ESCUELA POLITÉCNICA NACIONAL

ESCUELA DE INGENIERÍA

PREDICCIÓN DE LA DEMANDA ELÉCTRICA, UTILIZANDO UN MODELO DE APRENDIZAJE

PROYECTO PREVIO A LA OBTENCIÓN DEL TITULO DE INGENIERO EN SISTEMAS INFORMÁTICOS Y DE COMPUTACIÓN

ESTEBAN ISRAEL ALBÁN SILVA
DAVID ALBERTO ALEMÁN QUIMBIULCO

DIRECTOR: MSC. ING. CARLOS MONTENEGRO ARMAS

Quito, Julio 2006

DECLARACION

Nosotros, Esteban Israel Albán Silva y David Alberto Alemán Quimbiulco, declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido previamente presentada para ningún grado o calificación profesional; y, que hemos consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cedemos nuestros derechos de propiedad intelectual correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

Esteban Israel Albán Silva

David Alberto Alemán Quimbiulco

CERTIFICACION

Certifico que el presente trabajo fue desarrollado por Esteban Israel Albán Silva y David Alberto Alemán Quimbiulco, bajo mi supervisión.

Msc. Ing. Carlos Montenegro A.
DIRECTOR DE PROYECTO

AGRADECIMIENTO

Agradezco a mis Padres y hermanos por el apoyo incondicional que me han brindado, durante todas las etapas de mi vida.

Doy las gracias a mi compañero y Amigo, David, que me ha brindado lo más valioso en este proyecto, su amistad.

Agradezco, también a nuestro Tutor de Proyecto, quien con su meritoria ayuda, nos ha permitido terminar este trabajo.

Israel

AGRADECIMIENTO

Agradezco de manera especial a mi familia que siempre ha estado conmigo brindándome su apoyo.

Agradezco a mi amigo y compañero de proyecto Israel por su colaboración y su amistad brindada.

Agradezco a nuestro Director de Proyecto, por su ayuda para lograr la culminación de este trabajo.

David

DEDICATORIA

Dedico con mucho cariño, el trabajo realizado
A mis Queridos Padres y Hermanos,
A todos mis Amigos.

Israel

DEDICATORIA

Éste trabajo está dedicado a mis padres, a mis hermanos y amigos.

David

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

Índice de Figuras

INTRODUCCIÓN

En la actualidad la mayoría de las actividades que realiza el ser humano va acompañado de la tecnología que en los últimos años ha tenido un gran desarrollo, haciéndose indispensable su constante actualización y uso.

El hecho de que haya una mayor relación entre tecnología y ser humano, a dado lugar a que se busquen formas en que la máquina simule las actividades que realiza un hombre, esto a través de modelos de aprendizaje de máquina, los mismos en algunos casos han llegado a reemplazar el desempeño del ser humano.

Los modelos de aprendizaje de máquina cubren algunas áreas de aplicación y una de ellas es la predicción, en nuestro país no es de desconocimiento el problema de provisión de energía eléctrica que tenemos, y el cual se puede llegar tener previsión, en función de los diversos factores que influyen en la demanda eléctrica con el fin de evitar los efectos que tiene en la producción, industria y en si en la economía del país.

RESUMEN

Una de las ventajas de poder aplicar modelos de aprendizaje de máquina para la predicción, es que se pude tener alternativas de aplicación con el fin de tener a la mano varias opciones para tener un mejor análisis de la información.

El objetivo del presente trabajo es predecir la demanda de energía eléctrica en función de un modelo de aprendizaje alternativo, para lo cual se describe brevemente el contenido del mismo.

En el primer capítulo se realiza una descripción teórica de lo que involucra el aprendizaje de máquina, mencionando, algunos de los modelos de aprendizaje.

En el segundo capitulo se describe la situación actual del Sistema Nacional Interconectado, la forma de operación del mismo las empresas que conforman el mismo, además se conocen los modelos que realizan actualmente la predicción de la demanda.

En el tercer capítulo se realiza la selección del modelo de aprendizaje a utilizarse además del desarrollo de la metodología con la se realizó la aplicación, para terminar con un análisis de resultados obtenidos una vez aplicado el modelo de aprendizaje.

Para terminar en el cuarto capitulo, se presentan las conclusiones y recomendaciones obtenidas de la experiencia de haber realizado éste trabajo.

CAPITULO 1. MODELOS DE APRENDIZAJE PARA PREDICCIÓN

1.1 TEORÍA DEL APRENDIZAJE COMPUTACIONAL

1.1.1 ANTECEDENTES

Desde un principio, el objetivo de todo programa computacional es imitar a la inteligencia humana, pero acelerando los procesos que estos programas ejecuten, puesto que de esta manera los resultados serían más precisos, rápidos y confiables.

Debido a la complejidad de la Inteligencia Humana, la investigación orientada a entender el funcionamiento de la misma está avanzando, pero debido a la gran magnitud que esta involucra, es muy difícil conocer a ciencia cierta cuando se pueden alcanzar el objetivo planteado; es por eso que los programas computacionales en los que interviene el Aprendizaje todavía no llegan a la capacidad de imitar en un 100% las habilidades cerebrales del humano.

Sin embargo, de los resultados de las investigaciones realizadas se han llegado a crear programas que tienen similitudes operacionales al pensamiento humano, en las cuales los programas computacionales pueden llegar a trabajar de una manera parecida al humano; en algunos casos puede llegar a realizar la misma actividad con mejores resultados.

A medida que la tecnología avanza, es necesario que el hombre aproveche de ella, con el fin de obtener un mayor desarrollo de la productividad, de las tareas en las que esté involucrado para su propio beneficio y el de la sociedad a la que pertenece.

Cuando se presentan necesidades del ser humano, estas pueden ser correspondidas y solucionadas con los modelos de aprendizaje computacional conocidos o a la vez pueden crearse nuevos, que ayuden a cubrir esas limitaciones planteadas. De esta manera, existe una constante evolución del Aprendizaje de Máquina.

1.1.2 EVOLUCIÓN DEL APRENDIZAJE COMPUTACIONAL

La investigación en aprendizaje automático empezó a mediados del siglo XX y desde entonces ha evolucionado con distinto grado de intensidad, utilizando diferentes técnicas y haciendo énfasis en distintos aspectos y objetivos. Podemos distinguir tres periodos importantes, en esta disciplina de la Inteligencia Artificial, cada uno de los cuales se menciona a continuación:

- Técnicas de modelado neuronal y de decisión
- Aprendizaje orientado a conceptos simbólicos
- Sistemas de aprendizaje de conocimiento con exploración de varias tareas de aprendizaje

La principal característica del primer período fue el interés de construir sistemas de aprendizaje de propósito general que partan con poco o ningún conocimiento inicial de la estructura. La investigación estuvo orientada a la construcción de una gran variedad de máquinas basadas en modelos neuronales, con una estructura inicial aleatoria o parcialmente aleatoria. Estos sistemas fueron denominados Redes Neuronales o Sistemas Auto-organizativos. El aprendizaje en estos sistemas consiste en la realización de cambios incrementales en las probabilidades de que elementos del tipo neurona (típicamente unidades lógicas con umbral) puedan transmitir una señal.

Debido a la primitiva tecnología computacional de los primeros años, la mayoría de las investigaciones en esta área eran teóricas o relativas a la específica construcción de sistemas de hardware con propósito específico, tal como perceptrones, pandemonium y adelaine. El fundamento de estos trabajos fue hecho en la década de los cuarenta donde se descubrió la aplicabilidad de la lógica simbólica para el modelado de actividades del sistema nervioso. Otro tipo de investigación relacionada con el área es la concerniente a la simulación de procesos evolutivos, que a través de operaciones aleatorias de mutación y de “selección” natural pueden crear un sistema capaz de realizar un comportamiento inteligente.

De las prácticas realizadas en éstas áreas generó la nueva disciplina de Reconocimiento de Patrones y se dio inicio al desarrollo de sistemas de decisión en aprendizaje automático. En dichos casos, el aprendizaje es igualado con la adquisición de funciones lineales, polinomiales, o formas relacionadas con funciones discriminantes a partir de un conjunto de ejemplos de entrenamiento.

Uno de los sistemas más exitosos y conocidos dentro de esta clase fue el programa de juego de damas de Samuel. Este programa estaba capacitado para adquirir por medio de aprendizaje, un mejor nivel de desempeño. Algo diferente, pero relacionado, son las técnicas que utilizan métodos de decisión estadística para el aprendizaje de reglas de reconocimiento de patrones.

Como investigación paralela al modelado a través de redes neuronales y sistemas de decisión, se realizaron investigaciones relacionadas con teoría de control, sistemas de control adaptativos capaces de ajustar automáticamente sus parámetros con el objetivo de mantener un desempeño estable en presencia de perturbaciones.

Posteriormente, los resultados prácticos realizados por modelos basados en redes neuronales y sistemas de decisión encontraron ciertas limitaciones. Las altas expectativas mantenidas por los trabajos originales no se cumplieron, y la investigación en estas áreas comenzó a declinar. Algunos resultados teóricos revelaron fuertes limitaciones, como por ejemplo en el aprendizaje del tipo perceptrón simple.

El segundo periodo comenzó a surgir en los sesenta, a partir de los trabajos de psicólogos e investigadores en inteligencia artificial, sobre el modelado del aprendizaje humano. Estas investigaciones se basaban en estructuras lógicas o de grafos en vez de métodos numéricos o estadísticos. Los sistemas aprendían descripciones simbólicas que representaban un mayor nivel de conocimiento de las estructuras y conceptos adquiridos.

Una influencia importante en esta área de trabajo fue el sistema de aprendizaje estructural de Winston, como así también los trabajos realizados con el

objetivo de aprender conceptos estructurales a partir de ejemplos, incluyendo los programas de aprendizaje basados en lógica inductiva.

El tercer periodo representa la fase de investigación más reciente, comenzando a partir de mediado de los setenta. Las investigaciones en este sentido han sido orientadas al aprendizaje de conceptos a partir de ejemplos, utilizando una amplia variedad de técnicas, muchas de las cuales se orientan a los sistemas basados en conocimiento. Se ha hecho especial énfasis en el uso de conocimiento orientado a tareas y en las restricciones que este provee, que guían el proceso de aprendizaje. Además del énfasis en el aprendizaje a partir de ejemplos, se trabaja en aprendizaje a partir de instrucciones, por analogía y descubrimiento de conceptos y clasificaciones. En contraste con esfuerzos previos, las nuevas investigaciones tienden a incorporar heurísticas, y utilizar ejemplos de entrenamiento para el aprendizaje de conceptos.

1.1.3 CONCEPTOS DE APRENDIZAJE COMPUTACIONAL

Existen variedad de conceptos sobre Aprendizaje Computacional, de los que se resaltan los siguientes:

Una de los conceptos de Aprendizaje Computacional dice: “Un programa de computadora se dice que aprende de experiencia E con respecto a una clase de tareas T y medida de desempeño D, si su desempeño en las tareas en T, medidas con D, mejoran con experiencia E” [1]; por ejemplo, un programa de computación que aprende a jugar al ajedrez debería mejorar su desempeño, medido por su habilidad de ganar en la clase de tareas correspondientes a jugar partidas de ajedrez, a través de la experiencia obtenida jugando partidas.

Otro concepto sobre el Aprendizaje Computacional menciona: “Aprendizaje Computacional significa cambios adaptativos en el sistema: permite que el sistema ejecute la misma tarea con mayor eficacia para la siguiente ocasión” [2] ; en base a lo expuesto en los conceptos anteriores podemos manifestar que el Aprendizaje Computacional es una rama de la Inteligencia Artificial que tiene como objetivo desarrollar técnicas que permitan a las computadoras aprender.

---

[1] [Mitchell, 97].
[2] [Herbert Simun]

Es decir que se trata de elaborar programas capaces de generalizar comportamientos a partir de una información no estructurada suministrada en forma de ejemplos.

En muchas ocasiones el campo de acción del Aprendizaje Computacional se solapa con el de la Estadística, ya que las dos disciplinas se basan en el análisis de datos, para su posterior uso dentro de la aplicación para dar solución a un problema determinado, y dentro de ellos contempla el problema de la predicción.

Hay que tomar en cuenta que los programas de Aprendizaje de Máquina utilizan la experiencia como medio para mejorar el desempeño de la aplicación. Es por esta razón que la Experiencia ayuda a obtener soluciones de una manera más eficaz. Cuando se utiliza un programa que aplique el Aprendizaje Computacional, se debe considerar que el algoritmo de aprendizaje, pueda realizar una predicción que satisfaga las expectativas que motivó a ser utilizado. Para comprender aún mejor el concepto de aprendizaje se deben establecer tres actores importantes que permiten establecer el objeto y el escenario en el que se va ha desarrollar dicha acción, a continuación los mencionamos: agente, entorno y función de costo.

Agente y Entorno

Se denomina a un sistema o agente A, que interactúa con un entorno E, es decir el agente es el objeto que se convierte en móvil de aprendizaje y el entorno es el contexto o ambiente en el que va ha desenvolver y a tomar las características para aprender el agente. Por ejemplo: Un programa de ordenador que clasifica el correo electrónico como “spam” o “no spam”

Función de Costo

Si tenemos una “función de costo” f(A, E) que asigna un número, ésta proporciona un grado de optimalidad de la interacción del agente con su entorno, ésta permite ver hasta punto existe un vínculo entre los otros dos elementos. Ejemplo: Número de fallos que comete el ordenador clasificando los correos.

Una vez que se conocen a los principales elementos del aprendizaje, podemos decir que existe aprendizaje si, el agente A “aprende”, si es capaz de regular su comportamiento de tal forma que la función de costo f(A, E) es más óptima con el paso del tiempo. Ejemplo: El programa clasifica cada vez mejor el correo electrónico. Esto se establece desde un punto de vista fuera del sistema en el que no se define cómo hace eso el agente, ni en cómo obtiene la información adicional, ni en quién la proporciona a dicho sistema.

En resumen se puede concluir que siempre hay una función de costo que se debe optimizar. El aprendizaje consiste en la búsqueda de la hipótesis óptima que maximiza esta función.

1.1.4 APLICACIONES, UTILIDADES DEL APRENDIZAJE COMPUTACIONAL

En la actualidad existen algoritmos y teorías que implementan el Aprendizaje Computacional, los cuales han tenido mucho éxito en muchos campos como por ejemplo: detección de transacciones falsas, reconocimiento de voz, motores de búsqueda en el Internet, conducción de autos autónomos, etc…

Aplicación En Los Sistemas Productivos

El Aprendizaje Computacional en un sistema productivo integrado por computadora, es usado para la supervisión, planificación, secuenciación, cooperación y ejecución de las tareas de operación en centros de trabajo, agregado al control de los niveles de inventario y características de calidad y confiabilidad del sistema. Los factores mencionados determinan la estructura del sistema y su coordinación representa una de las funciones más importantes en el manejo y control de la producción.

En otras de las ramas en donde el AC es de mucha utilidad por tener mayor incidencia en los procesos productivos de la industria a nivel mundial, es el diseño de sistemas para la toma de decisiones basados en la optimización de parámetros, puesto que al optimizarlos se mejora la funcionalidad de los procesos que requieren solución.

Un programa computacional que aplique aprendizaje de máquina ayuda a mejorar el control de calidad de un proceso de producción ejecutando de una manera totalmente automática las labores de identificación de objetos y de control de calidad de los mismos, ya que ciertos factores imperceptibles a la vista del operador encargado de dicho control, pueden ser detectados y advertidos por el sistema colaborando de esta manera a reducir los errores.

Aplicación En Biología Molecular

El AC aplicado a la Biología Molecular se basa en construir modelos computacionales capaces de predecir la estructura y función de las secuencias biológicas, además de acceder a las principales bases de datos de las mismas. Asimismo nos permite predecir rasgos fenotípicos como por ejemplo el color del pelo, la inteligencia o la susceptibilidad al cáncer. Este enfoque se basa en la estructura jerárquica de los genes, su ordenamiento dentro del genoma, la función de las proteínas para las que codifican y las interacciones entre estas proteínas, que da lugar al mantenimiento del metabolismo, la reproducción y la forma.

Aplicación En Economía, Marketing Y Finanzas

El AC aplicado a herramientas inteligentes en el área de Finanzas puede utilizarse para escrutar entre los millones de datos que se generan en un banco en busca de patrones de comportamiento de sus clientes o para detectar tendencias en los mercados de valores.

Como otro caso de aplicación podríamos citar la concesión de créditos de una entidad financiera, en la que se analiza la información contable de las empresas que solicitan el crédito financiero, valorando si la empresa es solvente o presenta problemas y, a partir de ahí, decidir si merece un crédito o no.

Aplicación En Ingeniería, Control Y Robótica

Esta aplicación se basa en la identificación global de ambientes ejecutada por un robot móvil en base al entrenamiento de una red neuronal que recibe la información captada del medio ambiente por el sistema sensorial del robot

(ultrasonido). Se considera que el robot, a través de la red neuronal, tiene como única tarea maximizar el conocimiento del ambiente que se le presenta. De esta forma este modela y explora el ambiente eficientemente mientras ejecuta algoritmos de evasión de obstáculos.

El resultado de este estudio es de gran importancia en el campo de la robótica móvil debido a que: el robot adquiere una mayor autonomía del movimiento, se optimiza el uso del ultrasonido como detector de obstáculos y es una herramienta importante para el desarrollo de planificadores de trayectoria y controladores “inteligentes”.

Aplicación En Medicina Y Salud

En las ciencias relacionadas con la salud humana el AC es útil para el proceso de diagnóstico que es siempre complejo, ya que exige la valoración de múltiples factores que interactúan en el caso que se examina. Como en todas las áreas que atañen a la salud, la tarea diagnóstica es fundamental para arribar a una conducta terapéutica acertada. El diagnóstico es una actividad personal, individual y cognitiva donde el profesional eficiente articula sus conocimientos científicos y su experiencia clínica con sentido común. Para el aprendizaje del diagnóstico médico, además del conocimiento teórico adquirido, es necesaria la experiencia en casos clínicos y la práctica de consultorio puede verse reforzada mediante herramientas automatizadas basadas en alguna representación del conocimiento.

El AC tiene una larga tradición en el desarrollo de sistemas informáticos aplicados al área de la salud, dentro de los cuales, los sistemas de apoyo a la toma de decisión han ocupado un lugar importante. Estos sistemas son útiles como herramientas educativas y de entrenamiento, y hay varios desarrollos de software orientados a la formación continua de profesionales del área. Estos sistemas han utilizado distintas técnicas y siguen en continuo desarrollo.

1.2 TIPOS DE APRENDIZAJE

Dentro de la clasificación de los tipos de aprendizaje de máquina hemos visto conveniente agruparlos en dos tipos, los cuales mencionamos a continuación:

a. En función de la experiencia utilizada durante el entrenamiento

b. En función del método utilizado para aprender

1.2.1 EN FUNCIÓN DE LA EXPERIENCIA UTILIZADA DURANTE EL ENTRENAMIENTO

Debido a que la experiencia es un elemento fundamental cuando de aprendizaje computacional se trata, se ha visto la necesidad de tomar en cuenta éste factor y clasificarlo de la siguiente manera:

1.2.1.1 Aprendizaje Supervisado

El aprendizaje supervisado se basa en el intento por parte del programa computacional de pronosticar resultados con ejemplos conocidos, y es un método de uso habitual. El programa compara sus predicciones con la respuesta objetivo y aprende de sus errores. Utiliza un algoritmo de aprendizaje el cual tiene información del resultado asociado a cada ejemplo, es decir conoce lo que es correcto y lo que no lo es.

Si hacemos una analogía entre el Aprendizaje Supervisado Computacional y el método de enseñanza tradicional se puede decir que el profesor indica y corrige los errores del alumno hasta que éste aprende la lección.

Éste tipo de aprendizaje es típico en problemas de predicción, clasificación y modelos de series temporales, por ejemplo si se tiene el caso de diagnóstico de enfermedades, el programa que integre aprendizaje supervisado parte de una lista de síntomas para obtener un diagnostico correcto de dicha enfermedad.

1.2.1.2 Aprendizaje No Supervisado

El aprendizaje no supervisado es más efectivo para describir datos que para pronosticar resultados. El aprendizaje No Supervisado crea efectivamente sus propios métodos de interpretación y validación. No requieren suposiciones iniciales sobre qué constituye un grupo o cuántos grupos hay, ya que trabajan exclusivamente a partir de los patrones de los datos. Además éste tipo de

aprendizaje parte de cero, por lo que no se nota una influencia sobre qué factores son más importantes.

Si realizamos la analogía con el método de enseñanza tradicional para el caso del Aprendizaje No Supervisado, no hay un profesor que corrija los errores al alumno; recurre más al autoaprendizaje. El alumno dispone del material de estudio pero nadie lo controla. Si el entrenamiento es no supervisado, únicamente debemos suministrar los datos de entrada para que extraiga los rasgos característicos esenciales.

Éste tipo de aprendizaje utiliza un algoritmo que no tiene acceso a información sobre el resultado, es decir, solo se conocen los ejemplos, no sus resultados, Por ejemplo se aplica en problemas de clustering (clasificación sin clases predefinidas), o sea para agrupar los casos similares.

1.2.2 EN FUNCIÓN DEL MÉTODO UTILIZADO PARA APRENDER

Otro factor preponderante para clasificar el aprendizaje es el método utilizado para aprender y en base a esto, se realiza la siguiente clasificación:

1.2.2.1 Aprendizaje Inductivo

El aprendizaje inductivo es la capacidad de obtener nuevos conceptos, más generales, a partir de ejemplos. Este tipo de aprendizaje conlleva un proceso de generalización/especialización sobre el conjunto de ejemplos de entrada. Los algoritmos implementados son, además incrementales, es decir, el procesamiento de los ejemplos se realiza uno a uno. Esta característica, a diferencia del procesamiento por lotes (batch), permite visualizar el efecto causado por cada uno de los ejemplos de entrada, en el proceso de obtención del concepto final. Además de la generalización de conceptos, el programa permite clasificar conjuntos de ejemplos a partir de los conceptos obtenidos anteriormente. De este modo, se puede comprobar, para cada ejemplo de un conjunto dado, a qué clase pertenece dicho ejemplo.

El aprendizaje inductivo puede verse como el proceso de aprender una función, es decir, al elemento de aprendizaje se le da un valor correcto (o aproximadamente correcto) de una función a aprender para entradas particulares y cambia la representación de la función que está infiriendo, para tratar de aparear la información dada por la retroalimentación que ofrecen los ejemplos.

1.2.2.2 Aprendizaje Deductivo

Este tipo de aprendizaje se basa en adquirir una serie de generalidades sobre algo de lo que se conoce solo una parte. Trata de disminuir el tiempo requerido para utilizar el conocimiento disponible, en realidad no se está aprendiendo nada sino que se está optimizando la utilización de lo aprendido. Es decir, que cuando realiza el aprendizaje deductivo, se busca ampliar el conocimiento que se tiene de algo de lo cuál se sabe muy poco.

La descripción del aprendizaje deductivo, es que dada una definición inicial del concepto que queremos aprender y dado un ejemplo que creemos que pertenece a ese concepto, intentamos comprobar que ese ejemplo pertenece realmente al concepto a aprender, y para ello le aplicamos al ejemplo las reglas que inicialmente conocemos del concepto a aprender. De dicha aplicación obtenemos reglas más generales que el ejemplo pero más específicas que el concepto.

