diseño y estructuración de una metodología para el monitoreo y control del mercado de energía mayorista – MEM", Disponible en la página de la Superintendencia de Servicios Públicos Domiciliarios (SSPD), 2004

[9] [Reneses, 2004] J. Reneses, "Análisis de la operación de los mercados de generación de energía eléctrica a medio plazo", Tesis Doctoral, E.T.S. de Ingeniería (I.C.A.I.), Universidad Pontificia de Comillas, 2004, Dirigida por Julián Barquín Gil y Efraim Centeno Hernáez.

[10] [Bahon, Cecilio, 2005] Angulo Bahon, Cecilio, UPC. "Aprendizaje con máquinas núcleo en entornos de multiclasificación" http://www.tdx.cesca.es/TDX-0628101-141150/

[11] [Navarro 1, 2005] IMPROVEN CONSULTORES. "¿Qué es CRM?" http://www.improven-consultores.com/paginas/documentos_gratuitos/que_crm.php

[12] [Navarro 2, 2005] Eduardo Navarro, IMPROVEN CONSULTORES. "Vendiendo más y mejor, entendiendo CRM en la práctica" http://www.improven-consultores.com/paginas/documentos_gratuitos/CRMpractico.php

[13] Eduardo Navarro, IMPROVEN CONSULTORES. "Las realidades del CRM" http://www.improven-consultores.com/paginas/documentos_gratuitos/realidad_crm.php

[14] [Wikipedia, CRM, 2004] "CRM" http://es.wikipedia.org/wiki/CRM

[15] [López, 2005] Carlos López, Gestiopolis.com. "¿Sabes que es CRM?" http://www.gestiopolis.com/canales/gerencial/articulos/20/crm.htm

[16] [Cortijo 1, 2002] Francisco José Cortijo Bon. "Aprendizaje y reconocimiento de patrones", http://www-etsi2.ugr.es/depar/ccia/rf/www/tma1_00-01_www/node4.html

[17] [Cortijo 2, 2002] Francisco José Cortijo Bon. "Aproximaciones al reconocimiento de patrones" http://www-etsi2.ugr.es/depar/ccia/rf/www/tema1_00-01_www/node3.html

[18] [SPSS, 2005] "El análisis predictivo de SPSS" http://es.spps.com

[19] [CIER, 2003] CIER. "Atlas 2003 del desarrollo eléctrico de América del Sur"

[20] [Muñoz, Rodrigo, 2003] D. Muños-Díaz y F.S. Rodrigo, "Aplicación del análisis "cluster" para el estudio de la relación Nao-precipitaciones de invierno en el sur de la península ibérica", Depto. Física Aplicada, Universidad de Almería, 2003.

[21] [Cortes, Vapnik, 1995] C. Cortes and V. Vapnik. "Support vector networks. Machine Learning", (1995).

[22] [Vapnik, Stitson, 1996] V. Vapnik, M. O. Stitson, J. A. E. Weston, A. Gammerman, V. Vovk. "Theory of Support Vector Machines", Royal Holloway University of London, England, 1996.

[23] [Burges, 1998] Christopher Burges. "A tutorial on support vector machines for pattern recognition". Data Mining And Knowledge Discovery, 1998.

[24] [Albrecht, 2000] Karl Albrecht. El radar empresarial, 2000.

[25] [Berzal, de la Fuente, Gómez] David Berzal, Jose de la Fuente, Tomás Gómez. "Elaboración de estrategias competitivas de oferta para el mercado diario de energía", Universidad Pontificia de Comillas, Madrid.

[26] [Martínez, Martínez, 2005],Martínez Wendy L., Martínez Angel R. "Exploratory Data Análisis with Matlab", Chapman y Hall/ CRC, 2005

[27] [Cormane, 2006], Cormane Angarita, Jorge Andrés. "Modelo estadístico para la localización de fallas en sistemas de distribución de energía eléctrica". Universidad Industrial de Santander, Bucaramanga, 2006.

[28] [López de Castilla] Carlos López de Castilla Vásquez. "Clasificadores por redes bayesianas", Universidad de Puerto Rico, 2005.

[29] [Molina, 2002] Molina, Luis Carlos. "Feature Selection Algorithms: A Survey And Experimental Evaluation". Universidad Politécnica de Cataluña, 2002.

ANEXO A: Definición de variables

Durante el desarrollo de esta investigación se han utilizado variables que hacen parte e intervienen en la dinámica del MEM; para mayor claridad en cuanto a dichas variables a continuación se realiza una compilación de algunas de las definiciones dadas en el sistema NEON, el cual está a su vez basado en las resoluciones establecidas por la CREG.

- **Demanda Comercial** Considera la demanda propia de cada comercializador mas la participación en las pérdidas del sistema de transmisión nacional (STN) y los consumos propios de los generadores. Esta variable se da en kWh y la información requerida para su cálculo se obtiene mediante procesos de agregación de datos provenientes del Mercado de Energía Mayorista

- **Demanda Comercial No Regulada** Es la demanda de los comercializadores para atender sus clientes finales No Regulados, mas la participación en las perdidas del STN. Dicha información se obtiene como resultado de la operación del sistema interconectado nacional (SIN), con base en los datos reportados por los agentes del sector.