En realidad, de lo que se trata es de dada una definición funcional (muy general y vaga) de un concepto, y estudiando algunos ejemplos de ese concepto, obtener una definición más estructural (o más objetiva) de dicho concepto.

1.2.2.3 Aprendizaje Basado En Ejemplos (Instancias)

A diferencia de aquellos métodos de aprendizaje que construyen una descripción general, y explícita de la función objetivo a partir de los datos de entrenamiento, estos métodos simplemente guardan dichos datos. La generalización sobre estos ejemplos se pospone hasta que una nueva instancia debe ser clasificada. Cada vez que una nueva instancia es encontrada, se calcula su relación con los ejemplos previamente guardados

con el propósito de asignar un valor de la función objetivo para la nueva instancia. El aprendizaje basado en instancias incluye el vecino mas cercano y métodos de regresión pesados localmente que asumen que las instancias pueden ser representadas como puntos en el espacio euclideo. Incluyen también a los métodos de razonamiento basado en casos, que utilizan una representación más compleja y simbólica de los datos. Los métodos de aprendizaje basados en instancias son denominados "perezosos" pues dilatan el procesamiento hasta que una nueva instancia deba ser clasificada. Una ventaja de este retraso es que no se estima la función objetivo una vez para todo el espacio de instancias, sino que se hace en forma local y diferente para cada nueva instancia a clasificar.

1.2.2.4 Aprendizaje Por Refuerzo

Este tipo de aprendizaje se refiere a como un agente autónomo que actúa en un entorno, puede aprender a elegir acciones optimas que lo conduzcan a alcanzar objetivos. Este problema genérico cubre tareas tales como el aprendizaje del control de un robot móvil, aprendizaje de cómo optimizar operaciones en una factoría, y aprendizaje de cómo realizar jugadas en juegos de tableros. Cada vez que un agente realiza una acción en su entorno, un entrenador provee un premio o penalización que indica la bondad del estado resultante. Por ejemplo, cuando se entrena a un agente para jugar un juego, el entrenador debe proveer una recompensa positiva cuando el juego es ganado, negativa si se pierde y cero en los otros estados. La tarea del agente es la de aprender a partir de esta recompensa indirecta y retrasada, a elegir secuencias de acciones que produzcan la mayor acumulación de recompensas. Un ejemplo de estos tipos de algoritmos es el Q-learning, que permite adquirir estrategias de control óptimas a partir de recompensas retrasadas, aun cuando el agente no posee un conocimiento inicial del efecto de las acciones en el entorno. Estos algoritmos están relacionados con la programación dinámica, frecuentemente empleada en la resolución de problemas de optimización.

1.2.2.5 Otros Tipos

1.2.2.5.1 Aprendizaje Bayesiano

El aprendizaje Bayesiano conocido también como aprendizaje probabilístico provee un acercamiento a la inferencia. Esta basado en asumir que las cantidades de interés están gobernadas por distribuciones de probabilidad y que las decisiones optimas pueden ser realizadas por medio de razonamientos sobre estas probabilidades y datos observables. Provee una visión cuantitativa para pesar la evidencia que soporta distintas hipótesis. El aprendizaje Bayesiano provee las bases para el aprendizaje de algoritmos que manipulan directamente probabilidades, y un ámbito para analizar como operan otros algoritmos que no las manipulan explícitamente.

Actualmente existe algunas aplicaciones importantes en las que interviene el aprendizaje bayesiano de las cuales podemos destacar una, por ser muy conocida por la mayoría de personas que utiliza Internet: "El motor de búsqueda Google, en la cual un pequeño ejército de matemáticos bayesianos intentan encontrar patrones derivados de las conexiones y dependencias que emergen de los billones de links que el motor maneja, y de la forma en qué buscan y seleccionan resultados millones de usuarios" [3] .

1.2.2.5.2 Aprendizaje evolutivo

El aprendizaje de estos algoritmos esta basado en la simulación de la evolución, ya que los algoritmos evolutivos AE están inspirados en el concepto Darwiniano de evolución. Las hipótesis que se aprenden originalmente fueron representadas como cadenas de bits, cuya interpretación depende del tipo de aplicación. Dichas hipótesis pueden llegar a ser descriptas por expresiones simbólicas o aun por programas de computación. La búsqueda de una hipótesis apropiada comienza con una población, o colección, de hipótesis iniciales. Los miembros de la actual población pasan a la próxima generación, por medio de operaciones tales como mutación randómica o cruce, asociadas a procesos de evolución biológica. En cada paso, la hipótesis es evaluada en base a una medida de aptitud, y las mejores son seleccionadas en forma probabilística para pasar a la próxima generación. Estos algoritmos han sido aplicados en forma exitosa a una variada gama de tareas de aprendizaje y a otros problemas de optimización. Por ejemplo, han sido utilizados para

[3] http://www.infonomia.com/directorio/bayesiano.asp

aprender una colección de reglas de control de un robot y para optimizar la topología y los parámetros de una red neuronal. Uno de los representantes de esta clase de algoritmos son los algoritmos genéticos.

Por otra parte, los Algoritmos Evolutivos (AE) conforman una de las más importantes familias de modelos computacionales con aplicación en el campo del aprendizaje automático.

Aunque existen varios aspectos que tienen una influencia notable en la eficiencia y la eficacia del algoritmo, las dos tareas principales en la aplicación de un AE son el diseño de la codificación y la evaluación de los individuos.

1.2.2.5.3 Aprendizaje basado en árboles de decisión

Los árboles de decisión representan el conocimiento obtenido en el proceso de aprendizaje inductivo. A partir de un conjunto numeroso de prototipos puede verse la estructura resultante de la partición recursiva del espacio de representación. Esta partición recursiva se traduce en una organización jerárquica del espacio de representación que puede modelarse mediante una estructura de tipo árbol. Cada nodo interior contiene una pregunta sobre un atributo concreto (con un hijo por cada posible respuesta) y cada nodo hoja se refiere a una decisión (clasificación).

La clasificación de patrones se realiza en base a una serie de preguntas sobre los valores de sus atributos, empezado por el nodo raíz y siguiendo el camino determinado por las respuestas a las preguntas de los nodos internos, hasta llegar a un nodo hoja. La etiqueta asignada a esta hoja es la que se asignará al patrón a clasificar.

Es uno de los métodos más ampliamente utilizados para inferencias inductivas. Es un método para la aproximación de funciones de valores discretos, robustos frente a datos con ruido y capaces de aprender expresiones disjuntas. Existe una familia de algoritmos de aprendizaje de árboles de decisión que incluye los ampliamente utilizados: ID3, C4.5, See5 y ASSISTANT. Estos métodos buscan un espacio de hipótesis completamente expresivo y de esta manera evitan las

dificultades que surgen de espacios de hipótesis restringidos. Su sesgo inductivo es la preferencia de árboles pequeños por sobre los grandes.

El aprendizaje basado en árboles de decisión, también llamado de clasificación o de identificación, es uno de los métodos de aprendizaje más utilizado, puesto que su ámbito de aplicación no está restringido a un campo específico como por ejemplo al diagnóstico médico, juegos, predicción meteorológica, control de calidad, etc… ya que éste tipo de aprendizaje no está atado a alguna área determinada, sino que pueden aplicarse en diversos ámbitos. Los árboles de clasificación destacan por su sencillez para la representación del conocimiento.

1.2.2.5.4 Aprendizaje basado en redes neuronales artificiales

En principio éste tipo de aprendizaje fue una simulación de los sistemas nerviosos biológicos puesto que partieron de unidades básicas como es el de las neuronas conectadas unas con otras y al aplicar el concepto de dendritas y axones como en los sistemas nerviosos humanos.

Las Redes Neuronales Artificiales fueron originalmente una simulación abstracta de los sistemas nerviosos biológicos, formados por un conjunto de unidades llamadas "neuronas" o "nodos" conectadas unas con otras. Estas conexiones tienen una gran semejanza con las dendritas y los axones en los sistemas nerviosos biológicos.

El aprendizaje computacional basado en redes neuronales artificiales representar funciones usando redes de elementos con cálculo aritmético sencillo, y métodos para aprender esa representación a partir de ejemplos. La representación es útil para funciones complejas con salidas continuas y datos con ruido.

Este aprendizaje es robusto frente a la aparición de errores en los datos de entrenamiento y han sido aplicados con éxito a problemas tales como la interpretación de escenas visuales, reconocimiento del habla y estrategias de control de robots.

CAPITULO 2. PREDICCION DE LA DEMANDA ELECTRICA EN EL CENACE - ECUADOR

En este Capítulo se describe de una manera general la Situación Actual de la Energía Eléctrica en el Ecuador, así como los principales problemas que se tiene cuando no se distribuye de forma eficiente la Energía Eléctrica. También en este capítulo se refiere al papel en que está involucrado el CENACE en el país, además de recalcar las principales funciones de las que se encarga esta Organización. Luego además se realiza un análisis sobre los modelos que utiliza el CENACE para predecir la Demanda Eléctrica en la actualidad. Describiéndolos de una manera general para conocer principalmente el funcionamiento de los mismos.

2.1 SITUACION ACTUAL

2.1.1 SITUACION DEL SECTOR ELECTRICO NACIONAL

La actual situación del sector eléctrico ecuatoriano lo conforman la Generación, Transmisión y Distribución de la Energía Eléctrica; a todo este conjunto se lo denomina Sistema Nacional Interconectado SNI. Cada parte se encarga de tareas específicas que cubren la Demanda general del País. Se tiene que tomar en cuenta que la planeación de la operación de los recursos de generación y transmisión debe hacerse en forma integrada, buscando minimizar los costos de la operación del sistema y atender con seguridad, confiabilidad y calidad la demanda.

Generación

Los métodos para la Generación de Electricidad, se muestra a continuación:

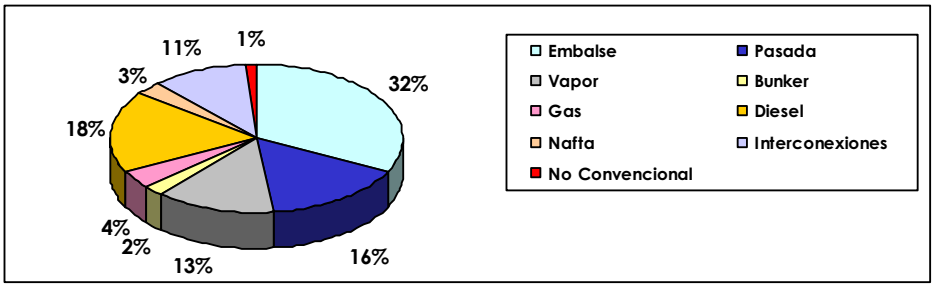


Figura 1 Métodos de Generación de Electricidad

Fuente: CENACE

De esto podemos darnos cuenta que el principal método de generación de Energía Eléctrica es la Hidroeléctrica (Embalse), razón por la cual se tiene mucha dependencia con el factor climatológico de nuestro país.

Entre las principales generadoras de electricidad en el País podemos citar las siguientes:

HIDROPUCARA	HIDROPAUTE
HIDRONACION	HIDROAGOYAN
M. POWER	ELECAUSTRO
CATEG-G	ELECTROGUAYAS S.A.
ELECTROQUIL	TERMOESMERALDAS S.A.
TERMOPICHINCHA	INTERVISA
EEQSA	EMPRESAS NO CONVENCIONALES

Tabla 1 Principales Generadoras del Ecuador

Fuente: CENACE

El Sistema Nacional Interconectado se encuentra operando al límite de los criterios de economía, seguridad y calidad, debido fundamentalmente a la alta indisponibilidad por mantenimiento correctivo del parque generador del país y a la falta de algunas obras a nivel de transmisión y Distribución. Se requiere lo siguiente:

- El cumplimiento del Plan de Expansión de Transmisión.
- El cumplimiento de los factores de potencia por parte de los Distribuidores y Grandes Consumidores.
- La construcción y operación urgente del patio de 69 kV de la subestación Salitral.
- Realizar las acciones correctivas por parte de las empresas cuyas unidades de generación no presentan la confiabilidad y disponibilidad adecuadas. En el mes de abril de 2005 se encontró alrededor de 500 MW de capacidad de generación indisponible por causas no programadas, situación que puso en riesgo la cobertura de las demandas de potencia y energía del sistema.

- La instalación de compensadores sincrónicos o compensadores estáticos de potencia reactiva en las áreas de Guayaquil y Quito.

- Ejecución de la implementación de los centros de control en las Empresas de Distribución, particularmente en las áreas de Quito y Guayaquil donde se encuentran ubicados los mayores centros de consumo del país.

- Ejecución de la Auditoria Integral de las protecciones del Sistema Nacional Interconectado.

Los Directores monitorean y miden las características del producto y verifican diariamente que los requisitos de éste han sido cumplidos.

El personal que ejecuta los procesos los controla mediante rutinas de validación y control incorporadas a los diferentes sistemas informáticos y en otras ocasiones mediante rutinas manuales que permite la corrección inmediata del servicio para garantizar la conformidad con los requisitos especificados.

Adicionalmente, las salidas de los procesos son entradas para otros, como se puede verificar en el Diagrama de Calidad. Estos procesos, que son clientes internos, a su vez verifican la conformidad de los productos que son entradas a su proceso mediante rutinas.

TRANSELECTRIC S.A.

"La Compañía Nacional de Transmisión Eléctrica, TRANSELECTRIC S.A., es una empresa privada que tiene como único accionista al Fondo de Solidaridad, comenzó a operar en abril de 1999, y su función es la transmisión de energía eléctrica para todo el país."[4]

Transelectric S.A. opera el Sistema Nacional de Transmisión Eléctrica, su objetivo fundamental es el transporte de energía eléctrica, garantizando el libre acceso a las redes de transmisión a todos los agentes del Mercado Eléctrico Mayorista (MEM).

[4] http://www.transelectric.com.ec/quienes.asp

Transelectric S.A. administra y opera un conjunto de subestaciones y líneas de transmisión en tensiones de 230 kV y 138 kV que transportan la energía producida por las centrales de generación hacia las empresas eléctricas de distribución en todo el país. A la fecha están conectadas al Sistema Nacional de Transmisión (SNT), en el territorio nacional 18 empresas de distribución y 11 empresas de generación. Se dispone de: 1251 km. de líneas de 230 kV, 1481 km. de líneas de 138 kV y 26 subestaciones a nivel nacional con una capacidad de transformación de 4382 MVA. La disponibilidad de las instalaciones está garantizada a través de 4 unidades de mantenimiento del SNT, distribuidas estratégicamente dentro del territorio nacional.

Empresas Generadoras

Termoesmeraldas	Termopichincha S.A.	Hidropaute
Hidroagoyán	Hidropucará	Electroguayas

Tabla 2 Empresas Generadoras

Fuente: CENACE

Empresas Distribuidoras

Empresa Eléctrica Quito S.A	Empresa Eléctrica Azogues C.A	Empresa Eléctrica Ambato S.A.
Empresa Eléctrica Milagro C.A.	Empresa Eléctrica Regional Centro Sur C.A.	Empresa Eléctrica Riobamba S.A.
Empresa Eléctrica Regional del Sur S.A.	Empresa Eléctrica Bolívar S.A.	Empresa Eléctrica Santo Domingo S.A.
Empresa Eléctrica Esmeraldas S.A.	Empresa Eléctrica Guayas - Los Ríos	Empresa Eléctrica Regional Manabí S.A
Empresa Eléctrica Regional del Norte S.A.	Empresa Eléctrica Regional El Oro S.A	Empresa Eléctrica Los Ríos C.A.
Empresa Eléctrica de Sucumbios S.A.	Empresa Eléctrica Península de Santa Elena C.A.	Empresa Eléctrica Provincial Cotopaxi S.A.

Tabla 3 Empresas Distribuidoras

Fuente: CENACE

2.1.2 PARTICIPACIÓN DEL CENACE EN LA ENERGÍA ELÉCTRICA DEL ECUADOR

Debido a los crecientes intercambios comerciales, así como también aportes y desarrollos tecnológicos, políticos (acuerdos de integración, redefinición rol Estado), ambientales (mejor uso recursos naturales), económicos (metodologías e instrumentos para creación y conformación de mercados) entre otros; influyen al aumento de la demanda eléctrica en nuestro país por tal situación fue necesaria la creación de un ente que permita realizar la gestión de control sobre los procesos y actividades para la administración de la Energía Eléctrica.

El sector eléctrico ecuatoriano se encuentra sumido en una profunda crisis, fundamentalmente de carácter económico, producto de factores endógenos y exógenos al mismo, que pone en peligro el suministro continuo y confiable de la energía eléctrica en las cantidades suficientes, y al menor costo posible, para el desarrollo de la sociedad ecuatoriana.

La Corporación Centro Nacional de Control de Energía – CENACE, fue establecida a partir de lo expuesto en el artículo 22 de la Ley de Régimen del Sector Eléctrico, publicada en el suplemento del Registro Oficial No. 43 del jueves 10 de octubre de 1996 y comenzó su funcionamiento el 1 de febrero de 1999. Se constituye como una Corporación civil de derecho privado de carácter eminentemente técnico, sin fines de lucro, cuyos miembros son todas las Empresas de: Generación, Transmisión, Distribución y Consumidores.

Misión

"La Corporación CENACE administra con seguridad, calidad, y economía, tanto el funcionamiento técnico del Sistema Nacional Interconectado e interconexiones internacionales, como el aspecto comercial del Mercado Eléctrico Mayorista (MEM), incluyendo las transacciones internacionales de electricidad, cumpliendo la normativa para satisfacer a sus clientes." [5]. Esto se consigue mediante la gestión de un talento humano calificado y comprometido,

[5] http://www.cenace.org.ec/

la disponibilidad de los sistemas tecnológicos de información, y del mejoramiento continuo del Sistema de Gestión de la Calidad.

Visión

Ser un organismo líder en la administración de mercados eléctricos mayoristas integrados, que asegure una alta calidad, confiabilidad, y economía del suministro de electricidad, propiciando el desarrollo socio-económico del país y de la Región Andina.

Función Principal

La función principal de la Corporación CENACE es tener a su cargo la administración de las transacciones técnicas y financieras del Mercado Eléctrico Mayorista – MEM, debe resguardar las condiciones de seguridad de operación del Sistema Nacional Interconectado – SNI, responsabilizarse por el abastecimiento de energía al mercado al mínimo costo posible, preservar la eficiencia global del sector y crear condiciones de mercado para la comercialización de energía eléctrica por parte de las empresas generadoras, sin ninguna discriminación entre ellas, facilitándoles el acceso al sistema de transmisión.

Esta función principal está expresada en el Artículo 23 de la Ley de Régimen del Sector Eléctrico. Entre otras funciones que están a cargo del CENACE para poder administrar de una manera eficiente el sector de la Energía Eléctrica, están las siguientes:

a) Recabar de todos los Actores del MEM sus planes de producción y mantenimiento, así como sus pronósticos de la demanda de potencia y energía de corto plazo;

b) Informar del funcionamiento del MEM y suministrar todos los datos que le requieran o que sean necesarios al Consejo Nacional de Electricidad - CONELEC;

c) La coordinación de la operación en tiempo real del SNI en condiciones de operación normal y de contingencia, ateniéndose a los criterios y normas de seguridad y calidad que determine el CONELEC;

d) Ordenar el despacho de los equipos de generación para atender la demanda al mínimo costo marginal horario de corto plazo de todo el parque de generación;

e) Controlar que la operación de las instalaciones de generación la efectúe cada titular de la explotación, sujetándose estrictamente a su programación;

f) Aportar con los datos que requiera el Director Ejecutivo del CONELEC para penalizar a los Generadores, de conformidad a lo señalado en el Reglamento respectivo, por el incumplimiento no justificado de las disposiciones de despacho impartidas;

g) Asegurar la transparencia y equidad de las decisiones que adopte;

h) Coordinar los mantenimientos de las instalaciones de generación y transmisión, así como las situaciones de racionamiento en el abastecimiento que se puedan producir;

i) Preparar los programas de operación para los siguientes doce meses, con un detalle de la estrategia de operación de los embalses y la generación esperada mensualmente de cada Central. ”

Estas, entre otras funciones garantizan, la calidad del servicio de energía eléctrica para el consumidor final, contribuyendo de esta manera al control de todo lo que tiene que ver con el sector eléctrico del País.

2.1.3 IMPORTANCIA DE LA PREDICCIÓN DE LA DEMANDA ELÉCTRICA EN EL ECUADOR

Debido a que el desarrollo de los pueblos está íntimamente relacionado con la productividad de los mismos, el consumo de energía eléctrica se vuelve indispensable para las máquinas y artefactos que pueden ayudar a incrementar la productividad. Y a la vez nos encontramos en un mundo globalizado, en

donde la energía eléctrica juega un papel fundamental, para el desarrollo de todos los pueblos. Es por esta razón que debemos estar conscientes que hay que satisfacer la demanda de toda la Nación, con índices de calidad, para satisfacer de una manera óptima las necesidades del usuario final.

Cuando se presentan problemas de exceso de capacidad de generación de potencia, o por el contrario de capacidad insuficiente, pueden tener costes muy elevados, no solo en términos económicos (tanto por la energía no facturada como por los perjuicios de esta índole que puedan sufrir los usuarios), sino también por la imagen que el cliente o abonado percibe en cuanto a la calidad del servicio que recibe.

La importancia de la predicción de la demanda de energía eléctrica surge, de forma obvia, de la incertidumbre asociada a una magnitud que se refiere al futuro. La mencionada predicción puede ayudar a determinar si, previsiblemente, se va a producir una carencia de capacidad generadora (y, en consecuencia, pudiera ser conveniente considerar la construcción de nuevas centrales de energía o simplemente impulsar la adopción de medidas de conservación de la energía) o, por el contrario, en el futuro existirá un exceso de capacidad que pudiera aconsejar la no utilización de parte del parque generador ya existente.

Según Kher, Sioshanci y Sorooshian (1987), “la industria energética es un sector de capital intensivo con inversiones a muy largo plazo”. Puesto que se necesita al menos una década para planificar y construir una nueva planta generadora, una previsión correcta de la demanda de energía eléctrica es un requisito imprevisible para lograr las metas previstas de calidad y fiabilidad del servicio, ya que la creciente dependencia de la electricidad aumenta los inconvenientes causados a los consumidores si se producen deficiencias en el suministro de energía eléctrica. La previsión de la demanda es una actividad esencial de los suministradores de energía eléctrica. Sin una adecuada representación de las necesidades futuras de generación eléctrica, los problemas de exceso de capacidad, o por el contrario de capacidad insuficiente, pueden tener costes sorprendentemente altos. La correcta previsión de la demanda también desempeña un importante papel en las

decisiones de una compañía eléctrica respecto a qué cantidad, y en qué época, será conveniente comprar (vender) energía a otras empresas del sector.

Es difícil que ambas desviaciones, por exceso o por defecto, no acaben repercutiendo sobre el usuario final. Si se produce una carencia de electricidad, el precio de ésta se incrementará y el abonado pagará más por la energía consumida. Si por el contrario, las erróneas predicciones se traducen en una superabundancia de energía, los costes asociados con la clausura de algunas plantas de potencia, u otros medios de disminuir el suministro, serán trasladados al consumidor. Pronosticar la demanda de energía, es una de las tareas más importantes y de mayor complejidad en el sector eléctrico, a escala mundial.

Existen al menos tres motivos para modelar la demanda de energía. En primer lugar, el suministro razonablemente fiable de energía es vital para el funcionamiento de la economía moderna. En segundo lugar, la ampliación de los sistemas de suministro de energía requiere muchos años. En tercer lugar, las inversiones necesarias en tales sistemas son altamente intensivas en capital, representando, en algunos países una considerable proporción de su Producto Interno Bruto.

Por otra parte, los índices habituales de fiabilidad eléctrica miden la probabilidad de que se pueda prestar el servicio requerido y, por tanto, los programas de planificación del sector eléctrico se basan, fundamentalmente, en las predicciones que se realicen sobre las puntas de potencia y energía demandada y, de acuerdo con ellas, se pueden adoptar las eventuales decisiones de incrementar la capacidad generadora instalada. Es obvio que, cuanto más acertadas sean tales predicciones, menores serán los riesgos de incurrir en inversiones innecesarias y/o de causar insatisfacción a los usuarios.

Estos pronósticos, son un insumo fundamental para cualquier agente que se dedique a la operación de mercados de energía de cualquier tamaño, y de ellos depende la determinación de las metas y objetivos empresariales, además de servir como insumo fundamental para él:

- Cálculo de balances eléctricos

- Planear la operación y ejecutarla
- Elaborar planes de expansión y de inversión
- Estimación de compras de energía requeridas y su costo
- Presupuestar los ingresos por ventas para cada sector de consumo.
- Calcular el margen esperado por la compañía.

Si bien es cierto que existen incontables formas para hacer pronósticos, desde los más simples hasta los más sofisticados, se debe tener en cuenta que todo pronóstico es en esencia la sugerencia de una sola posibilidad. Esto confirma que por más sofisticado que sea el método utilizado, el comportamiento del mercado energético no se ajusta al pronóstico hallado, sino que por el contrario, es simplemente la alternativa más viable que se encuentra, luego de hacer una serie de razonamientos basados en la información disponible

Se espera que cada vez, el planeamiento sea más preciso y así obtener mejores resultados en el sector eléctrico, satisfaciendo la creciente demanda de energía, de manera que se cumpla con las exigencias de calidad y continuidad en el servicio exigidas por el consumidor.

Parámetros de la demanda

Se ha mencionado en diferentes ocasiones, que el objetivo principal de la planeación en los sistemas de energía eléctrica, es satisfacer las necesidades de los consumidores bajo los parámetros de calidad, confiabilidad y seguridad.

Calidad

El suministro de energía a un usuario debe tener las siguientes características: ninguna distorsión de la onda, libre de ruido, magnitud constante, frecuencia constante y debe tener continuidad, es decir, no debe presentarse interrupciones de pequeña o larga duración, en donde la confiabilidad es un problema superado y la calidad de la potencia eléctrica se identifica principalmente con la calidad del voltaje.

Confiabilidad

Medida de habilidad de un sistema de potencia para suministrar la electricidad en todos los puntos de utilización con estándares aceptables de calidad y en la cantidad deseada. Esta confiabilidad puede describirse en términos de atributos de adecuación, seguridad, integridad y reposición de envío.

Seguridad

Medida de la habilidad de un sistema de potencia para responder adecuadamente, ente perturbaciones súbitas tales como: corto circuitos o salidas imprevistas de componentes. Así mismo, los enormes avances de nuestra época han sido posibles, fundamentalmente, debido al uso de la energía eléctrica. Un país con energía es un país con futuro.

2.2 MODELOS UTILIZADOS EN LA ACTUALIDAD.

En el presente se describen los modelos que usan en el CENACE para realizar la predicción de la Demanda Eléctrica.

2.2.1 MODELO DE REDES NEURONALES

La correlación existente entre la generación y la demanda de la energía eléctrica hace que su previsión sea llevada de tal manera que se tomen en cuenta la mayor parte de los elementos involucrados en éste ámbito, además debido a su importancia se debe tener diferentes enfoques que permitan dar un mejor criterio al momento de predecir la demanda eléctrica existente en el país.

En la actualidad existe una gran tendencia a utilizar modelos que permitan la resolución de problemas que no pueden ser descritos fácilmente mediante un enfoque algorítmico tradicional, es así que el modelo de redes neuronales permite manejar entre otras cosas, tomas de decisiones para la predicción de la demanda de la energía eléctrica. El análisis de los datos reales obtenidos tanto de las generadoras como del consumo (demanda eléctrica) permite definir a éstos datos como series de tiempo, para las redes neuronales estás series de tiempo se constituyen en relaciones no lineales.

En conclusión las redes neuronales son modelos no lineales que permiten representar una serie usando redes de elementos con cálculo aritmético sencillo, para aprender esa serie a partir de ejemplos.

La metodología aplicada está en función de las metas que se quiera alcanzar al realizar el pronóstico de corto plazo a mediano plazo lo que involucra tomar en cuenta las variables energéticas, demandas individuales, factores que afectan al perfil de la demanda para una buena predicción, tomar en cuenta todos éstos factores permite obtener un mejor análisis.

La red Neural aprende los factores involucrados en el comportamiento de la demanda, lo que ha permitido que éstas características sean analizadas por las redes neuronales, ya que ésta relación en el modelo no es explícito, los valores de entrada-salida de la red en intervalos regulares sirven para la obtención del modelo. No existe limitación en el número de variables de entrada.

El nivel de aprendizaje de la red se considera de acuerdo a las relaciones existentes en la serie de datos, esto por la interacción de los neuronios, los mismos que por su interacción forman una red neural, activándose por una señal de entrada generada por una señal de salida.

La forma de relacionar los datos dentro de la red se realiza a través de una distribución de los neuronios, así se tiene un grupo de entrada para la lectura de los datos de la serie, un grupo intermedio para la determinación de la relación no-lineal de los datos de la serie y por último un grupo de salida que permiten obtener el pronóstico de la demanda. Ésta topología se la conoce como Red Multilayer Perceptrón (MLP).

Los neuronios de la entrada dependen del número de valores considerados en la correlación que poseen con la serie a ser pronosticada, los del grupo intermedio dependen del nivel de interacción deseada, el número de conexiones establecidas en la red dependen de la característica de la serie evaluada, puesto que cada interconexión constituye una vía de transferencia de efectos positivos o negativos para cada valor ingresado en la red.

La función tanh(x) permite indicar el nivel de activación de cada neuronio, la ponderación de la conexión está dada por los pesos de los estados de entrada-salida con el conjunto de entrenamiento entrada-salida de la red.

La calidad de los datos ingresados determinan la habilidad de aprendizaje de la red, éstos datos deben ser normalizados antes de su ingreso, de no realizar ésta normalización podría existir un desbalance en la red.

Algoritmo de entrenamiento de la red neural

Conjunto de datos de entrada salida normalizados

Una misma función de activación para todos los neuronios diferenciable

La función objetivo a ser minimizada

Número de neuronios del grupo intermediario fijos

Inicialización del conjunto de pesos de la red, aleatoriamente con una distribución uniforme

El desempeño de la red neural se determina por:

- El grado de aprendizaje de la red
- La topología de la red, especialmente en los neuronios involucrados en el grupo intermedio.

2.2.2 MODELO ESTADÍSTICO ARIMA CON SPSS

Otra metodología aplicada en la obtención del pronóstico de demanda del Mercado Eléctrico Ecuatoriano en el corto plazo con un horizonte de 24 y 168 horas, prediciendo en períodos diarios de hora a hora, así como predicciones de los días de la semana, es por medio de un modelo de series temporales.

Este método de predicción utiliza el modelo estadístico ARIMA (Autoregressive Integrated Moving Average) (Promedio Móvil integrado Auto Regresivo.)

El método ARIMA o llamado también serie histórica o serie cronológica es una sucesión de valores observados de una variable referidos a períodos de tiempo generalmente regulares. El análisis de una serie temporal consiste en hacer uso de estos datos para elaborar un modelo que describa adecuadamente el comportamiento de esta variable en el pasado, y permita realizar predicciones satisfactorias.

Para la obtención de estimaciones con propiedades estadísticas adecuadas de los parámetros de un modelo ARIMA, es necesario que la serie muestral que utilizamos para la estimación sea estacionaria en media y varianza.

Estacionariedad de una Serie

Para que resulte posible la obtención de una estimación con propiedades estadísticas, la serie debe ser estacionaria; es decir que la serie sea invariable con respecto al tiempo. Cuando las características de la serie cambian con respecto al tiempo, el proceso no es estacionario.

Una posibilidad para definir la estacionariedad de una serie, es verificar que las propiedades del proceso no son afectadas por cambios de origen temporal, es decir, cuando al realizar un mismo desplazamiento en el tiempo de todas las variables de cualquier distribución conjunta finita, resulta que esta distribución no varía.

En resumen, la hipótesis de estacionariedad implica que la media y la varianza del proceso sean constantes y que las autocovarianzas, y las autocorrelaciones dependen solamente del retardo, y no del momento del tiempo. Para obtener un pronóstico de la serie de tiempo se utilizan algunas funciones del programa estadístico SPSS, en el CENACE se ha creado una barra de menús personalizados llamada “Estimación de Demanda” la cual se encuentra estructurada para cada día de la semana. Permitiendo cargar los datos del día en estudio y con otras opciones permite obtener el pronóstico de la demanda. La barra de herramientas personalizada consta de 14 botones (2 por cada día) en la cual si deseamos obtener la proyección de demanda para un día en particular el primer botón de ese día carga la serie de datos correspondiente y el segundo botón permite la ejecución de la proyección.

Manejo de datos y pruebas de modelos

Se puede modificar las series establecidas, dentro del ambiente del SPSS para ensayos, el estudio ó análisis que se tenga no afecta ni a los datos, ni a los modelos establecidos para el proceso diario de la demanda.

Series con varias estacionalidades

Una de las principales ventajas de usar el SPSS es que se puede tener estimaciones de demanda con modelos multiestacionales, como es el caso de la programación semanal en los que tenemos los datos de la semana, con sus siete días y cada uno con sus 25 horas.

Interacción con otros ambientes de Windows

Se puede importar ó exportar datos con otros utilitarios como archivos de excel, asccii, etc. los cuales permiten la generación de reportes según requiera el usuario.

Procedimiento para proyecciones en Días Normales, Feriados y Especiales

Días Normales

Analiza y evalúa el pronóstico obtenido y si es necesario se realiza un ajuste adicional considerando los días en los que se prevé acontecimientos especiales como partido de fútbol, suspensión de clases, etc.

Días Feriados y Especiales

Es necesario obtener la demanda del día de análisis dándole el mismo tratamiento que un día normal. Más, para obtener el pronóstico de la demanda horaria requerida se hace necesario.

Identifica en la estadística de la demanda días semejantes al del estudio, así como, los días normales cercanos al día de estudio y los semejantes para obtener factores de relación con cada uno de ellos.

CAPITULO 3. DESARROLLO DE LA APLICACIÓN USANDO UN MODELO DE APRENDIZAJE

En éste capítulo se describe las principales características y actividades del modelo de aprendizaje a ser utilizado para la predicción de la demanda eléctrica.

Además se define y detalla la metodología de desarrollo con que se elaboró la aplicación. Por lo que se establece utilizar para el desarrollo de la aplicación el PUD proceso Unificado de Desarrollo y como metodología el UML Lenguaje Unificado de Modelado, para lo cual se procede a realizar una introducción de conceptos que permitan finalmente realizar la adaptación del caso de estudio.

3.1 SELECCIÓN DEL MODELO DE APRENDIZAJE

Introducción

En el Ecuador se mantiene un registro diario de la demanda eléctrica por lo que es importante el análisis de la información en función de los factores que influyen de manera directa o indirectamente su variación para poder prever las respectivas acciones cuando así se requiera.

El estudio de ésta temática y el deseo de conocer los posibles factores determinantes que expliquen el comportamiento de la demanda eléctrica han conducido a la búsqueda de modelos alternativos aún no tomados en cuenta en la investigación del mismo.

La problemática referente a éste tema integra en la práctica la totalidad de la demanda eléctrica, siendo objeto de análisis en la presente investigación al considerar que se puede contribuir a un mayor y mejor conocimiento de esta realidad.

Para llegar a comprender cual es el modelo de predicción aplicable al caso de estudio es necesario revisar algunos criterios en los que se suele utilizar los respectivos modelos al realizar éste tipo de casos.

Tipos de Problemas de Predicción

Clasificación

El problema de clasificación es aquel que consiste en asignar una clase a un objeto, una etiqueta de clasificación, luego de determinarse una clasificación en función de los datos involucrados.
Ejemplo: clasificar un producto como “bueno” o “malo” en un test de control de calidad

Regresión

El problema de regresión es una generalización del problema de clasificación consiste en asignar un número real como salida, luego de determinarse un análisis y tratamiento de los datos.
Ejemplo: predecir la temperatura que habrá la semana que viene

Clustering (agrupamiento)

El problema de clustering se basa en organizar objetos en grupos que tengan sentido. Así se halla una agrupación de objetos que puede ser jerárquica.
Ejemplo: organizar plantas en una taxonomía de especies

Descripción

El problema de descripción representa un objeto en términos de una serie de primitivas. Realizando una descripción estructural o lingüística.
Ejemplo: etiquetar una señal ECG en términos de complejos P, QRS y T

Como se menciona, existen modelos aplicados a cubrir ciertos problemas de predicción, en el caso de la predicción de la demanda eléctrica como se nombró en el Capítulo 2, se hace referencia a un modelo estadístico y un modelo basado en redes neuronales, razón por la que se ha seleccionado un

modelo aún no tomado en cuenta. Éste modelo es el de árboles de decisión, está basado en el aprendizaje inductivo supervisado, que utiliza el algoritmo de clasificación C5 de Quinlan.

Ahora de mencionan términos que son comunes dentro del modelo seleccionado, lo que permite tener familiarización y entendimiento de éste.

Atributo

Un atributo es cualquier aspecto distintivo, cualidad o característica utilizada para denominar a un objeto. Éste puede tomar valores simbólicos, como por ejemplo: color, sexo; o numéricos como por ejemplo: altura, temperatura.

La combinación de los atributos se representa en un vector columna de d dimensiones llamado vector de atributos. El espacio de atributos es el espacio de d dimensiones definido por este vector, los objetos se representan como puntos del espacio de atributos.

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_d \end{bmatrix} \quad VectordeAtributos$$

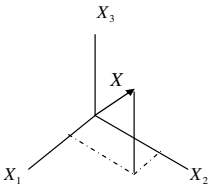


Figura 2 Espacio de atributos

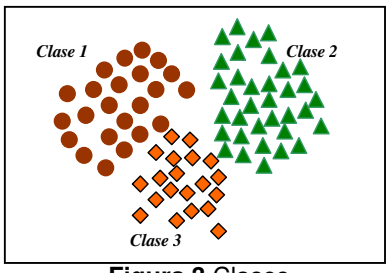


Figura 3 Clases

Fuente: Sánchez-Montañez, Teoría y Aplicaciones a Problemas de Predicción

En un buen vector de atributos se debe considerar la calidad del mismo que esta en función de la capacidad de discriminar ejemplos de clases diferentes. Así los atributos de ejemplos de la misma clase deberían tener valores similares y los atributos de ejemplos diferentes clases deberían tener valores diferentes.

Figura 4 Atributos buenos **Figura 5** Atributos malos

Fuente: Sánchez-Montañez, Teoría y Aplicaciones a Problemas de Predicción

Propiedades Relacionadas Con Los Atributos

Separabilidad Lineal

Esta propiedad nos permite distinguir entre dos atributos los cuales deben ser separables linealmente, es decir la distinción de los valores de los atributos sean fácilmente diferenciables y no sean confundibles así claramente se podrá determinar que valores pertenecen a cada atributo, sin embargo aun cuando loa atributos no puedan cumplir separabilidad lineal se puede lograr una algunos métodos son capaces de aprender.

Separabilidad No Lineal

Ésta propiedad permite distinguir los valores de los atributos en su espacio que aunque no cumplen una separabilidad lineal son claramente diferenciables.

Altamente Correlacionados

Ésta propiedad permite establecer a un atributo en un valor específico, cuando otros atributos tienen similares valores específicos.

Multi-modal

Ésta propiedad involucra a las propiedades antes mencionadas.

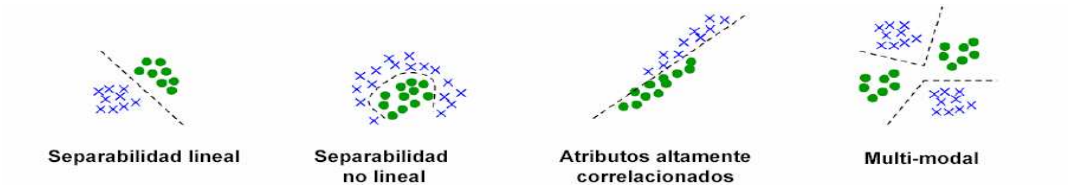


Figura 6 Relaciones entre atributos

Fuente: Sánchez-Montañez, Teoría y Aplicaciones a Problemas de Predicción

Clasificador

Un clasificador es el encargado de separar el espacio de atributos en regiones de decisión, cada una de las regiones separadas con una clase asignada. Los límites que se lleguen a establecer entre las diferentes regiones se denominan fronteras de decisión. La clasificación del vector de atributos, consiste en determinar a qué región de decisión pertenece, y asignarle la clase correspondiente o asignada.

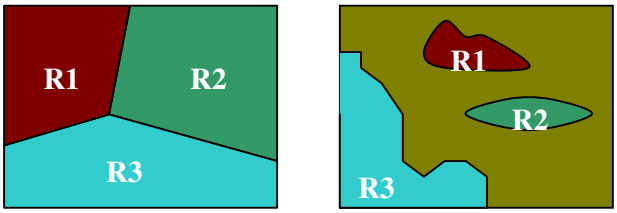


Figura 7 Clasificación de regiones (atributos)
Fuente: Sánchez-Montañez, Teoría y Aplicaciones a Problemas de Predicción

A un clasificador se puede ver como un conjunto de funciones discriminantes en la que cada función es encargada de determinar a que clase asignada pertenece cada atributo.

EXPLICACIÓN DEL MODELO

A continuación se mencionan los principales conceptos y una explicación del modelo de Árboles de Decisión, en el ANEXO Árboles de Decisión se realiza una ilustración que complementaria a ésta explicación.

Definición, Objetivo y Condiciones para su Aplicación

Definición

Un árbol de decisión es un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas.

Objetivo

Aprender de un árbol de decisión consistente con los ejemplos para posteriormente clasificar ejemplos nuevos.

Lo que se busca al aplicar un modelo basado en árboles de decisión es aproximar el valor de una variable (continua o discreta) a partir de otras variables.

Para poder aplicar el modelo es necesario de la definición de una variable dependiente. Dependiendo del tipo de la variable dependiente, los árboles reciben distinto nombre.
Así, los árboles son denominados de acuerdo a la variable dependiente que se utilice, así serán árboles de regresión si la variable dependiente es continua y árboles de clasificación si la variable dependiente es discreta.

Se debe buscar las variables y más concretamente las modalidades de las mismas que más separan los distintos valores que toma la variable dependiente.
El árbol construido es el que se utiliza para clasificar o predecir nuevos casos. El árbol no podrá clasificar nuevos casos que presenten categorías de variables que no se hayan utilizado en la construcción del mismo

Condiciones para su aplicación

1.- Tipo de variables:

El tipo de variables que se puede utilizar en éste modelo pueden ser de cualquier tipo de variable ya sea continua o discreta, numérica o categórica. Es recomendable tener variables con un normal número de valores.

2.- Outliers y Missings:

Los árboles de decisión es una técnica poco sensible a los outliers, los resultados no cambian significativamente por la presencia de outliers. Aunque esto depende de la manera con que se analicen los datos.
Los árboles de decisión pueden tratar a los valores missings como tal, los árboles los reagrupan formando una categoría nueva de la variable. Aunque esto depende de la manera con que se analicen los datos.

Principales Factores del Modelo

A continuación se describe el modelo basado en árboles de decisión explicando brevemente los principales factores dentro del modelo:

1.- Elementos de un árbol.
2.- Construcción de un árbol
3.- Criterio de división
4.- Conjunto de Reglas
5.- Condiciones de crecimiento de un árbol
6.- Interpretación de los nodos hoja
7.- Realización de predicciones

1.- ELEMENTOS DEL ÁRBOL

Los principales elementos que posee un árbol son los siguientes:
Raíz: Se trata del nodo inicial en el que se encuentran todas las observaciones (datos).
Nodo: Grupo de observaciones que cumplen unas condiciones determinadas.
Nodo hijo: Los nodos resultantes de dividir un nodo superior
Rama: cada uno de los diferentes caminos que unen los nodos padres con los nodos hijos.
Nodo hoja: nodo sin hijos.

2.- CONSTRUCCIÓN DE UN ÁRBOL

La construcción del árbol empieza en el punto de partida, es decir parte de un nodo raíz, de ahí que todas las observaciones están agrupadas en la raíz. El crecimiento del árbol esta en función de un nodo que se divide en dos o más hijos aplicando una condición no arbitraria siguiendo un criterio de división sobre una de las variables independientes. Para culminar en un fin de árbol, que son los nodos hojas, y no se divide más, el árbol no sigue creciendo por esta rama. Además el conjunto de observaciones que pertenecen al mismo nodo hoja presentan las mismas características (mismo comportamiento de las variables independientes).

El conjunto de observaciones se va separando de arriba a abajo, desde la "raíz" del árbol hasta los nodos "hoja", para ello, utiliza cada vez un conjunto de reglas excluyentes.
Una observación pertenece a un único nodo hoja y a medida que el árbol crece se generan más nodos (reglas), también el número de observaciones en cada uno de ellos es más pequeño y se observan más diferencias entre ellos (comportamiento de la variable dependiente con respecto al nodo padre). Ver apartado Construcción en el ANEXO Árboles de Decisión.

3.- CRITERIO DE DIVISIÓN

El criterio de división debe buscar divisiones que discriminen mejor con el fin de obtener los mejores resultados en base al análisis que se realice.
Por lo general se selecciona la variable más discriminante donde buscan el corte de la misma también más discriminante. El criterio de división no es más que una función f, de forma que al aplicarla sobre la probabilidad de la variable de análisis elija el "mejor" corte. Esta función f puede tener diversas formas y el criterio de división depende de la misma, además también se depende del algoritmo que se utilice para construir el árbol de decisión.

Al tratarse de variables discretas se divide un nodo en función de los distintos valores que toma. Ejemplo: Tipo de Demanda, en caso que se haya definido como valores Alta, Media y Baja.