- **Demanda Real De Energía** Es la demanda de energía calculada con base en la generación real y los intercambios de energía, no incluye las pérdidas de transmisión. La información proviene directamente del Mercado de Energía Mayorista (MEM)

- **Disponibilidad De Generación** MW Es la cantidad de potencia neta que un generador puede suministrar al sistema durante un intervalo de tiempo determinado. Se mide en MW y estos datos son obtenidos de los eventos de recursos de generación reportados por los agentes del CND.

- **Embalse Ofertable** Es el margen resultante de la diferencia entre el nivel actual de un embalse y el nivel establecido como reserva (Minimo Operativo Superior MOS). El embalse ofertable mide la cantidad de energía hidráulica del país, disponible para transar en el Mercado Mayorista. Estos datos se calculan con base en la información histórica de cada río.

- **Generación Ideal** Se llama así al despacho de generación que resulta de considerar una red de transporte inexistente. Esta información se obtiene mediante procesos de agregación de datos provenientes del Mercado de Energía Mayorista (MEM)

- **Generación Programada** Es el despacho de generación esperado de las centrales, que resulta de considerar la red de transporte real, la unidad en que se mide esta variable es kWh. Al igual que en la generación ideal la información proviene de procesos de agregación de datos provenientes del MEM.

- **Generación Real** Es la generación neta de cada una de las plantas en sus puntos de frontera, incluyendo las importaciones Internacionales, la fuente de esta información es directamente el MEM, presenta un retraso de 5 días.

- **Generación Programada** Esta variable corresponde al despacho de generación esperado de las centrales; resulta de considerar la red de transporte real. La información se obtiene mediante procesos de agregación de datos provenientes del MEM.

- **Mínimo Operativo Inferior (% Energía)** Es el nivel mínimo requerido en un embalse, expresado en términos del porcentaje de energía, como reserva energética para cubrir condiciones hidrológicas criticas. Los niveles mínimos operativos son definidos mediante acuerdos.

- **Mínimo Operativo Inferior (Energía)** Corresponde al límite operativo de un embalse, por debajo del cual el precio de oferta de las plantas asociadas debe ser mayor que el CR01 del Sistema Interconectado Nacional en cada hora; la unidad de medida de esta variable es MWh. Estos niveles mínimos operativos también son definidos mediante acuerdos.

- **Mínimo Operativo Superior (% Energía)** Es el nivel mínimo requerido en un embalse, expresado en términos del porcentaje de energía, como reserva energética para cubrir condiciones hidrológicas criticas. Los niveles mínimos operativos son definidos mediante acuerdos.

- **Mínimo Operativo Superior (Energía)** Hace referencia al límite operativo de un embalse por debajo del cual sólo se permite utilizar la energía almacenada si todas las unidades térmicas están despachadas. Los niveles mínimos operativos son definidos mediante acuerdos

- **Nivel Del Embalse (Volumen)** Se define como la reserva de agua almacenada en un embalse de acuerdo con la cantidad de agua almacenada en el mismo. Ésta variable se mide en Millones de m^3. Los valores correspondientes a esta reserva de energía son reportados directamente por los agentes propietarios y/o administradores de los recursos energéticos.

- **Nivel Del Embalse (% Energía)** Es el porcentaje de reserva de energía de un embalse respecto a su capacidad total de acuerdo con la cantidad de agua almacenada. El porcentaje de reserva de energía se calcula con base en la reserva actual del embalse y su capacidad máxima.

- **Nivel Del Embalse (Energía)** Corresponde a la reserva de energía de un embalse de acuerdo con la cantidad de agua almacenada en el mismo. Las unidades usadas para medir esta energía son MWh. La reserva de energía es reportada directamente por los agentes propietarios y/o administradores de los recursos energéticos.

- **Costo Equivalente En Energía (CEE)** Esta unidad está dada en pesos por Kilowatt hora ($/kWh). Este costo es estimado del cargo por capacidad, la información necesaria para el cálculo de esta variable proviene directamente del Mercado de Energía Mayorista

- **Costo Equivalente En Energía (CERE)** Es el costo equivalente real de energía del cargo por capacidad: La información proviene directamente del Mercado de Energía Mayorista

- **Precio En Bolsa Nacional** En condiciones normales de operación, corresponde al precio de oferta incremental más alto de las plantas flexibles programadas en el despacho ideal para la hora de liquidación. Esta medido en pesos por kilowatt hora ($/kWh). La información acerca de esta variable proviene directamente de la Gerencia del Mercado de Energía Mayorista de ISA

- **Recaudo Cargo Por Capacidad** Es el valor recaudado de cargo por capacidad por cada submercado de generación, es medido en pesos.