Las variables continuas dividen un nodo en función de tramos de valores. Ejemplo: Demanda: Cantidad de días con demanda inferior a 500MWh

Todos los criterios de división buscan una división s, con el menor valor de:

$$I(s) = \sum_{j=1..n} p_j f(p_j^1, p_j^2, ..., p_j^c) \quad donde:$$

$n = número\ de\ nodos\ hijos\ de\ la\ división$

$p_j = probabilidad\ de\ caer\ en\ el\ nodo\ j$

$p_j^1 = proporción\ de\ elementos\ de\ clase\ 1\ en\ el\ nodo\ j$

$p_j^c = proporción\ de\ elementos\ de\ clase\ c\ en\ el\ nodo\ j$

Criterio de división a partir de la definición de la función de impureza f

Criterio	$f(p_1, p_2, ..., p_c)$
Error esperado	$1 - \max(p_1, p_2, ..., p_c)$
GINI (CART)	$1 - \sum (p^i)^2$
Entropía ((gain), ID3,C4.5,C5)	$-\sum p^i \log(p^i)$
DKM	$2(\pi\, p^i)^{1/2}$
X^2 (CHAID)	$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{w} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

Tabla 4 Criterio de división

Fuente: Sánchez-Montañez, Teoría y Aplicaciones a Problemas de Predicción

4.- CONJUNTO DE REGLAS

El modelo de árboles de decisión también se puede entender como un conjunto de reglas organizadas de forma jerárquica.

La división de un nodo padre en sus hijos es debida a la aplicación de una condición sobre alguna de las variables independientes. Cada rama o camino desde el nodo raíz hasta cualquier nodo hoja, corresponde a una regla (conjunto de condiciones).

5.- CONDICIONES DE CRECIMIENTO DE UN ÁRBOL

El crecimiento de un árbol de decisión está en función del número de las variables utilizadas así como también del número de valores correspondientes a cada variable.

Si se tiene un árbol demasiado grande, se puede deber a que el modelo es demasiado específico y se corre el peligro de que se ajuste demasiado a los datos de entrenamiento, en éste caso se puede dar una mala clasificación de nuevos casos produciéndose la posibilidad de clasificar ruido en caso de haberlo.

Para solucionar éste inconveniente se puede acortar reglas(ramas del árbol) con el fin de obtener un árbol más corto y más general.

Prepoda y Post poda

Se trata de procesos que acortan reglas de un árbol.

Prepoda

La prepoda se realiza durante la construcción del árbol, además de busca determinar el criterio de parada. Ejemplo:
Condición de crecimiento a 10 niveles máximos de profundidad
Condición sobre el número de observaciones mínimo en cada nodo final

Post poda

La post poda se realiza después de la construcción del árbol, consiste en eliminar los nodos de abajo a arriba hasta un cierto límite. Con lo que se realiza una visión completa del modelo. La post poda viene desencadenada por:

- Una excesiva complejidad del modelo (muchas ramas y nodos hoja).
- Por falta de interpretabilidad de negocio de algunas de las reglas finales.
- Por cambios importantes en la probabilidad de impago asociada al nodo al cambiar el conjunto de datos (validación).

En ambos casos los criterios de poda están basados en:

- El número mínimo de observaciones por nodo para poder dividirlo
- El número mínimo de observaciones por nodo hoja
- La máxima profundidad del árbol
- Técnicas más sofisticadas (error esperado, criterio Minimum Description Length, etc).

6.- INTERPRETACIÓN DE LOS NODOS HOJAS

Para la interpretación de los nodos hojas dependen del tipo de variable objetivo que se tenga.

Si se tiene una variable objetivo discreta, se debe tomar en cuenta para su interpretación:

- Los valores posibles de la variable dependiente
- El reparto de las modalidades de la variable dependiente
- El número de observaciones en el nodo hoja
- El reparto en porcentaje, del total de las observaciones que caen en este nodo en categorías de la variable dependiente
- El valor que se predice a un nuevo caso que cae en este nodo hoja

Para una variable objetivo continua:

- La media de la variable objetivo del total de observaciones que hay en este nodo hoja
- El número de observaciones en el nodo hoja
- Desviación típica de la variable objetivo del total de observaciones que hay en éste nodo hoja

7.- REALIZACIÓN DE PREDICCIONES

Para la realización de predicciones se deben tomar en cuenta las siguientes observaciones:

- Se tienen tantas predicciones como nodos hojas hay.
- Se tienen tantas predicciones distintas como reglas finales componen el árbol.

Predicción de un nuevo caso

Con el fin de definir la predicción para un nuevo caso se aplica el siguiente procedimiento:

Recoger el valor de las variables independientes que constituyen el modelo final para este nuevo caso

Asociarle, a través del conjunto de reglas, el nodo hoja correspondiente (uno y solo uno por la propia construcción del árbol de decisión)
Predecir el valor de la variable objetivo en función de los datos del nodo hoja correspondiente.

Predicción para una Variable Objetivo Discreta

La predicción en caso de tratarse de una variable objetivo discreta tiene el siguiente tratamiento:

Se determina la variable objetivo a ser evaluada.
Se realiza el reparto, en porcentaje, de las observaciones que caen en éste nodo en categorías de de la variable dependiente.
Si cumple con el conjunto de reglas correspondientes, entonces se predice el valor de la variable objetivo del nodo correspondiente.

Existe un caso particular de la probabilidad de la variable objetivo discreta, en la que ésta es una variable objetivo binaria (0/1), así el valor que de predice es una probabilidad, para un nuevo caso se le asigna la probabilidad de que en el futuro, el valor de la variable objetivo sea 1.

Para extraer la probabilidad del nodo hoja, se evalúa el porcentaje de estimación para cada clase y se opta por la de mayor valor.

Predicción para una Variable Objetivo Continua

La predicción en caso de tratarse de una variable objetivo continua tiene el siguiente tratamiento:

Para un nuevo caso que cae en el nodo hoja tendrá como predicción de la variable objetivo la media del nodo hoja, si se realiza así la predicción se constituye en una variable discreta. Puesto que hay un número finito de nodos finales los distintos valores que se predicen son limitados. Entonces, las predicciones constituyen una variable discreta que “estima” el valor de una variable continua (variable objetivo).

Validación de Ajuste

La validación de ajuste advierte que en la aplicación del modelo debe existir una comprobación de los resultados obtenidos, esto se realiza a través de:

Matriz de clasificación

La matriz de clasificación indica el número de aciertos del modelo teniendo en cuenta una ponderación de costes en caso necesario. (coste de clasificar mal).

Verificación de las Reglas

Se trata de contrastar si sobre el conjunto de validación las reglas discriminan del mismo modo o intensidad. Además podemos consultar con expertos la validez de las reglas.

Algoritmo C5/See5

El algoritmo de inducción de reglas y árboles de decisión See5 de Quinlan es una técnica procedente del campo de la Inteligencia Artificial que, a través del aprendizaje inductivo, obtiene las características que más diferencian a los atributos objeto de estudio. Dicho algoritmo procede del Concept Learning System, ideado por Hunt et al. (1966), y representa una extensión de los algoritmos ID3 y C4.5. Quinlan está considerado como una referencia mundial en el ámbito de la Inteligencia Artificial habida cuenta la repercusión de los modelos construidos.

Estos métodos inductivos de Inteligencia Artificial realizan particiones binarias sucesivas en el espacio multidimensional de las variables explicativas, para así construir un árbol de clasificación, de tal forma que en cada partición se selecciona la variable que aporta una mayor cantidad de información atendiendo a una medida de entropía. A partir del árbol se elaboran unas reglas clasificadoras de sencilla interpretación, las cuales permitirán definir las características diferenciales de los atributos considerados. Una mayor descripción del algoritmo se adjunta en el Anexo Árboles de Decisión.

Ventajas

- Superan a los sistemas expertos en que no necesitan de la intervención del experto humano para la inferencia de las reglas clasificadoras, ya que éstas se elaboran automáticamente.

- Frente a las redes neuronales presentan la ventaja de que las reglas son mucho más comprensibles por el usuario que la topología de una red, puesto que una red neuronal, por sencilla que sea, es un modelo de "caja negra", que no permite valorar la importancia relativa de cada una de las variables explicativas.

- La utilización del algoritmo presenta importantes ventajas frente a las técnicas estadísticas, pues su principal característica reside en su mayor flexibilidad, al no requerir ninguna de las hipótesis iniciales sobre la estructura de los datos y sus interrelaciones, ni estar sujeta a las restricciones requeridas por la mayor parte de las técnicas estadísticas paramétricas, como son la normalidad de las distribuciones de las variables consideradas y la igualdad de las matrices de varianza-covarianzas. Estos requisitos no son verificados para la muestra analizada, circunstancia que justifica el uso de esta técnica. Asimismo, resulta mucho más fácil de interpretar, y además, como así se concluye en los estudios referidos, ofrece resultados superiores si el número de datos considerados no es muy amplio.

Resultados

La información que se proporciona hace referencia a las reglas clasificadoras, las cuales constituyen una simplificación bastante precisa del árbol. Para cada una de las reglas se muestra el número de veces que ha sido aplicada, es decir, el número de datos que la verifican y el número de errores de clasificación cometidos. De forma separada al conjunto de reglas, finalmente se ofrece una matriz de clasificación que resume la eficiencia clasificadora lograda. Tal y como se han descrito los niveles de la variable dependiente, las clases o agrupaciones se definen de acuerdo al mejor criterio de la persona

encargada de hacer el análisis de la demanda. Una más detallada descripción del modelo se adjunta en el Anexo Árboles de decisión.

Análisis de Datos

Introducción

Para poder predecir de manera eficaz, se tiene que tomar en cuenta algunos factores involucrados con el análisis de los datos, como los atributos a intervenir, sean estos continuos o discretos, el objetivo de la predicción, y los resultados a obtener.

Cuando se realiza un análisis de datos que intervienen en un estudio, esto involucra emitir resultados cuantificables o cualitativos de dicho estudio o experimento. La claridad de dicha presentación es de vital importancia para la comprensión de los resultados y la interpretación de los mismos. A la hora de representar los resultados de un análisis de datos se lo debe realizar por medio de una tabla, un diagrama o un gráfico los mismos que pueden ayudarnos a mostrar de un modo más eficiente nuestros datos. El análisis de los datos, de ser el caso, nos permite presentar la información de forma que ésta se pueda visualizar de una manera más sistemática y resumida. Los datos que nos interesan dependen, en cada caso, del tipo de variables que estemos manejando.

Para variables discretas, se debe conocer la frecuencia y el porcentaje del total de casos que corresponden en cada categoría. Una forma muy sencilla de representar gráficamente estos resultados es mediante diagramas de barras o diagramas de sectores. En los gráficos de sectores, también conocidos como diagramas de pastel, se divide un círculo en tantas porciones como clases tenga la variable, de modo que a cada clase le corresponde un arco de círculo proporcional a su frecuencia absoluta o relativa.

Requisitos para el modelo de Aprendizaje

El modelo de Aprendizaje funciona mediante el uso de datos históricos (casos anteriores), por lo que es necesario, crear y utilizar una base de datos donde

consten dichos datos históricos con el fin de conocer el comportamiento de los datos.

La base de datos que se utiliza para este trabajo, ha sido creada en función a los datos proporcionados por el CENACE. Estos datos constan del valor de la demanda registrada para cada hora en punto del día, a más de las 19h30 puesto que es en este instante es donde se registran comúnmente los valores más altos de la demanda eléctrica durante un día.

Los datos entregados, constan desde el 01 de enero del 2003 hasta el 11 de noviembre del 2005. Es por esta razón que la base de datos consta de 27196 registros.

Propiedades de la base de datos (datos históricos)

Para la elaboración de la base de datos se contempla las siguientes características que se mencionan a continuación:

Los casos deben ser medidos en las mismas condiciones que los que queremos predecir.

El histórico que consta en la base de datos, sirven para evaluar el caso de estudio, que es la predicción de la Demanda Eléctrica. Es por esto que se utiliza atributos que están relacionados con la predicción. Estos están sujetos a las mismas condiciones cuando se quiera realizar la predicción.

Los atributos a utilizar para la Predicción de la Demanda Eléctrica se establecieron a través de los datos históricos proporcionados por el CENACE y son los siguientes:

Atributo	Descripción
Dia	Atributo que indica un día específico con respecto a un mes.
Mes	Atributo que se refiere al mes con respecto a un año.
Anio	Atributo que representa el año en el que fue registrado un valor determinado.
DiaLaborable	Atributo que indica si un día es laborable o no lo es.

	Entendiendo por laborable cuando es un día de Lunes a Viernes que se trabaje normalmente.
Feriado	Atributo que especifica el nombre de un feriado Nacional, en el cual no se labora.
NombreDia	Atributo que indica el nombre de un día en la Semana.
ValorDemanda	Atributo que indica el valor numérico registrado del valor de la demanda eléctrica. Los valores de este atributo están dados en MW.
Clima	Atributo que se refiere al estado del clima en determinado mes del año.
Predicción	Clase resultado que da información en forma de rangos, sobre la demanda eléctrica a predecir.

Tabla 5 Descripción de Atributos

Fuente: Los Autores

Los casos no deben tener ningún sesgo relevante frente a los casos a predecir.

Hay que tener en cuenta que los atributos que se utilizan para la Predicción deben estar en función del caso de estudio y de los resultados que se quieran obtener, es así que la base de datos es variada en cuanto a datos a utilizar, los mismos que pueden ser un factor importante en el diagnóstico.

Esta propiedad se ha tomado en cuenta para establecer los atributos que intervienen en la predicción de la demanda eléctrica y que fueron mencionados en el punto anterior.

Codificación o transformación de las variables

La codificación o transformación de variables, permite definir las variables a utilizar dentro de un modelo, en función del tipo de análisis y la conveniencia de su manejo para una mejor obtención de resultados.

A continuación se menciona factores a tomarse en cuenta para realizar dicha transformación:

- Utilidad
- Tipificación de las variables

- Discretización de las variables
- Creación de variables dicotómicas
- Otras transformaciones

Utilidad

Cuando se va a trabajar con modelos se conoce que para cada uno de ellos se utiliza un determinado tipo de variable, además de los supuestos(normalidad, linealidad, multicolinealidad, etc…)
A través de los análisis descriptivos se puede obtener información que requiere tratamiento de datos, ya que los datos iniciales de partida pueden no representar todos los aspectos de negocio deseables.

Las acciones que se pueden llevar a cabo es la transformación (tipificación de variables, operadores, etc.) y creación de variables (variables elaboradas, discretización, variables binarias, etc…)

Tipificación de las variables

Si se desea realizar la comparación de dos variables se necesita que éstas tengan la misma escala de medida. La escala de medida de la variable puede influir en el papel de la misma en el modelo. Se puede establecer una correspondencia de sus valores con los de otra variable con distribución normal, con media 0 y varianza 1, a la que se llama variable normal tipificada.

Proceso de Tipificación de las Variables

El proceso de tipificación de variables continuas consiste en obtener una nueva variable con el fin de obtener mejor resultados en su tratamiento, a continuación se describe el proceso de obtención:

- Calcular la media de la distribución de la variable X
- Calcular la desviación típica de la distribución de la variable X
- Crear un nueva variable mediante la siguiente fórmula:

$$newX = \frac{\left(X - \overline{X}\right)}{Std(X)} \qquad \text{Fórmula 1}$$

Discretización de las Variables

Consiste en convertir una variable continua a una variable discreta, ésta discretización se realiza por varios motivos, que a continuación se mencionan:

1.- El error de medida de la variable es grande
2.- Existen cortes de la variable significativos
3.- El modelo a aplicar solo acepta variables discretas

Proceso de Discretización de las Variables

En el caso de que exista variable dependiente (objetivo), existen técnicas que indican por donde separar la variable y cuantas clases hacer. Éstas técnicas buscan los cortes maximizando un comportamiento diferente de la variable dependiente en cada una de las clases generadas (entropía, información mutua, árboles de decisión, ...)

En el caso de que no exista variable dependiente, la discretización más sencilla es aquella que realiza intervalos del mismo tamaño. Por lo que por lo general se suele utilizar los percentiles, el máximo y el mínimo de la distribución.
También se puede discretizar utilizando intervalos de la misma amplitud siguiendo criterios del negocio. Hay que tomar en cuenta que una mala discretización de los datos puede hacer variar mucho los resultados. La forma de discretización de la variable a predecir se describe en el Anexo Cálculo de Intervalos

Creación de Variables Dicotómicas

Se define como variable dicotómica aquella que solo admite dos categorías que definen opciones o características mutuamente excluyentes u opuestas tales como (Y=SI, Y=NO); (Y=0, Y=1), (Y=Encendido, Y=Apagado). La creación de una variable dicotómica, consiste en convertir una variable discreta a numérica. Ésta creación se da principalmente cuando el modelo a aplicar solo acepta variables numéricas.

Realización:

Uso de variables dicotómicas (variable binaria 0 o 1).
Creación de tantas variables dicotómicas como categorías distintas tenga la variable. La variable dicotómica toma el valor 0 o 1 dependiendo de si la variable discreta toma o no un determinado valor. Para el caso de estudio, se transformó solo la variable Mes, asignando el número 1 para Enero y consecutivamente ascendente para los demás meses.

Ventajas de las Transformaciones

Al realizar la transformación de variables se pueden obtener los siguientes beneficios:

Creación de nuevos indicadores del negocio

En éste caso se pueden obtener variables más elaboradas que permitan determinar nuevos aspectos que mejoren el entendimiento del tratamiento de la información analizada, ejemplos: creación de ratios, tratamiento de fechas, etc… Ejemplo:
A partir del mes del año creación de variables estacionales (fines de semana, momento del mes, etc…).

Mejora de los resultados de un modelo

Si se encamina a las variables de tal manera que a las variables involucradas se le pueda dar el sentido al caso de estudio. Así también cuando se tiene una buena interpretación de las variables también así será fácil interpretar los resultados.

Realización de Transformación

Variables numéricas:

La transformación en caso de las variables numéricas se lo realiza en base a operaciones matemáticas básicas de uno o más argumentos. Sumas, restas, producto, máximo, media, cuadrado, etc…

Variables categóricas o discretas:

La transformación en caso de las variables categóricas o discretas se lo realiza en base a operaciones lógicas. Conjunciones, disyunciones, negación, igualdad...

No es recomendable la generación de variables que el modelo sea capaz de obtener, porque se estaría redundando en su contenido por ejemplo:

Si se realiza la transformación de una variable discreta, en los siguientes términos:

Rango_1 = verdadero si demanda < 400 MHW, es poco útil si utilizamos un modelo de árbol de decisión, puesto que el propio modelo nos genera reglas.

PROBLEMAS EN LA CONSTRUCCIÓN DEL MODELO

Los problemas que se pueden llegar a tener cuando se construye con un determinado modelo. Hay que tener en cuenta que un modelo no siempre se ajusta en forma perfecta por lo que existe un margen de error, entonces el modelo lo que busca es minimizar el error cometido.

En algunos casos se produce un fenómeno cuando el modelo se ajusta demasiado a los datos utilizados para la construcción del mismo.

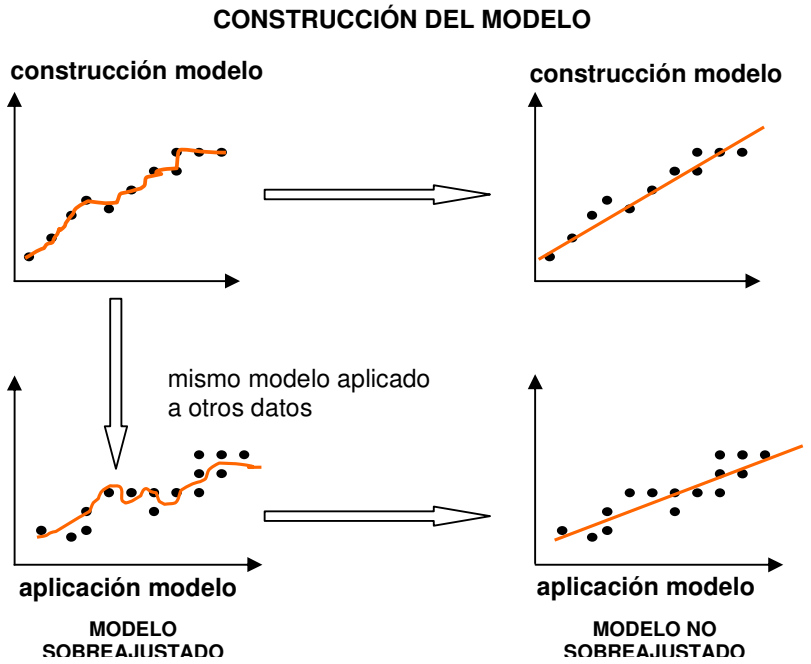


Figura 8 Construcción del modelo

FUENTE: Sánchez-Montañez, Teoría y Aplic. a Problemas de Predicción

Frente a un modelo sobre ajustado no cabe esperar buenos resultados cuando se utilizan datos nuevos

Para evitar el sobre ajuste es necesario la utilización de un conjunto de validación para comparar el comportamiento de la función de error entre el conjunto de validación y el conjunto de entrenamiento.

El sobre-ajuste de parámetros impide realizar una generalización apropiada para nuevos valores de las variables de entrada.

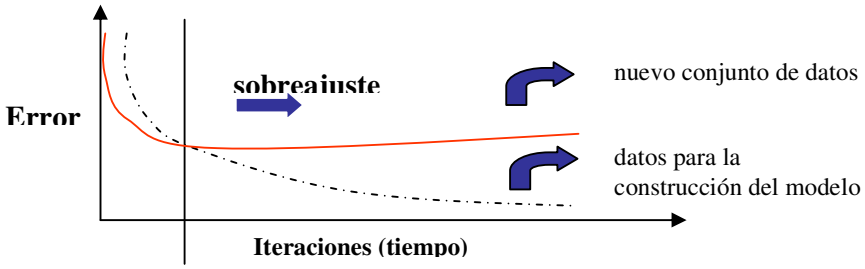


Figura 9 Sobre ajuste de parámetros

Fuente: Sánchez-Montañez, Teoría y Aplicaciones a Problemas de Predicción

Aplicado al caso de estudio

Tomando en cuenta lo descrito anteriormente acerca de la tipificación y codificación de variables se menciona el proceso que se aplica al caso de estudio. Mencionando los valores que van a tomar de aquí en adelante los diferentes atributos.

Cada atributo a utilizar, así como también la clase resultado; es decir la predicción, toman valores únicos, los cuales pueden ser valores numéricos u opción.

A continuación se determina el tipo de atributo y sus respectivos valores:

Atributo Día

Este atributo se refiere al número que tiene cada día en el mes. El Atributo Día es de tipo Opción, es decir que tiene valores que se repiten continuamente, los valores que tiene el atributo Día son los siguientes:

Valores del Atributo Día
1, 2, 3, 4, 5, 6, 7, 8, , 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23,24, 25, 26, 27, 28, 29, 30, 31.

Tabla 6 Valores del atributo Día

Fuente: Los Autores

Atributo Hora

Este atributo se refiere a la hora que tiene cada día. Este atributo es de tipo Opción, los valores de este atributo son los siguientes:

Valores del Atributo Hora
00h00, 01h00, 02h00, 03h00, 04h00, 05h00, 06h00, 07h00, 08h00, 09h00, 10h00, 11h00, 12h00, 13h00, 14h00, 15h00, 16h00, 17h00, 18h00, 19h00, 19h30, 20h00, 21h00, 22h00, 23h00, 24h00

Tabla 7 Valores del atributo Hora

Fuente: Los Autores

Los valores de este atributo están considerados de esta manera puesto que en el CENACE, se registra la demanda nacional de energía cada hora y a las 19h30 por considerarse que en esa hora existe la mayor demanda de energía durante el día.

Atributo Mes

Este atributo se refiere al número que corresponde a cada mes en un año. El Atributo Mes es de tipo Opción, se discretizó según lo mencionado en discretización de variables, los valores que tiene el atributo Mes son los siguientes:

Valor del Atributo Mes	Descripción
1	Mes de Enero
2	Mes de Febrero
3	Mes de Marzo
4	Mes de Abril
5	Mes de Mayo
6	Mes de Junio
7	Mes de Julio
8	Mes de Agosto
9	Mes de Septiembre
10	Mes de Octubre

11	Mes de Noviembre
12	Mes de Diciembre

Tabla 8 Valores del atributo Mes

Fuente: Los Autores

Atributo Anio

Este atributo se refiere al número de año y es de tipo numérico, pues puede tomar cualquier valor de cualquier año. Consta de tres valores puesto que los datos históricos proporcionados cubren estos años. Al corresponder al tipo numérico puede tomar cualquier valor de cualquier año.

Valor del Atributo Anio	Descripción
2003	Año 2003
2004	Año 2004
2005	Año 2005

Tabla 9 Valores del atributo Anio

Fuente: Los Autores

Atributo DiaLaborable

El atributo DiaLaborable expresa si un día de la semana es un día laborable o no lo es, es decir si un día es un Día Feriado o es un día del fin de Semana, este día será no Laborable. Los valores que corresponden a este atributo son:

Valor del Atributo DiaLaborable	Descripción
Si	Si es un día no feriado de Lunes a Viernes
No	Si es un día en que no se labore normalmente.

Tabla 10 Valores del atributo DiaLaborable

Fuente: Los Autores

Atributo Feriado

Este atributo se refiere al nombre de un feriado nacional. Los valores de este atributo son:

Valor del Atributo Feriado	Descripción	Fecha
Nodf	No es ningún día Feriado	Sin fecha
anio nuevo	Feriado de Año Nuevo	1 Enero
Carnaval	Feriado de Carnaval	Sin fecha

dia trabajo	Feriado del Día del Trabajo	1 Mayo
dia difuntos	Feriado del Día de los Difuntos	2 Noviembre
fiestas cuenca	Feriado de las fiestas de Cuenca	3 Noviembre
fiestas quito	Feriado de las Fiestas de Quito	6 Diciembre
independencia guayaquil	Feriado de la Independencia de Guayaquil	9 Octubre
dia madre	Festividad del Día de la Madre	2do Domingo Mayo
dia independencia	Feriado por el Día de la Independencia	10 Agosto
anio viejo	Feriado de Año Viejo	31 Diciembre
Semana santa	Feriado de Semana Santa. Viernes Santo	Sin fecha
fiestas guayaquil	Feriado de Fiestas de Guayaquil	25 Julio
Navidad	Feriado de Navidad	25 Diciembre

Tabla 11 Valores del atributo Feriado

Fuente: Los Autores

Si el valor del Atributo Feriado contempla cualquiera de las opciones mencionadas a excepción de “nodf”, entonces el valor del atributo DiaLaborable es igual a “no”.

Atributo NombreDia

El atributo NombreDia se refiere al nombre de los Días de la semana y toma los siguientes valores:

Valor del Atributo NombreDia
Lunes
Martes
Miércoles
Jueves
Viernes
Sábado
Domingo

Tabla 12 Valores del atributo NombreDia

Fuente: Los Autores

Atributo Clima

El atributo Clima se refiere a la descripción del clima en el Ecuador dependiendo del mes del año. Los valores que se detallan a continuación del atributo Clima son los siguientes:

Valor del Atributo Clima	Descripción
Seco	Corresponde cuando el mes del año es: Mayo, Junio, Julio, Agosto, Septiembre, Octubre, Noviembre.
Lluvioso	Corresponde cuando el mes del año es: Diciembre, Enero, Febrero, Marzo, Abril.

Tabla 13 Valores del atributo Clima

Fuente: Los Autores

Una vez descritos los atributos, así como también los valores de los mismos, es necesario describir el atributo a predecir, en este caso lo llamaremos Clase.

Atributo Predicción

El atributo o clase resultado es Predicción. Esta clase es de tipo opción, es decir que el resultado final de la predicción nos dará en forma de Rangos o intervalos, los mismos que pueden ser modificados por el usuario, mediante el uso de la aplicación. Para modificar los Rangos de la clase Predicción se hace referencia al valor de la demanda. Los posibles valores de esta clase están limitados a los valores mínimo y máximo del valor de la demanda.

La sintaxis en la que se presentan los rangos para la predicción, es básicamente la siguiente:

Rango_01, Rango_02, ….. Rango_n
Donde n es el último valor para un rango seleccionado.

Atributo ValorDemanda

Este atributo se refiere al valor de la demanda expresado en MW; es de tipo numérico, por lo que el valor de este atributo depende de la demanda de energía que se registra cada hora en cada día del año.

A los atributos antes mencionados, se los denomina Atributos Base, por considerarlos atributos característicos para la predicción de la Demanda Eléctrica.

Debido a que el modelo acepta variables discretas no se ha realizado una completa discretización en todos los atributos.

Minimización de la redundancia en las variables

Esta propiedad indica que se evite codificar con diferentes valores el mismo significado. Por ejemplo si queremos referirnos al mes de Enero, este solo se debe codificar con el número 1, pero no hay que codificarlo con el número 2 o con otro número u otra letra. Así, como en el ejemplo se procedió a realizar la denominación de los valores de los atributos, de tal manera que éstos no sean redundantes.

Definición inconfusa de los valores

Se debe evitar el codificar con el mismo valor cosas diferentes. Por ejemplo no podemos codificar unas veces el Mes del año como 1,2,3,.... y otras veces como Enero, Febrero, Marzo,...; a su vez los valores del atributo codificado solo deberán tomar el valor que se estableció en la codificación.

Para la creación de la base de datos que se utilizará para la Predicción de la Demanda Eléctrica, se han tomado en cuenta las propiedades antes mencionadas.

Características de la Base de Datos

Para la realización de la base de datos para el caso de estudio se tomaron en cuenta las siguientes características:

En la tabla donde se registran los Atributos Base:

- Existen tantos registros como casos.

- Hay tantas columnas como atributos (atributos independientes + clase resultado).
- En cada celda, hay un solo valor, que puede ser de tipo opción o de tipo numérico.
- Todos los valores de la columna “clase resultado” están llenados apropiadamente.
- De no existir valores en alguna de las columnas, a estos valores se los denomina ‘missing values’. En los valores de los Atributos Base, no existen missing values, puesto que todos los datos entregados por el CENACE constan de un valor.

Depuración de la Base de Datos

En una base de datos hay que tener en cuenta que pueden existir casos en los que podamos encontrarnos con Outliers y Missings, a continuación se describen cada uno de los mismos:

Outliers

A los outliers se les considera como datos atípicos, es decir son observaciones con características diferentes a las demás. Su principal problema radica en que son elementos que pueden no ser representativos de la población pudiendo distorsionar seriamente el comportamiento de los modelos y test estadísticos.

Es muy importante la identificación de los outliers dentro de la población, porque un pequeño número de casos atípicos puede alterar totalmente la distribución de una variable y sus estadísticos. Aún cuando los outliers representan por lo general datos atípicos, también podría darse el caso en que éstos puedan representar un conjunto de la población, es decir pueden ser indicativos de características válidas de un determinado segmento.

La detección de los Outliers, en el caso de variables discretas se puede realizar a través de un grafico que muestre las frecuencias de las variables. Para las variables continuas los valores de las variables que en valor absoluto están a más de:

2.5 desviaciones típicas, si el numero de datos < 80
3 desviaciones típicas, si el numero de datos >= 80

Naturaleza y depuración de los outliers

Una vez realizada la detección del dato atípico se busca el motivo por el que éste aparece, dependiendo la naturaleza se determina la acción concreta para su depuración.

Naturaleza	Tratamiento si hay pocos outliers	Tratamiento si hay muchos outliers
Error de Procesamiento	Eliminación de ejemplos (filas)	Declararlos como missings
Dato Extraordinario	Eliminación de ejemplos	Presencia de varias subpoblaciones Búsqueda de una muestra representativa Replanteamiento problema
Desconocida	Eliminación de ejemplos	Identificación clara de los outliers Replicar el análisis con y sin ellas. Calibrar su influencia. Declararlos como missings

Tabla 14 Tratamiento de Outliers

Fuente: Sánchez-Montañez, Teoría y Aplicaciones a Problemas de Predicción

Missings

Los missings o datos ausentes son valores desconocidos (no introducidos) de los datos. Los datos ausentes son muy habituales en el tratamiento de datos ya que proceden generalmente de la respuesta parcial (por ejemplo en las encuestas) o la mala introducción del valor de las variables. Los missings necesitan ser reemplazados por varias razones:

- El modelo a utilizar no los trata bien o bien rechaza los registros con valores ausentes de forma automática.

- No permiten crear de forma coherente nuevas variables e indicadores de negocio.

Naturaleza y Tratamiento de missings

Una vez detectados los datos ausentes, se debe decidir qué hacer con ellos, la acción a llevar a cabo, depende del motivo por el que se tienen datos ausentes, es decir, la naturaleza y aleatoriedad del mismo. La naturaleza de la existencia de los missings pueden deberse a que se trate de una fenómeno no aleatorio donde los valores faltantes puedan considerarse significativos o no, así también por que se trate de un fenómeno aleatorio donde se tenga valores incompletos.

Naturaleza	Tratamientos si hay pocos missings	Tratamiento si hay muchos missings
Fenómeno no aleatorio	Eliminación de ejemplos	Eliminar variable si es posible. Utilizar modelos que los traten. Para el caso de datos sin sentido, imputarles un valor "diseñado" fuera de rango del valor de la variable.
Fenómeno aleatorio	Eliminación de ejemplos. Ignorar si el modelo a utilizar es robusto o datos faltantes . Reemplazar el valor (métodos de imputación)	Eliminar variable si es posible

Tabla 15 Tratamiento de datos ausentes (Missings)

Fuente: Sánchez-Montañez, Teoría y Aplicaciones a Problemas de Predicción

Métodos De Imputación De Valores Missing

Dentro del tratamiento de valores ausentes tenemos los siguientes métodos de imputación aplicable a los tipos de variable, así:
Si la variable es continua: reemplazar el missing por la media o mediana.
Si variable es discreta: reemplazar por la moda.

Selección y Particionamiento de datos

Dentro del contexto del proceso de selección y particionamiento de datos se deben definir los siguientes conceptos para su entendimiento:

Conjunto de Partida

El conjunto de Partida es aquel que contiene todos los valores de los respectivos atributos característicos.

Conjunto de Entrenamiento o Aprendizaje

El conjunto de Entrenamiento o Aprendizaje es aquel que contiene los valores de los atributos que van hacer utilizados para que el modelo de aprendizaje pueda aprender de éstos.

Conjunto de Validación

El conjunto de Validación es aquel que contiene los valores de los atributos para validar o comprobar, el acierto que tiene el modelo de aprendizaje una vez que haya aprendido con el conjunto de Entrenamiento.

Conjunto de Prueba

El conjunto de Prueba es aquel que contiene nuevos valores de los atributos para hacer uso del modelo de aprendizaje.

Un ejemplo de selección del conjunto de datos podría ser en el que, del mismo conjunto de datos de partida se extrae el conjunto de entrenamiento y el de validación (66% y 33% respectivamente).

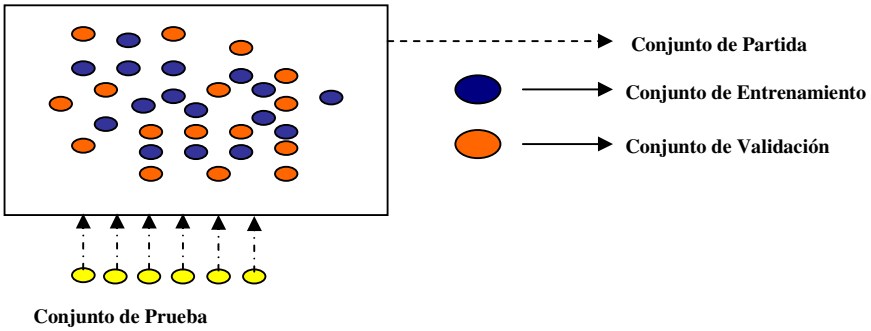


Figura 10 Conjunto de datos

Fuente: Sánchez-Montañez, Teoría y Aplicaciones a Problemas de Predicción

Muestra

Una muestra es un conjunto de casos o individuos procedente de una población.

Población

La población, también llamada universo o colectivo es el conjunto de elementos de referencia sobre el que se realizan las observaciones.

Características de la Muestra

1. La muestra debe ser representativa de la población de estudio. Para cumplir esta característica la inclusión de sujetos en la muestra debe seguir una técnica de muestreo.
2. El número de sujetos que componen la muestra suele ser inferior que el de la población, pero suficientes para que la estimación de los parámetros determinados tenga un nivel de confianza adecuado. Para que el tamaño de la muestra sea idóneo es preciso recurrir a su cálculo.

Ventajas de la elección de una muestra

1. Reducción de costes: Al estudiar un pequeña parte de la población, los gastos de recogida y tratamiento de los datos serán menores que si los obtenemos del total de la población.
2. Rapidez: Al reducir el tiempo de recogida y tratamiento de los datos, se consigue mayor rapidez.
3. Viabilidad: La elección de una muestra permite la realización de estudios que serían imposible hacerlo sobre el total de la población.

Nivel de Confianza

El nivel de confianza es la probabilidad de que el parámetro a estimar se encuentre en el intervalo de confianza.

Análisis de la Muestra

Es necesario tomar en cuenta los siguientes factores para analizar la muestra:

- Necesidad
- Tamaño muestral
- Técnicas de muestreo
- Representatividad de la muestra

NECESIDAD

Es necesaria la selección de una muestra puesto que por lo general el estudio sobre toda la población involucra la inversión de recursos humanos, materiales o económicos, por lo que es importante que la muestra seleccionada sea lo más representativa posible.

TAMAÑO MUESTRAL

El tamaño muestral es el número de sujetos que componen la muestra extraída de una población, necesarios para que los datos obtenidos sean representativos de la población.

Factores a tener en cuenta para el cálculo del tamaño muestral:

Para el cálculo del tamaño muestral se consideran:

El Conjunto de entrenamiento, validación y prueba

El caso en que la muestra es el punto de partida de construcción de un modelo, es necesario tener en cuenta en el tamaño muestral la partición a realizar de los datos.

La Generalización de la muestra a la población

El modelo se construye sobre la muestra, por lo que hay que fijar el porcentaje de confianza con el cual se quiere generalizar, comúnmente en las investigaciones de mercado se busca un 95%.

El Porcentaje de error

El porcentaje de error consiste en elegir una probabilidad de aceptar una hipótesis que sea falsa como si fuera verdadera, o a la inversa (rechazar la hipótesis verdadera por considerarla falsa), comúnmente se aceptan entre el 4% y el 6% como error.(E=0.04 , 0.06).

Fórmula para calcular el tamaño muestral

Tamaño de población desconocido $$n = \frac{Z^2}{4E^2}$$ Fórmula

Tamaño de población conocido $$n = \frac{Z^2 N}{4NE^2 + Z^2}$$ Fórmula

n: tamaño muestra; *Z*: valor normal asociado a la confianza fijada

E: porcentaje de error; *N*: tamaño población

Calculo de Z a partir del nivel de confianza fijado

Si se tiene un nivel de confianza fijado en un 95%.

Utilización de la distribución normal
Determinar el valor Z de la distribución normal que encierra bajo la curva el 95% de la población

Utilizando tablas distribución normal: Z = 1.96, $P(-Z < z < Z) = 0.95$.

Ejemplo:

Muestra representativa para nuestros datos:

N=3000
Error al 5% E=0.05
Nivel de confianza al 95%
Z=1.96
Aplicando la fórmula para N conocido n = tamaño muestral = 341

Muestreo con reemplazo

El muestreo con reemplazo se refiere a que a una observación seleccionada puede ser elegida nuevamente de la población.

Muestreo sin reemplazo

El muestreo sin reemplazo se refiere a que a una observación seleccionada no puede ser elegida nuevamente de la población.

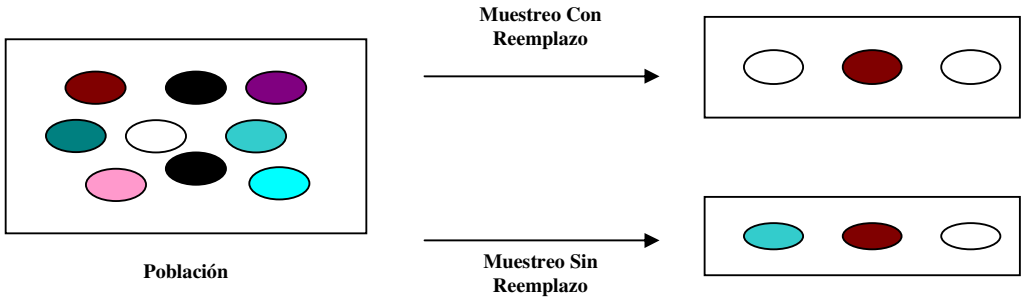


Figura 11 Muestreo con y sin reemplazo

Fuente: Sánchez-Montañez, Teoría y Aplicaciones a Problemas de Predicción

TÉCNICAS DE MUESTREO

La técnica de Muestreo permite realizar la selección de la muestra a partir de una población. Entre las principales técnicas de muestreo tenemos:

Aleatorio Simple

Para ésta técnica de muestreo, cualquier observación tiene la misma probabilidad de ser extraída, ésta selección puede ser con o sin reemplazo (utilización de números aleatorios).

Aleatorio Estratificado

Para ésta técnica es necesario conocer estratos o grupos diferenciados, es decir se debe obtener una muestra con suficientes elementos en todos los estratos o grupos, ya que es necesario saber cuantas observaciones se requieren en cada estrato.

Para generar la muestra es suficiente con realizar sobre cada estrato un muestreo aleatorio simple sin reemplazo para conseguir los elementos que se quiera.

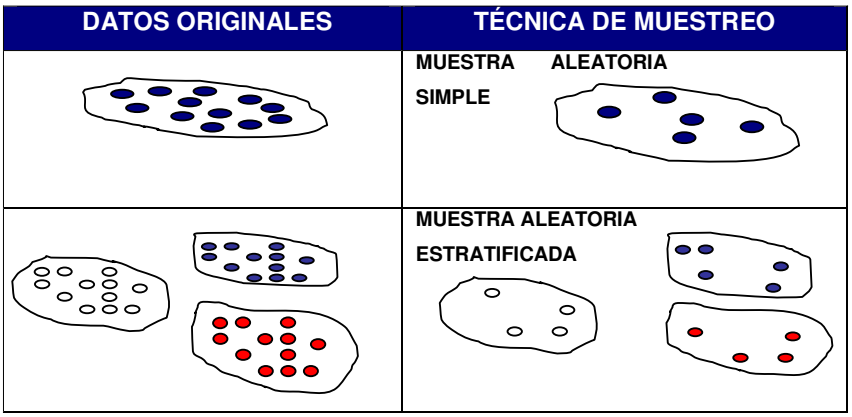


Figura 12 Técnicas de muestreo

Fuente: Sánchez-Montañez, Teoría y Aplicaciones a Problemas de Predicción

Representatividad de la Muestra

Una vez extraída la muestra es necesario medir su representatividad, es decir, si tiene mismas características que la población total. Si la muestra es:

1.-Simple: medir representatividad del total de la muestra

2.-Estratificada: medir representatividad del total y de cada estrato

Medir representatividad de la muestra

La medición de representatividad de la muestra se basa en comparar el comportamiento de las variables entre las observaciones que pertenecen a la muestra y la población total.

CONTRASTE DE HIPÓTESIS

Para realizar el contraste de hipótesis, en que se trata de ver la asociación o relación de los diferentes atributos, la manera de ver esta relación se aplica de la siguiente forma dependiendo si se trata de variables continuas o variables discretas.

Contraste de hipótesis para variables continuas

Para realizar contrastes de igualdad de media y varianza. Ver definiciones en el Anexo de Probabilidad y estadística adjunto en el cd.

Se quiere saber si la variabilidad que se observa en los datos se debe al azar o existen diferencias significativas entre pertenecer o no a la muestra.

Para ello se mide si hay una diferencia en la varianza de la distribución y a posteriormente se mide si hay una diferencia significativa entre la media de la distribución.

Contraste de hipótesis para variables discretas

Con el contraste de hipótesis, se pretende detectar si hay una diferencia significativa en el reparto de una variable categórica en función de si se cogen todas las observaciones o solo las que pertenecen a la muestra. Éste contraste se realiza a través de la prueba del chi-cuadrado.

PRUEBA CHI-CUADRADO

Esta prueba puede utilizarse incluso con datos medibles en una escala nominal. La hipótesis nula de la prueba Chi-cuadrado postula una distribución de probabilidad totalmente especificada como el modelo matemático de la población que ha generado la muestra. Para realizar este contraste se disponen los datos en una tabla de frecuencias. Para cada valor o intervalo de valores se indica la frecuencia absoluta observada o empírica (O_i). A continuación, y suponiendo que la hipótesis nula es cierta, se calculan para cada valor o intervalo de valores la frecuencia absoluta que cabría esperar o frecuencia esperada ($E_i = n * p_i$, donde n es el tamaño de la muestra y p_i la probabilidad del i-ésimo valor o intervalo de valores según la hipótesis nula). El estadístico de prueba se basa en las diferencias entre la O_i i y E_i y se define como:

$$x^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Este estadístico tiene una distribución Chi-cuadrado con k-1 grados de libertad si n es suficientemente grande, es decir, si todas las frecuencias esperadas son mayores que 5. En la práctica se tolera un máximo del 20% de frecuencias inferiores a 5. En el Subtema Comportamiento de Datos de éste Subcapítulo se muestra los resultados de la aplicación de ésta prueba.

Muestra no representativa

En el caso de no obtener en el estudio una muestra representativa, es necesario la generación de una nueva muestra y nuevamente medir representatividad de la nueva muestra. Para el respectivo análisis de los datos del caso de estudio, se ha tomado en cuenta las características de los datos, los mismos que se han subdivido en tres categorías, para un mejor análisis y una mejor comprensión del comportamiento de la demanda eléctrica:

- Días Laborables.
- Días Feriados.
- Fines de Semana.

Entendiendo cada uno de la siguiente manera:

Días Laborables: Dentro de este grupo se toman en cuenta todos los días laborables, es decir los días de Lunes a Viernes en los que no ha ocurrido ningún feriado.
Días Feriados: A este grupo pertenecen los días que ha ocurrido cualquier feriado nacional, de los antes descritos, sin importar el día en que haya acontecido.
Fines de Semana: En este grupo están todos los días Sábados y Domingos, pero no están los días Sábados y Domingos que hayan sido días feriados.

Para un mejor conocimiento de cómo están dispuestos los datos dentro de estos tres grupos se realiza el análisis de cada atributo respecto a cada categoría:

Días Laborables

Existen 18616 registros de un total de 27196 equivalentes a un 68.45% del total registrado, pertenecientes a los valores que corresponden a los días Laborables; los mismos que pertenecen a las horas en que se registra la demanda.