- **Capacidad Efectiva Neta** Máxima cantidad de potencia neta que puede suministrar una unidad de generación en condiciones normales de operación. Incluye las menores y cogeneradores, su unidad de medida son los Gigawatts (GW). La información se obtiene del Centro Nacional de Despacho.

- **Despacho ideal**. Es la programación de generación que se realiza a posteriori por el Sistema de Intercambios Comerciales (SIC), la cual atiende la demanda real con la disponibilidad real de las plantas de generación. Este despacho se realiza considerando la oferta de precios por orden de méritos de menor a mayor, sin considerar las diferentes restricciones que existen en el sistema, excepto por las condiciones de inflexibilidad de las plantas generadoras."

- **Despacho programado**. Es el programa de generación que realiza el Centro Nacional de Despacho (CND), denominado Redespacho en el Código de Redes, para atender una predicción de demanda y sujeto a las restricciones del sistema, considerando la declaración de disponibilidad, la oferta en precios y asignando la generación por orden de méritos de menor a mayor."

- **Despacho real**. Es el programa de generación realmente efectuado por los generadores, el cual se determina con base en las mediciones en las fronteras de los generadores."

- **Disponibilidad Comercial**. Es la disponibilidad calculada por el SIC, la cual considera la declaración de disponibilidad de los generadores, modificada cuando se presenten cambios en las unidades de generación en la operación real del sistema"

- **Inflexibilidad de Unidades**. Una unidad es inflexible cuando las características técnicas de la unidad hacen que genere en una hora a pesar de que su precio de oferta es superior al costo marginal del sistema, o cuando se modifica la disponibilidad declarada después de la hora de cierre de las ofertas y antes del período de reporte de cambios para el redespacho.

ANEXO B: Filtrado de variables

El estudio del filtrado de variables [29] o selección del subconjunto de variables para la inducción de un modelo clasificador, se conoce como FSS (Feature Subset Selection).El objetivo al tratar de resolver el problema FSS es el de detectar aquellas variables que son irrelevantes y/o redundantes para un problema clasificatorio dado. "Se considera que una variable predictiva es irrelevante cuando el conocimiento del valor de la misma no aporta información alguna que despeje incertidumbre sobre la variable clase. Una variable predictiva se dice redundante cuando su valor puede ser determinado a partir de otras variables predictivas". [28].

La selección de variables también es importante debido a la no monotocidad de los modelos clasificatorios en relación con el número de variables predictoras. La no monotocidad de la probabilidad de éxito de un sistema clasificador se debe al hecho de que no por construir un modelo clasificatorio con una variable añadida a las ya existentes, la probabilidad de éxito que se va a obtener con este nuevo modelo clasificador deba superar a la del modelo actual.

Dentro del FSS, existe un grupo de métodos conocidos como indirectos o filter. Estos hacen uso de heurísticos para determinar el subconjunto de atributos óptimo. Un heurístico es una regla matemática que es capaz de guiar un proceso de búsqueda hacia una solución.

Entre las principales características se encuentra la rapidez de cálculo, hecho que hace que en conjuntos de datos con alta dimensionalidad las aproximaciones filter sean consideradas como óptimas y utilizadas por diferentes autores. Dentro de la selección de variables tienen gran importancia los métodos que asignan coeficientes de relevancia a cada atributo, estableciendo un ranking entre ellos, tal como se realizo en este trabajo.

- **Ranking de Variables**:

Ranking realizado de forma univariada, es decir, solo se tiene en cuenta la relación existente entre el atributo que está siendo analizado y la variable supervisada o variable clase. Mediante la aplicación de distintos heurísticos derivados de diferentes medidas de divergencia entre funciones de distribución, se va generando un coeficiente o peso para cada uno de los atributos.Una vez generados todos los coeficientes, los atributos son ordenados en función de éstos, obteniendo así el ranking que se buscaba. [29].

- **Información Mutua**:

Este método fue el empleado en la investigación para llevar a cabo el ranking de las variables y lo que permite es, calcular la relación que existe entre las variables y otra dada; en el caso de este trabajo, la relación se calculó respecto al precio de oferta fijado por cada central.

Esta medida se basa en la cantidad de incertidumbre que el conocimiento de una variable es capaz de despejar con respecto al estado en el que se encuentre la segunda, conocida comúnmente como *variable clase*. La medida

puede tomar valores en el intervalo [0,1]; valores cercanos a 1 indican alta correlación entre las variables analizadas, mientras que, valores cercanos a 0 indican independencia entre ellas.

La expresión matemática mostrada a continuación muestra como se lleva a cabo el cálculo de la Información Mutua. La variable a evaluar es *'X'* y el número de estados de estados que puede tomar es r_x; La variable clase se representa por *'C'* y es el número de estados que puede tomar es *'*r_c*'*.

$$I_p(X;C) = \sum_{i=1}^{rx}\sum_{j=1}^{rc} P(x_i, c_j)\log\frac{P(x_i, c_j)}{P(x_i)P(c_j)}.$$

Donde:

$P(x_i)$ 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

y,

$P(x_i,c_j)$ 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:

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