Días Feriados

Para este grupo se han obtenido 1846 registros de un total de 27196 equivalentes a un 6.79% del total registrado; los mismos que corresponden a los registros de los valores de la demanda para los días feriados.

Fines de Semana

En este grupo se tienen 6734 registros de un total de 27196, equivalentes a 24.76% del total registrado; datos que corresponden a los valores registrados para cada hora en los fines de semana.

COMPORTAMIENTO DE LOS DATOS

Aplicación de la Prueba del Chi Cuadrado

A continuación se muestra la aplicación de la prueba del chi cuadrado para ver la asociación entre los atributos discretos utilizados.

DiaLaborable - NombreDia

	NOMBRE DIA					
DiaLaborable	Lunes	Martes	Miércoles	Jueves	Viernes	Total
Si	3718	3744	3744	3744	3666	18616
No	156	130	156	156	234	832
Total	3874	3874	3900	3900	3900	19448

Tabla 16 Chi Cuadrado NombreDia – DiaLaborable

Fuente: Los Autores

Planteamiento de hipótesis:

Ho= No hay asociación entre las variables
H1= Si hay asociación entre las variables

Valor obtenido del X^2 **=** 1.475
Grados de libertad = (col-1)*(fil-1) = 4

Probabilidad de un valor superior para grado de libertad 4 = 9.49

Valor obtenido del X^2 < Probabilidad de un valor superior para grado de libertad 4, **entonces se acepta Ho.**

La aplicación de la prueba del chi-cuadrado para ver la asociación entre los atributos discretos se encuentra en el Anexo Chi-Cuadrado adjunto en el cd, aplicación de la prueba del chi cuadrado. Se muestra una tabla resumen de los resultados obtenidos:

Atributo 1	Atributo 2	X²	Grado Libertad	Distribución chi cuadrado	Asociación
Dia laborable	NombreDia	1.475	4	9.49	Independiente
Fin de semana	DiaLaborable		1	3.84	No aplica
Dia laborable	Feriado	165.27	1	3.84	Dependiente
Mes	Clima	1030.23	11	19.68	Dependiente
Hora	Predicción	38342.04	125	152.093	Dependiente

Tabla 17 Chi Cuadrado Resultados obtenidos

Fuente: Los Autores

En la asociación de los atributos Fin de Semana y DiaLaborable se determina no aplica puesto que no cumplen por lo recomendado por la prueba.
Los otros resultados se han obtenido de acuerdo al contraste de hipótesis de la prueba.

3.2 METODOLOGÍA DE DESARROLLO

Para la realización de una aplicación de software sea ésta de diferente alcance desde la más pequeña hasta la más completa aplicación de sistemas integrados, es necesario realizar una serie de etapas ordenadas con la intención de lograr la obtención de un producto de software con buena calidad. Durante el proceso de software, las necesidades de un usuario son traducidas en requerimientos de software y estos requerimientos después son

transformados en el diseño implementado en el código para certificar su uso operativo. Para el desarrollo de la aplicación se utiliza, el PUD proceso Unificado de Desarrollo y como metodología el UML Lenguaje Unificado de Modelado, se realiza una generalización de conceptos que se describen en el ANEXO PUD incluido en el cd.

3.3 DESARROLLO DE LA APLICACIÓN PARA LA PREDICCIÓN DE DEMANDA ELÉCTRICA

ANALISIS DEL SISTEMA

ESPECIFICACIÓN DE REQUERIMIENTOS DE LA APLICACIÓN

Fase de Inicio

Introducción

En esta actividad del proceso, se describen los requerimientos que debe cumplir la aplicación, se tiene como referencia los Diagramas de Caso de Uso. Además se exponen diferentes factores a ser tomados en cuenta, como el propósito, alcance, definiciones, referencias.

Propósito

El propósito es permitir mediante una aplicación predecir la demanda eléctrica utilizando un modelo de aprendizaje.

Alcance

Se permitirá a los subsistemas de la aplicación que estén directamente relacionados con el Diagrama de Casos de Uso que se presenten en la respectiva etapa del proceso, tomando en cuenta el propósito de la realización de la aplicación.

Definiciones, Acrónimos y Abreviaciones

Las principales definiciones, acrónimos y abreviaciones utilizadas en el desarrollo de éste proceso se encuentra en el Glosario del Documento.

Referencias

Las referencias básicamente están establecidas en los Diagramas de Caso de Uso y el Glosario del Documento.

DESCRIPCIÓN GENERAL

- La Corporación CENACE es la entidad que administra con seguridad, calidad, y economía, tanto el funcionamiento técnico del Sistema Nacional Interconectado e interconexiones internacionales, como el aspecto comercial del Mercado Eléctrico Mayorista (MEM), incluyendo las transacciones internacionales de electricidad, cumpliendo la normativa para satisfacer a sus clientes.

- Con estos antecedentes se crea la necesidad de tener una alternativa para la predicción de energía eléctrica, usando un método cualitativo para éste efecto.

- La aplicación permite pronosticar la demanda eléctrica en función de un modelo de aprendizaje, da paso a la selección de uno o varios atributos cualitativos que de una u otra forma influyen en la demanda eléctrica en general, se preparan los respectivos archivos necesarios para la utilización del modelo de aprendizaje, también se puede ver gráficamente y cuantitativamente los valores de los diferentes atributos que intervienen, para tener una idea mas clara del comportamiento de los datos, finalmente se interpretan los resultados obtenidos comparándolos con el modelo de predicción actualmente usuado.

Requerimientos Específicos

Funcionalidad

Los requerimientos funcionales del Sistema son:

- Validar Usuario.- Verifica la existencia del usuario ingresado, a través de su nombre de usuario y su clave de ingreso.

- Predecir Demanda Eléctrica.- Utiliza los datos históricos para una selección de atributos influyentes en la demanda eléctrica, prepara los archivos para el uso del método de aprendizaje, toma los datos obtenidos del modelo y expresa los resultados obtenidos comparándolos con el modelo usado actualmente.

- Preparar datos.- Generación de archivos con extensiones .names, .data, importantes para poder ser utilizados por el programa del modelo de aprendizaje. Ésta generación está en función de algunos criterios de segmentación que define el usuario.

- Utilizar See5.- Permite manejar el modelo de aprendizaje, incluido en el Programa See5, el cual permite utilizar los archivos generados y necesarios para su utilización.

- Gestionar Pronóstico.- Registro, actualización y eliminación de pronósticos con sus respectivos atributos, para poder contestar con información histórica dichos pronósticos y atributos. La información registrada puede ser gestionada por el Administrador de la aplicación.

Confiabilidad

- Disponibilidad.- La aplicación debe estar disponible cuando el Administrador lo requiera.

Desempeño

- Capacidad.- El número de usuarios que utilizan la aplicación no es mayor por lo que no se controla concurrencia.

Soporte

- Se utilizarán estándares de programación y de base de datos, para facilitar el mantenimiento de la aplicación.

Restricciones de Diseño

Utilización del motor de base de datos Microsoft SQL Server 2000.

3.3.1 MODELO DE CASOS DE USO

Definición de Actores

DESCRIPCIÓN DE ACTORES	
Actores	**Descripción**
Administrador	Este actor no tendrá restricciones para el uso de la aplicación. Tiene acceso a toda la información, así como la administración de usuarios, realización de la predicción de la demanda y gestión de pronósticos.

Tabla 18 Descripción de Actores

Fuente: Los Autores

Diagrama de Casos Usos

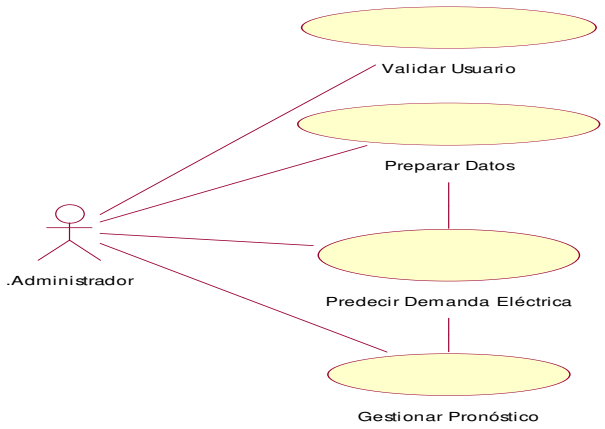


Figura 13 Diagrama de casos de uso

Fuente: Los Autores

ESPECIFICACIÓN DEL DIAGRAMA DE CASOS DE USO

DESCRIPCION DE CASOS DE USO	
Caso de Uso	**Descripción**
Validar Usuario	El sistema valida al usuario a través de un login y password, verifica su existencia. Si el usuario es un Administrador tiene acceso al manejo de todo el sistema.
Preparar datos	El sistema realiza la generación de archivos necesarios para que sean utilizados por el modelo de

	aprendizaje. Además posibilita realizar una segmentación de los valores de los atributos.
Predecir Demanda Eléctrica	El sistema permite predecir la demanda eléctrica mediante el uso de datos históricos que se cargaron al sistema, a su vez hace uso de los archivos generados en el caso anterior para luego obtener los resultados del modelo de aprendizaje. También se tiene una visualización de datos segmentados cuantitativamente y gráficamente.
Utilizar See5	El sistema permite ejecutar el método de aprendizaje para la utilización de los archivos generados en el caso anterior.
Gestionar Pronóstico	El sistema permite al Administrador registrar la información de los pronósticos y su respectiva definición de atributos para la carga de valores de los mismos en caso de que se trate de usar un nuevo pronóstico.

Especificación de Casos de Uso

DESCRIPCIÓN DEL CASO DE USO VALIDAR USUARIO	
Caso de Uso	Validar Usuario
Descripción	El Administrador puede verificar la existencia del usuario en el sistema y permite validar su password.
Actores	Administrador
Pasos	1. El Administrador digita su login y password. 2. El sistema verifica la existencia del usuario. 3. El sistema valida el password ingresado y busca el nombre del servidor, para establecer las seguridades y permisos del sistema. 4. El Administrador ingresa al menú principal del sistema.
Variaciones	En 3: 1. Si no existe el usuario el sistema lo notifica. 2. El sistema solicita nuevamente el ingreso del login y

	password. 3. Si no esta correcto el nombre del servidor el sistema lo notifica. En el caso de que el usuario no pueda ingresar se debe solicitar al Administrador se lo registre.

DESCRIPCIÓN DEL CASO DE USO PREPARAR DATOS	
Caso de Uso	Preparar datos
Descripción	El Administrador puede generar los archivos utilizados por el modelo de aprendizaje.
Actores	Administrador
Pasos	1. Previamente una vez que se ha utilizado el Caso de uso Validar Usuario. 2. El Administrador selecciona el origen de los datos, estos son los atributos proporcionados por defecto y alguno creado por el Administrador. 3. Si el Administrador selecciona los atributos proporcionados por defecto, puede realizar la segmentación o diferenciación de los valores de los atributos. 4. El Administrador puede elegir los atributos que van a intervenir en el proceso de predicción. 5. El Administrador solicita al Sistema iniciar el proceso generar archivo .names 6. El Administrador guarda el archivo .names. 7. El sistema genera el archivo.data y se guarda en el mismo directorio del archivo .names
Variaciones	En 2: 1. Si el Administrador selecciona la opción de un pronóstico diferente al dado por defecto, éste no puede segmentar los datos seleccionados. En 7:

	1. Se puede solamente dejar generados los archivos .names y data para ser usados en el caso de uso Predecir Demanda Eléctrica.

DESCRIPCIÓN DEL CASO DE USO PREDECIR DEMANDA ELÉCTRICA	
Caso de Uso	Predecir Demanda Eléctrica
Descripción	El Administrador puede pronosticar la demanda eléctrica en función de los datos históricos.
Actores	Administrador
Pasos	1. El Administrador selecciona el pronóstico. 2. El sistema carga los archivos .names y . data generados en el caso de uso anterior. 3. El Administrador solicita al Sistema iniciar el proceso de preparar atributos. 4. El Administrador solicita utilizar el caso de uso Utilizar See5 que contiene el modelo de aprendizaje. 5. El Administrador solicita al sistema obtener los resultados. 6. El Administrador ingresa los datos de los atributos para realizar la predicción. 7. El sistema entrega el resultado de la predicción en función de los datos ingresados. 8. El Administrador visualiza los datos obtenidos, tiene la opción de guardar los resultados para futuros evaluaciones.
Variaciones	En 2: 1. Una vez cargado el archivo .names generado en caso de uso anterior, se puede visualizar importante información referente a los valores de los atributos que se seleccionaron tanto cuantitativamente como gráficamente.

DESCRIPCIÓN DEL CASO DE USO UTILIZAR SEE5	
Caso de Uso	Utilizar See5
Descripción	El Administrador hace uso del programa que contiene el modelo de aprendizaje.
Actores	Administrador
Pasos	1. El Administrador solicita al sistema ejecutar el programa del modelo de aprendizaje. 2. El Administrador ubica el archivo .data generado en el caso de uso Preparar Datos. 3. El Administrador selecciona los modos de construcción del modelo de aprendizaje. 4. El programa que utiliza el modelo de aprendizaje See5, entrega los resultados.
Variaciones	En 2: 1. De no existir un archivo .data el programa See5 no podrá realizar la respectiva elaboración del árbol.

DESCRIPCIÓN DEL CASO DE USO GESTIONAR PRONÓSTICO	
Caso de Uso	Gestionar Pronóstico
Descripción	El Administrador puede gestionar pronósticos creación de pronósticos, atributos y opciones.
Actores	Administrador
Pasos	1. El Administrador realiza la gestión de un pronóstico. 2. El Administrador guarda el pronóstico. 3. El Administrador define los atributos. 4. El Administrador guarda lo atributos. 5. El Administrador guarda los valores respectivos para el pronóstico y para los atributos. 6. Se utiliza El caso de Uso Preparar Datos.
Variaciones	Ninguna

Fase de Elaboración

En ésta fase se trata de conseguir una mayor especificación de los requerimientos que tiene la aplicación a través de la identificación de actores y Casos de uso que pueden aparecer en el contexto de la realización del trabajo.

La aplicación en desarrollo no tiene mucha extensión ya que se ha dispuesto la intervención de un solo actor, Administrador y de los casos de uso respectivos descritos anteriormente. Con esto se contempla tener un modelo de casos de uso bien estructurado que permita tener varias iteraciones hasta lograr el propósito planteado. Hay que tomar en cuenta que en ésta fase se realiza una depuración del modelo de caso de uso habiéndose realizado las modificaciones necesarias y convenientes.

Fase de Construcción

En ésta etapa se identifica a los actores y Casos de uso que no fueron identificados en las dos etapas anteriores, es necesario tener en cuenta todos los factores influyentes en la aplicación para tomar la decisión de un nuevo actor o un nuevo caso de uso o la modificación de alguno de ellos para no caer el la redundancia de nombre o actividades. En el caso de la aplicación en desarrollo, la recopilación de requisitos que corresponden a esta etapa no ha dado como resultados la identificación de nuevos actores y tampoco de nuevos casos de uso.

3.3.2 ANÁLISIS DE LOS REQUERIMIENTOS

Fase de Inicio

En ésta fase se clasifica los casos de uso encontrados de tal manera que se pueda construir un modelo inicial de análisis.

En el caso de la aplicación a desarrollarse, mediante la especificación de los requerimientos del Caso de uso se realiza un prototipo de análisis.

De ahí que del caso de uso encontrado en la fase inicial del flujo de trabajo, se obtienen las siguientes realizaciones de caso de uso:

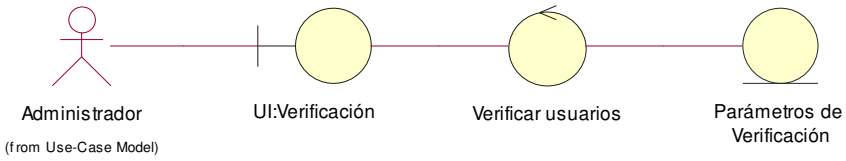


Figura 14 Diagrama de Clases de una realización del Caso de Uso: Validar Ingreso

Fuente: Los Autores

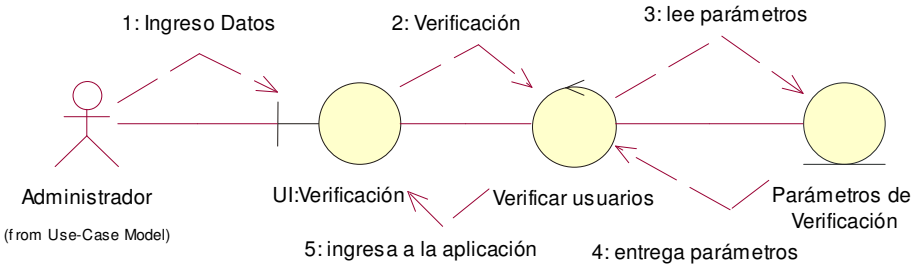


Figura 15 Diagrama de Colaboración de una realización del Caso de Uso: Validar Ingreso

Fuente: Los Autores

FLUJO DE SUCESOS – DIAGRAMA DE COLABORACIÓN DE LA REALIZACIÓN DEL CASO DE USO VALIDAR INGRESO
El Administrador escribe un login, password y nombre del servidor a través de IU Verificación quien pide al objeto Verificar Usuario se encargue de validar (la validación consiste en verificar que un Administrador este registrado para acceder a la aplicación), si lo escrito cumple con las normas se puede ingresar a la aplicación.

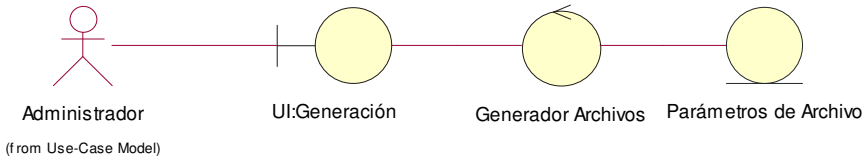


Figura 16 Diagrama de Clases de una realización del Caso de Uso: Preparar datos

Fuente: Los Autores

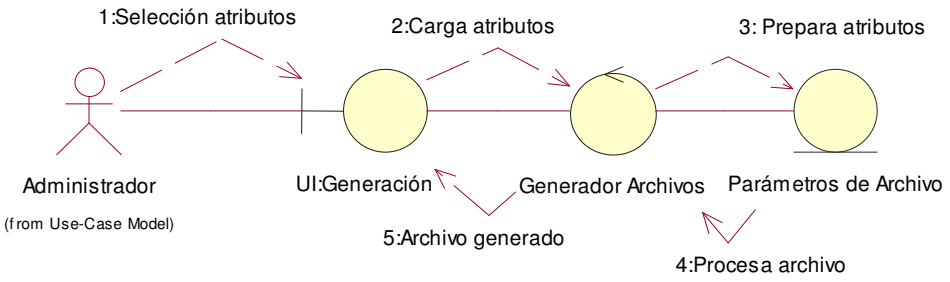


Figura 17 Diagrama de Colaboración de una realización del Caso de Uso: Preparar datos

Fuente: Los Autores

FLUJO DE SUCESOS – DIAGRAMA DE COLABORACIÓN DE LA REALIZACIÓN DEL CASO DE USO PREPARAR DATOS
El Administrador selecciona a través de la la IU Generación la selección de los atributos mostrados para la predicción, esto se lleva a cabo por medio de la colaboración del objeto IU Predicción, los atributos seleccionados son cargados nuevamente, se procesa los atributos en el archivo, para finalmente mostrar el archivo final para ser guardado.

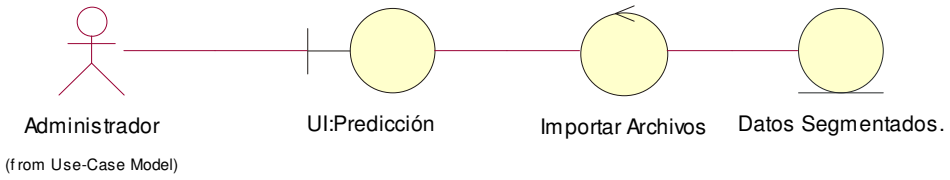


Figura 18 Diagrama de Clases de una realización del Caso de Uso: Predecir Demanda Eléctrica

Fuente: Los Autores

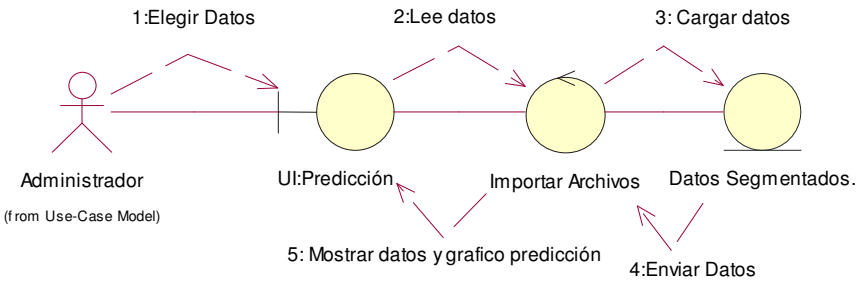


Figura 19 Diagrama de Colaboración de una realización del Caso de Uso: Predecir Demanda Eléctrica

Fuente: Los Autores

FLUJO DE SUCESOS – DIAGRAMA DE COLABORACIÓN DE LA REALIZACIÓN DEL CASO DE USO PREDECIR DEMANDA ELÉCTRICA
El Administrador selecciona los datos utiliza el objeto UI Predicción para luego realizar la lectura de los datos, realiza la importación con los datos necesarios para obtener los atributos a ser analizados, se cargan los datos y luego son mostrados para su preparación o a su vez se presentan los resultados de la predicción.

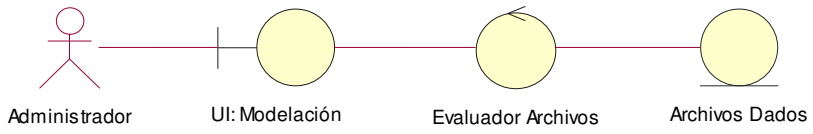


Figura 20 Diagrama de Clases de una realización del Caso de Uso: Utilizar See5

Fuente: Los Autores

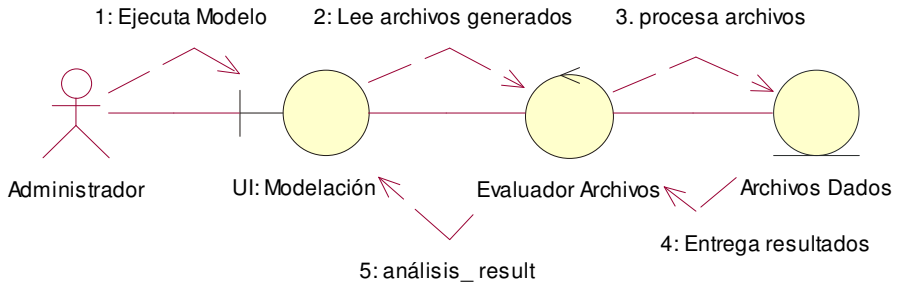


Figura 21 Diagrama de Colaboración de una realización del Caso de Uso: Utilizar See5

Fuente: Los Autores

FLUJO DE SUCESOS – DIAGRAMA DE COLABORACIÓN DE LA REALIZACIÓN DEL CASO DE USO UTILIZAR SEE5
El Administrador utiliza el objeto IU Modelación para ejecutar el programa que contiene See5, para ello se debe identificar los archivos generados mediante el caso de uso Preparar Datos. Se ubica el directorio de almacenamiento de los archivos, mediante el programa See5 seleccionamos clasificadores, el algoritmo procesa los archivos, y son entregados los resultados que pasan a ser analizados en el caso de uso predecir demanda.

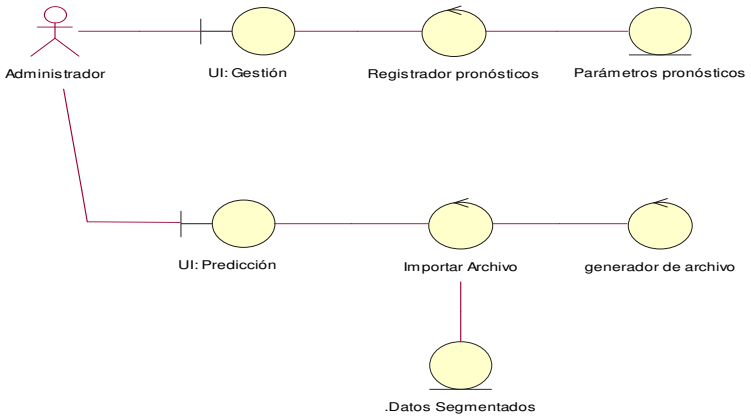


Figura 22 Diagrama de Clases de una realización del Caso de Uso: Gestionar Pronóstico

Fuente: Los Autores

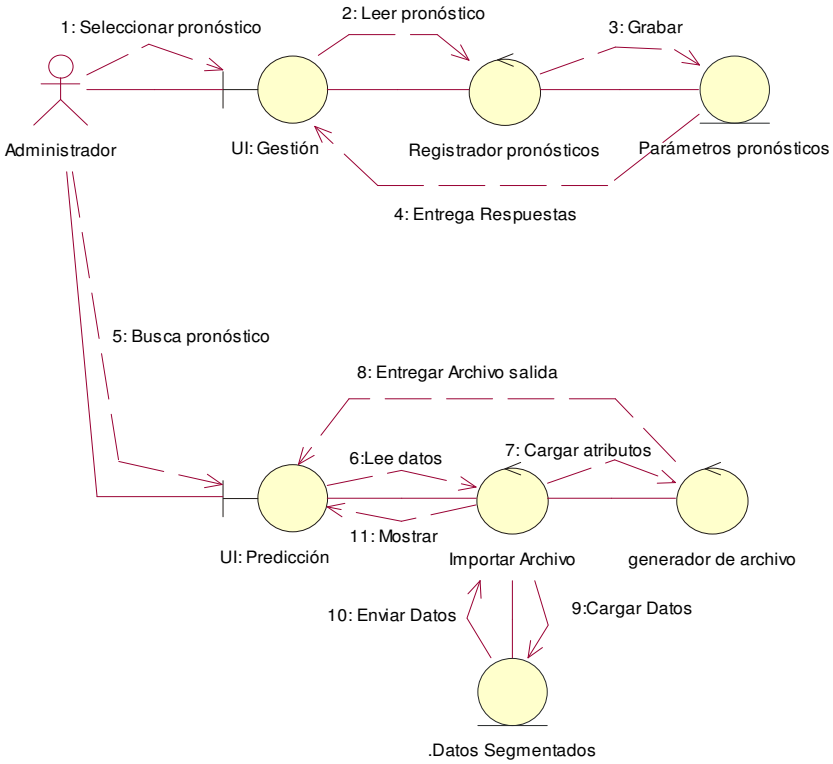


Figura 23 Diagrama de Colaboración de una realización del Caso de Uso: Gestionar Pronóstico

Fuente: Los Autores

FLUJO DE SUCESOS – DIAGRAMA DE COLABORACIÓN DE LA REALIZACIÓN DEL CASO DE USO GESTIONAR PRONÓSTICO
El Administrador utiliza el objeto IU Gestión registrar el pronóstico con sus respectivos atributos para ello antes de este paso se validar el ingreso del usuario mediante el caso de uso Validar Usuario luego Después de esto el

objeto IU Gestión le permite acceder a la administración de los pronósticos el mismo que será comparado con datos en caso que se tengan previamente preparados luego de esto tanto como los atributos como los datos serán guardados, luego pasan a ser utilizados por el caso de Uso predicción demanda y preparar datos.

Fase de Elaboración

Ésta fase define de manera más concreta aquellos casos de uso que no fueron tomados en cuenta para la aplicación en desarrollo. El modelo de análisis queda finalizado con la descripción de los casos de uso que no tienen tanta influencia en lo que respecta al negocio. Para el caso de la aplicación en realización no existen otros casos de usos que se deban analizar y describir.

Fase de Construcción

Con el objetivo de permitir observar el modelo en agrupamiento más simples, ésta etapa permite encontrar paquetes de análisis ya que así, se logra separar los elementos que se detallaron anteriormente. Al identificar los paquetes, éstos deben estar basados en requisitos funcionales y en el dominio del problema, además de establecer la mayor parte de casos de uso que se encuentren relacionados directamente.

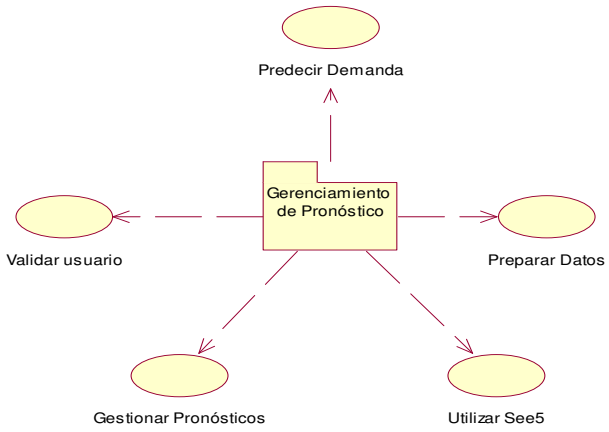


Figura 24 Identificación del Paquete de Análisis Gerenciamiento de Pronóstico a partir de los casos de uso

Fuente: Los Autores

Análisis de la Aplicación

Los datos que se manejará dentro de la aplicación estarán conformados por información, proveniente del repositorio de base de datos. El manejo de esta información se lo realizará por medio de interfaces, las cuales mostrarán la información. La información a presentarse en la aplicación es exacta y actualizada, debido a que todas las operaciones a realizarse están vinculadas a transacciones con el repositorio de datos.

3.3.3 DISEÑO DEL SISTEMA

FASE DE INICIO

Ésta fase permite realizar el diseño de la realización de los casos de uso para lo cual se toma en cuenta sus entradas y salidas de datos, de tal manera que se pueda orientar a una implementación generalizada.El diseño de la aplicación en desarrollo se ejecutará en las capas de presentación, lógica de negocio y servicio de datos. Como se muestra en la figura.

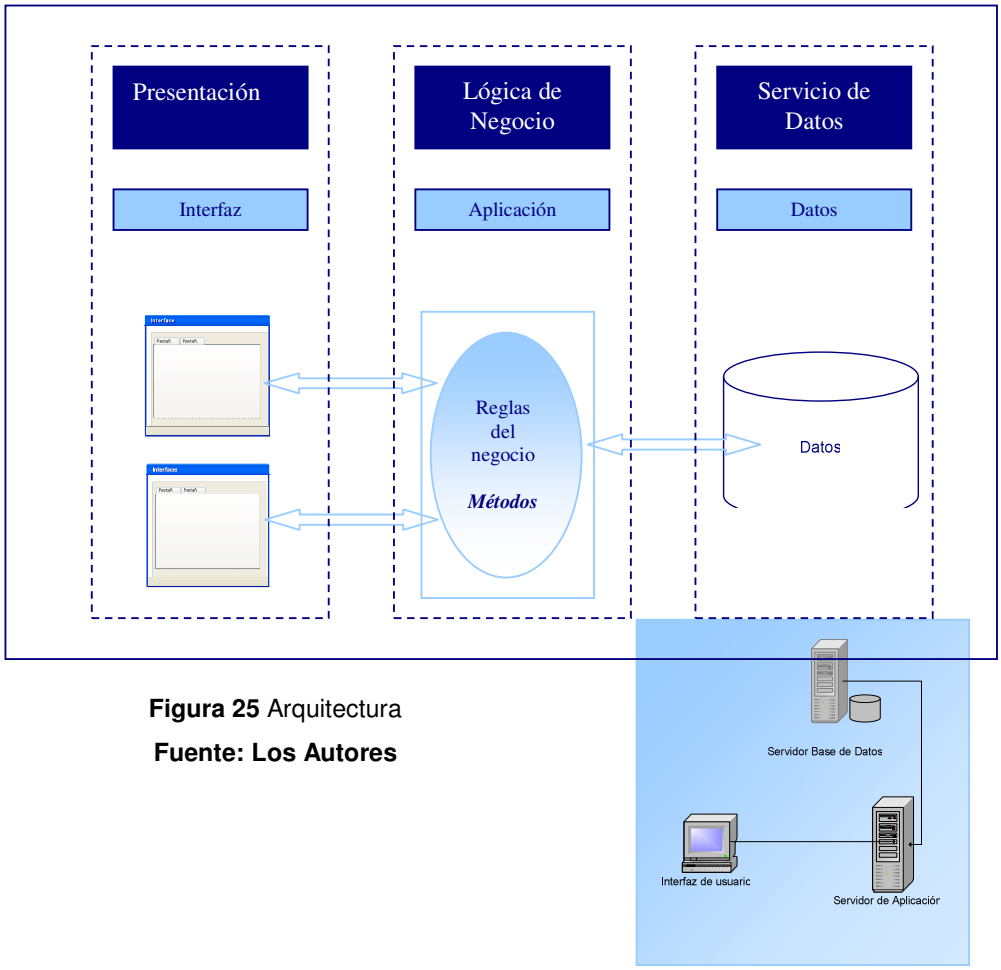


Figura 25 Arquitectura

Fuente: Los Autores

Servidor de Base de Datos

Se encarga del control de la apertura de las bases de datos, así como también se almacenarán las tablas, procedimientos definidos y datos propios de la aplicación.

Servidor de Aplicaciones

En el servidor de aplicaciones se encuentra la Lógica del Negocio, por lo general es el que enlaza la comunicación entre la información almacenada en los servidores de bases de datos y el usuario.

Interfaz de Usuario

Es el medio a través del cual se accede al sistema, de contar con las claves y permisos necesarios para el ingreso al mismo.

3.3.4 FASE DE ELABORACIÓN

Hay que tomar en cuenta las posibles operaciones que brindará cada unos de los subsistemas a utilizarse. Mediante la identificación de los paquetes de análisis que se determinó anteriormente se diseña los subsistemas en función de los casos de uso significativos.

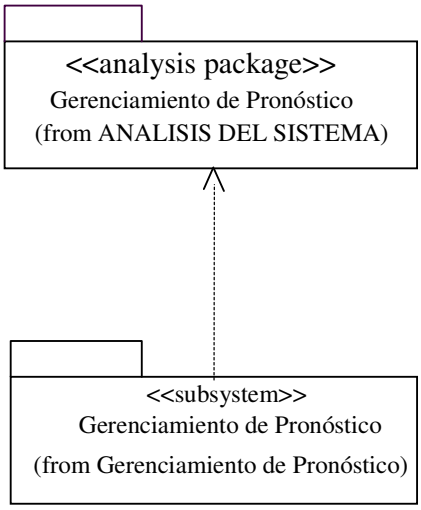


Figura 26 Identificación de subsistemas de diseño a partir de paquetes de análisis

Fuente: Los Autores

3.3.4.1 FASE DE CONSTRUCCIÓN

En esta fase, de acuerdo a los subsistemas encontrados se diseñan los casos de uso, los mismos que son representados por diagramas de colaboración y secuencia, esto nos permiten observar la forma de realización de sus eventos y los ambientes de un caso de uso.

Al finalizar esta fase se busca obtener un modelo de diseño el cual permita obtener un modelo que servirá para desarrollar la aplicación.

A continuación se presenta el diseño de los casos de uso con sus respectivas realizaciones:

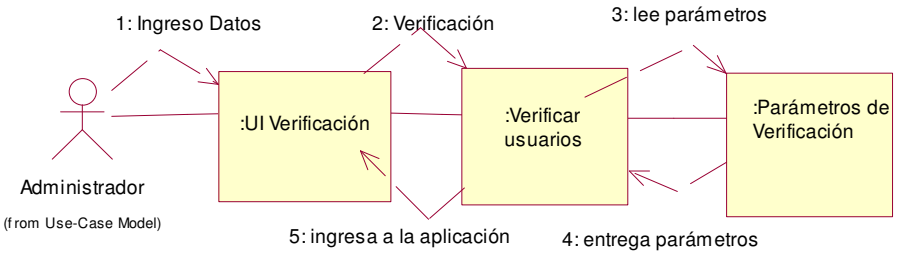


Figura 27 Diagrama de Colaboración de una realización del Caso de Uso: Validar Usuario

Fuente: Los Autores

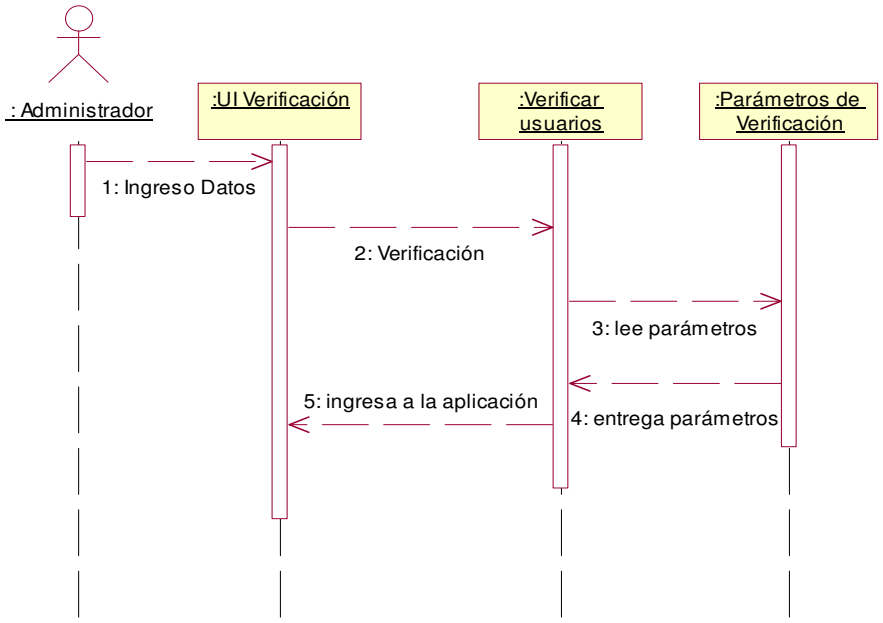


Figura 28 Diagrama de Secuencia de una realización del Caso de Uso: Validar Usuario

Fuente: Los Autores

FLUJO DE SUCESOS – DIAGRAMA SECUENCIA DE LA REALIZACIÓN DEL CASO DE USO VALIDAR USUARIO
El Usuario Administrador por medio del objeto IU Verificación ingresa el login, password, nombre de servidor este interactúa con el objeto Verificar usuario para verificar si el usuario está registrado, si esto es posible el objeto Parámetros de Verificación se comunica con el objeto Verificación para permitir el ingreso a la aplicación.

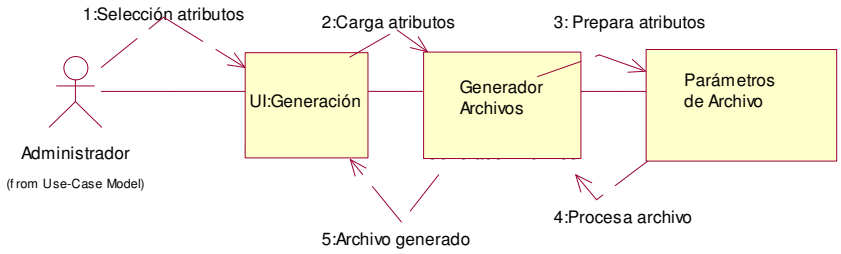


Figura 29 Diagrama de Colaboración de una realización del Caso de Uso: Preparar Datos

Fuente: Los Autores

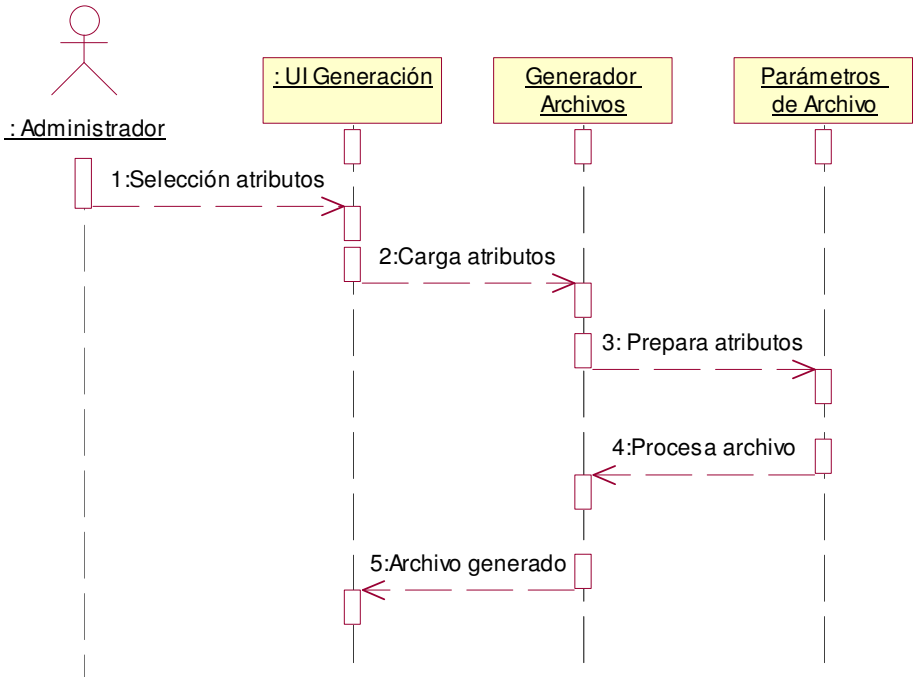


Figura 30 Diagrama de Secuencia de una realización del Caso de Uso: Preparar Datos

Fuente: Los Autores

FLUJO DE SUCESOS – DIAGRAMA SECUENCIA DE LA REALIZACIÓN DEL CASO DE USO PREPARAR DATOS
El Usuario por medio del objeto IU Generación selecciona los atributos de un determinado pronóstico, este interactúa con el objeto Generador de Archivos para verificar si se cumplen las reglas para elaborar los archivos(.names, .data) necesarios para ser procesados, si esto se cumple el objeto Parámetros de Archivos se comunica con el objeto IU Generación con los archivos generados.

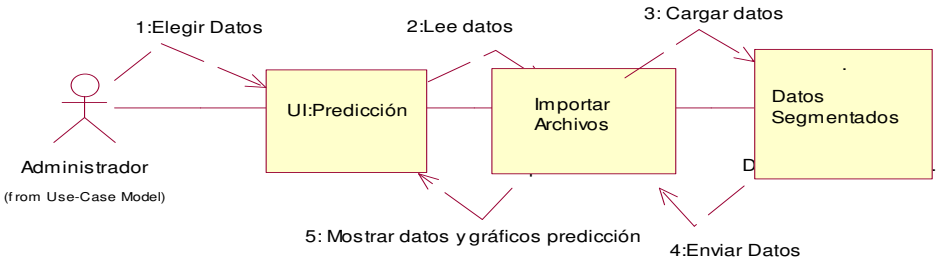


Figura 31 Diagrama de Colaboración de una realización del Caso de Uso: Predecir Demanda

Fuente: Los Autores

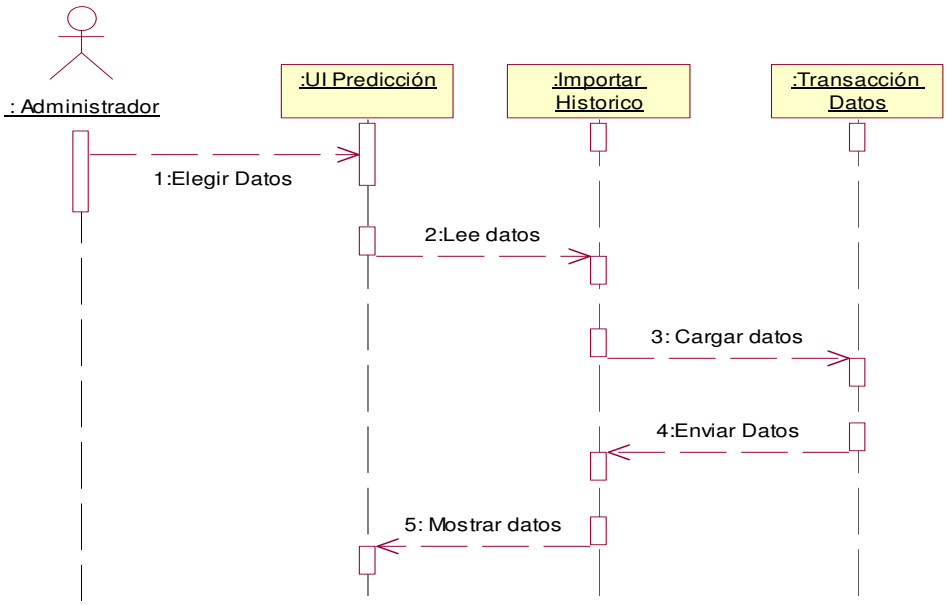


Figura 32 Diagrama de Secuencia de una realización del Caso de Uso: Predecir Demanda

Fuente: Los Autores

FLUJO DE SUCESOS – DIAGRAMA SECUENCIA DE LA REALIZACIÓN DEL CASO DE USO PREDECIR DEMANDA ELÉCTRICA
El Usuario Administrador por medio del objeto IU Predicción selecciona los datos, éste interactúa con el objeto Importar Archivo para cargar la información de la fuente de datos seleccionada si se cumplen las normas de control para cargar los datos el objeto Datos Segmentados se comunica con el objeto IU Predicción para mostrar los datos.

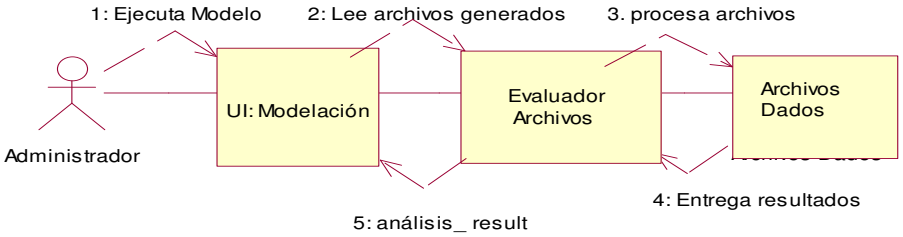


Figura 33 Diagrama de Colaboración de una realización del Caso de Uso: Utilizar See5

Fuente: Los Autores

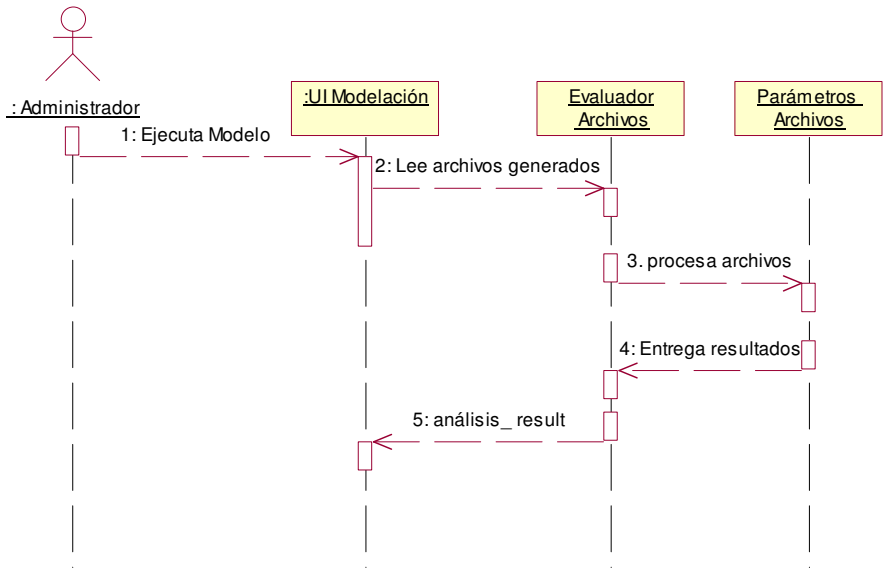


Figura 34 Diagrama de Secuencia de una realización del Caso de Uso: Utilizar See5

Fuente: Los Autores

FLUJO DE SUCESOS – DIAGRAMA SECUENCIA DE LA REALIZACIÓN DEL CASO DE USO UTILIZAR SEE5
El Usuario Administrador por medio del objeto IU Modelación toma los archivos generados, este interactúa con el objeto Evaluador de Archivos para procesar el contenido de los archivos con el cumplimiento de las reglas del modelo de aprendizaje, si esto es posible el objeto Parámetros de Archivo se comunica con el objeto IU modelación entregando el resultado del análisis.

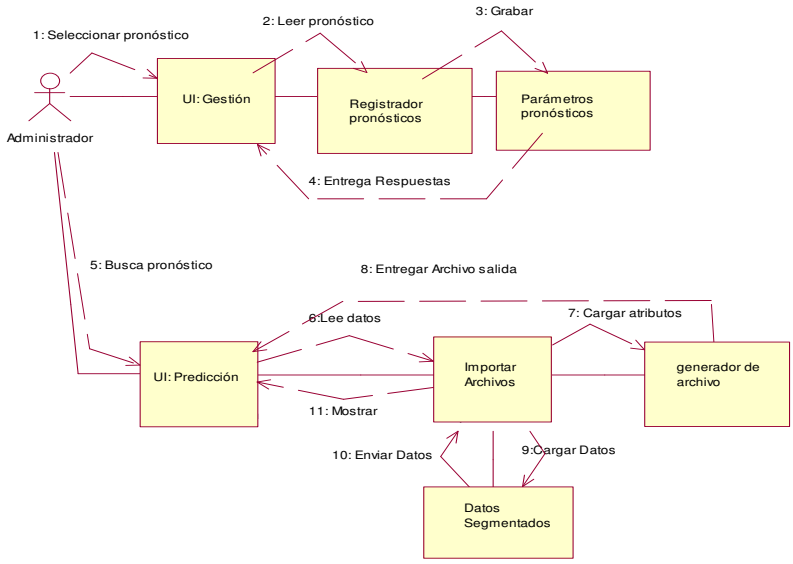


Figura 35 Diagrama de Colaboración de una realización del Caso de Uso: Gest. Pronóstico

Fuente: Los Autores

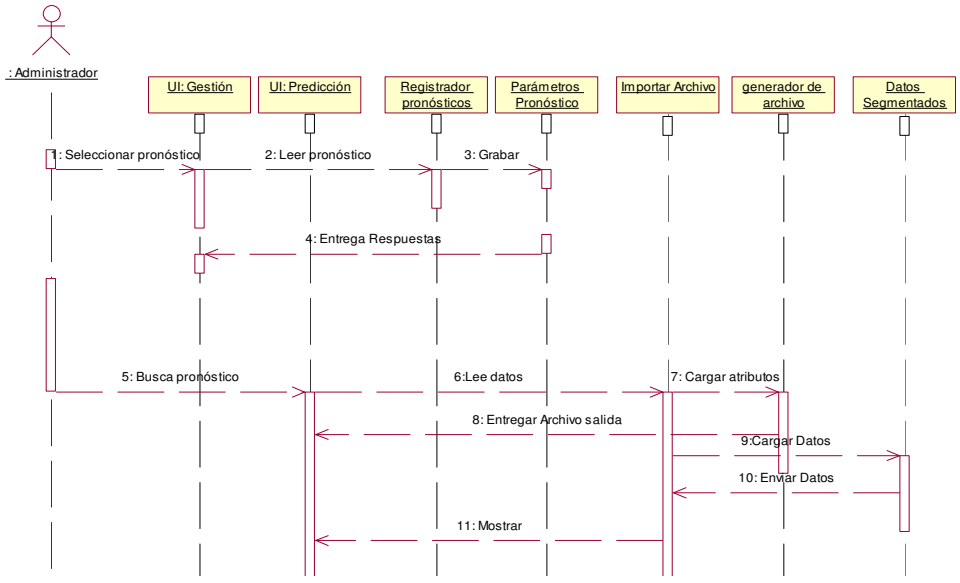


Figura 36 Diagrama de Secuencia de una realización del Caso de Uso: Gestionar Pronóstico

Fuente: Los Autores

FLUJO DE SUCESOS – DIAGRAMA SECUENCIA DE LA REALIZACIÓN DEL CASO DE USO GESTIONAR PRONÓSTICO
El Usuario Administrador por medio del objeto IU Gestión selecciona un pronóstico, este interactúa con el objeto Registrador de pronósticos el cual graba su registro con sus respectivos atributos, si esto es posible el objeto Parámetros Pronóstico se comunica con el objeto UI Gestión para advertir del registro. El usuario Administrador por medio del objeto IU Predicción carga los datos históricos del pronóstico, si esto es posible el objeto generador de archivo entrega archivo de salida para mostrar el resultado de la predicción.

3.3.5 DIAGRAMA DE CLASES

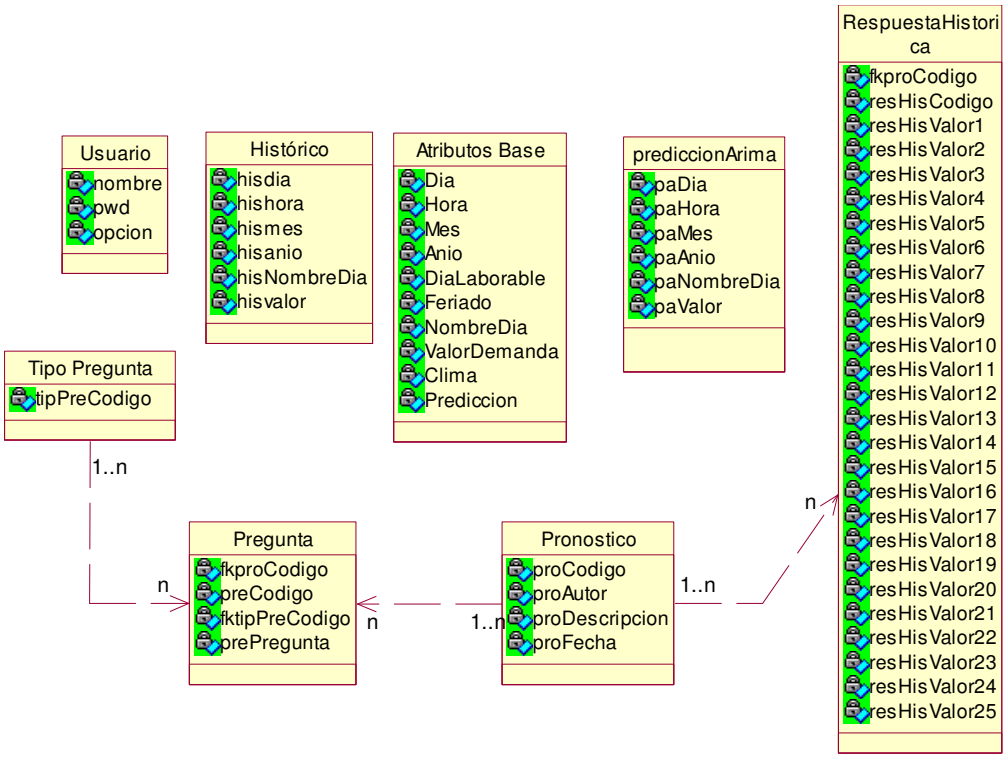


Figura 37 Diagrama de Clases

Fuente: Los Autores

DESCRIPCION DE CLASES	
Clase	**Descripción**
USUARIO	Clase que contiene los datos de los usuarios que van a Administrar la aplicación.

HISTÓRICO	Clase que contiene los datos históricos que permiten describir el gráfico de la demanda en función de algunos valores.
ATRIBUTOS BASE	Clase que contiene los atributos característicos en la demanda eléctrica.
PREDICCION ARIMA	Clase que contiene los atributos y sus respectivos valores de la predicción realizada con el modelo Arima en el período de prueba y predicción.
PRONÓSTICO	Clase que contiene los pronósticos con sus respectivos atributos que son utilizados para la carga de datos.
TIPO DE PREGUNTA	Clase que contiene un atributo que distingue el tipo de pregunta que se va ha realizar, necesarios para identificar la pregunta.
PREGUNTA	Clase que contiene los atributos que se quieran definir para pronósticos referentes a la demanda eléctrica.
RESPUESTA HISTÓRICA	Clase que contiene los valores de los atributos definidos en el pronóstico con sus respectivas preguntas.

Tabla 19 Descripción de Clases

Fuente: Los Autores

3.3.6 DISEÑO DE LA APLICACIÓN

Mapa de Navegación

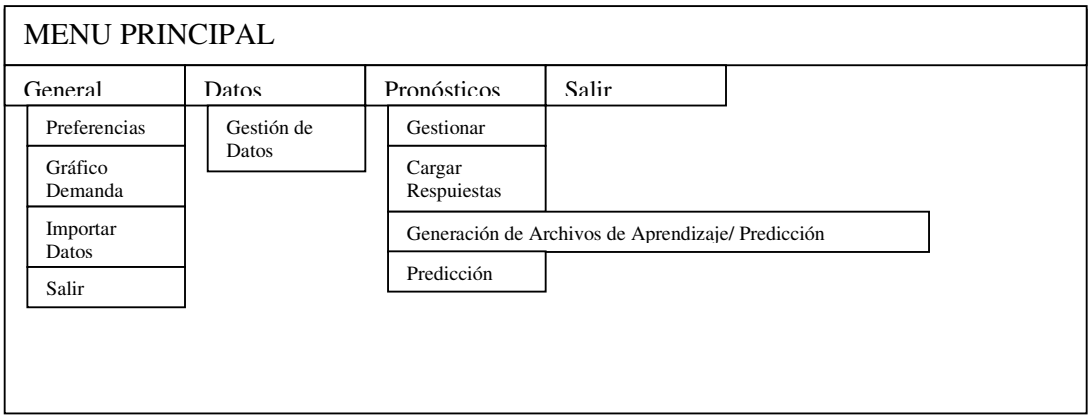


Figura 38 Mapa de Navegación

Fuente: Los Autores

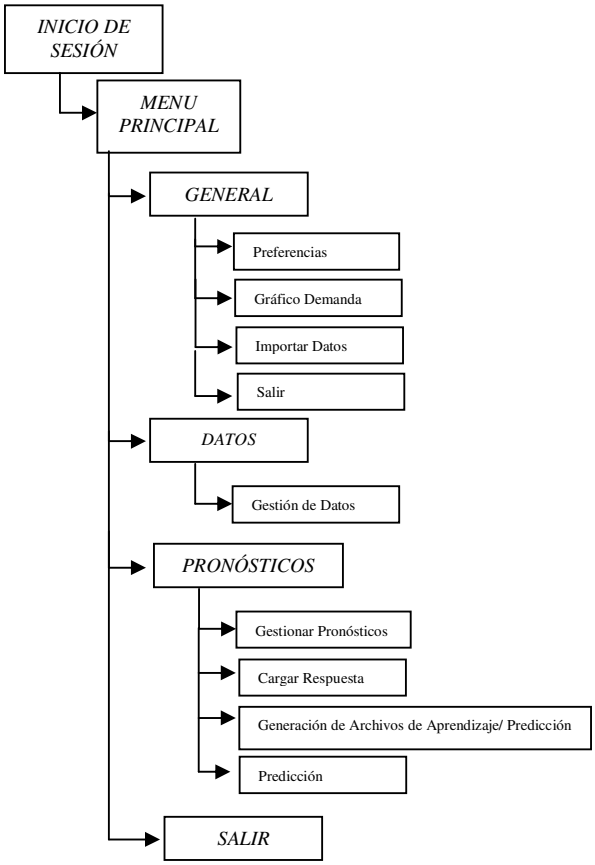


Figura 39 Esquema de Navegación de la Aplicación

Fuente: Los Autores

Diseño Prototipo de Interfaces

3.3.7 ENTRADA Y SALIDA DE DATOS

Las diferentes interfaces que intervendrán en la aplicación tienen el siguiente esquema:

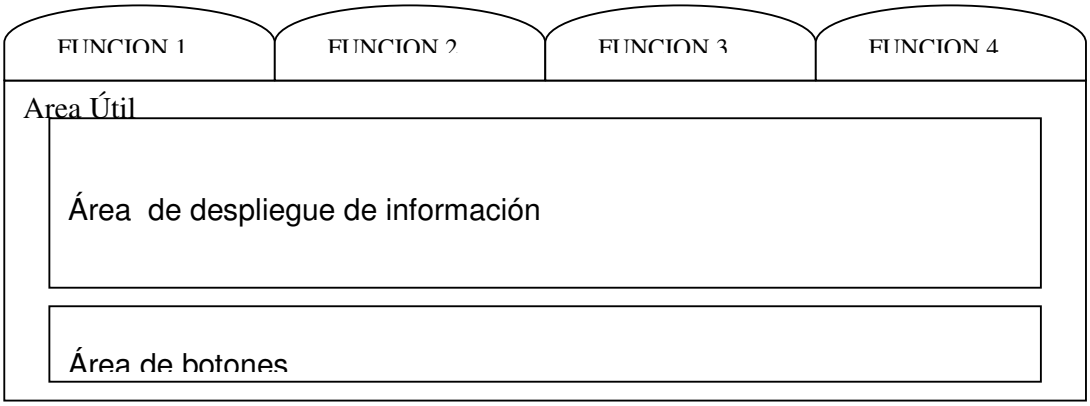


Figura 40 Esquema de las interfaces

Fuente: Los Autores

En éste esquema tenemos el área útil, que es el área de la pantalla donde se va a mostrar el Área de despliegue de información, donde se mostrará la información al usuario y el Área de Botones donde se dará el control sobre la información al usuario por medio de botones. El prototipo de interfaz anterior no consta de un área de despliegue de mensajes ya que los mensaje ya sean estos de error o de información se manejarán con pantallas independientes. Dentro de la aplicación se consideran el prototipo de las siguientes interfaces:

Inicio de Sesión

Esta interfaz permitirá al Actor Administrador, para ello dispondrá de los siguientes elementos:

Área donde se indica el Título de la Interfaz, Campos donde el Usuario ingresará su Nombre de Usuario de Correo Electrónico y su respectiva contraseña, y el nombre del Servidor donde se encuentre la aplicación.

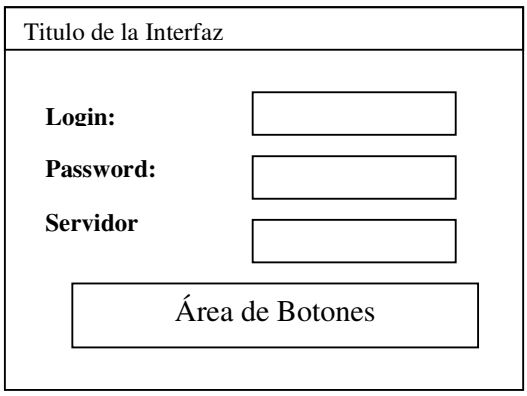


Figura 41 Pantalla de Inicio de Sesión

Fuente: Los Autores

Preferencias

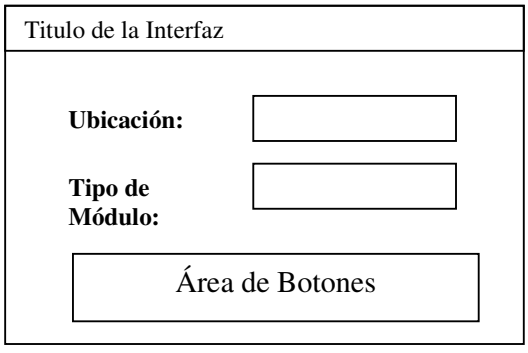


Figura 42 Pantalla de Preferencias

Fuente: Los Autores

Esta interfaz permitirá al Actor Administrador definir la ubicación en la que se encuentra ubicado el ejecutable del programa que utiliza el modelo de aprendizaje, así también permite definir si se va ha trabajar con una versión completa o con una versión demostrativa del programa. Para nuestro estudio se trabaja, con una versión demostrativa del programa See5.

En el área de botones se podrá ejecutar comandos que permitan guardar o cancelar las definiciones.

Gráfica de Demanda

Esta interfaz permitirá al Actor Administrador generar la gráfica de la curva de la demanda eléctrica en un día determinado por el Administrador, así también los principales datos de una demanda.

El área de opción se usará para seleccionar la fecha de la cual se quiera graficar la curva de la demanda así también se podrá elegir el tipo de gráfico que se quiera ver.

En el área de gráfico se podrá visualizar la curva de la demanda eléctrica en las diferentes horas son su respectiva potencia de demanda.

En el Área de datos se muestra la información en detalle de la demanda en el día seleccionado, así también el mínimo, máximo y promedio valor de la demanda en ése día.

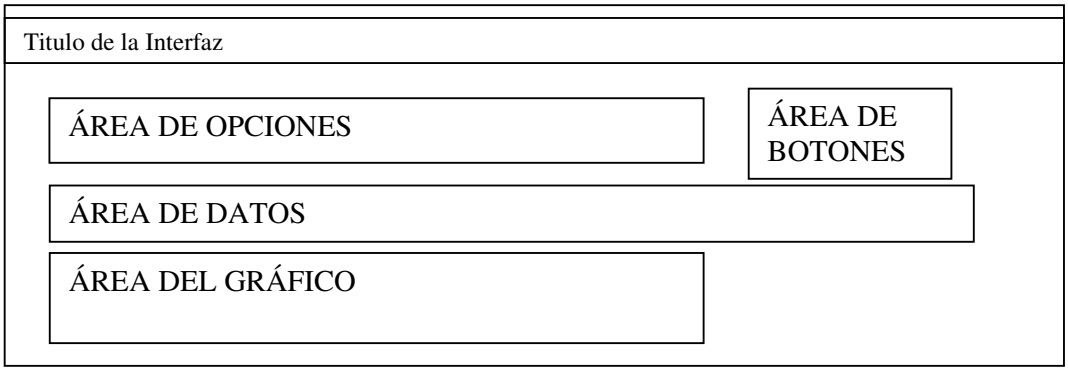


Figura 43 Pantalla del gráfico de la demanda eléctrica

Fuente: Los Autores

Generación de Archivos

Esta interfaz permitirá al Actor Administrador generar los Archivos Names, Data necesarios para ser utilizados por el modelo de aprendizaje.

Las diferentes áreas descritas en la figura como: Área de Datos, Área de Archivo generado y Área de Botones se repiten para los diferentes controles de ficha descritos como Archivos Names, Archivos Data, Ejecutar Predicción, Resultado.En el Área de datos se muestra la información en detalle del pronóstico seleccionado.

En el área de archivo generado se muestra la información de los datos una vez procesados listo para ser guardados.En el área de botones se podrá ejecutar comandos que permitan generar y guardar los archivos.

El área de opción y área de selección, solo aparece en el control de ficha Archivos Names, la primera se usará para seleccionar la fuente de datos, es decir desde una tabla o un pronóstico predefinido. La segunda permite seleccionar los atributos a ser analizados, además de permitir elegir la manera en que se van a cargar los datos para el análisis. El Área de Archivo de salida es utilizando en el control de ficha Ejecutar predicción, una vez que se haga uso del programa See5.

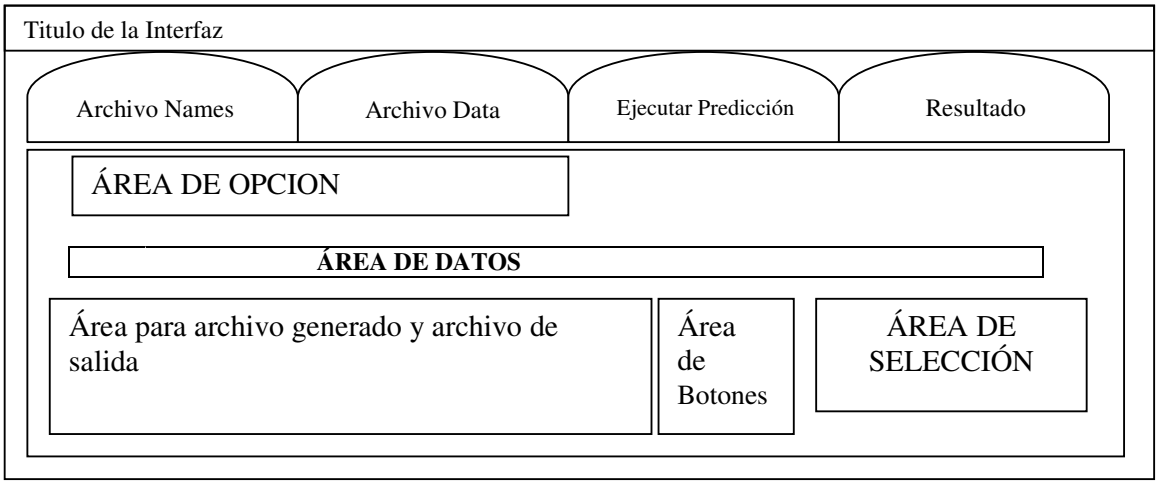


Figura 44 Pantalla de Generación de Archivos

Fuente: Los Autores

Gestión de Pronósticos

Esta interfaz permitirá al Actor Administrador gestionar los pronósticos de manera que pueda tener un pronóstico y al mismo pueda cargarle los datos de información, realizando una verificación de correspondencia entre los datos a cargarse y los atributos del pronóstico.

El área de opción se usará para seleccionar la fuente para la carga de datos, es decir desde un pronóstico predefinido o desde la base datos.

En el Área de datos se muestra la información en detalle del pronóstico seleccionado.

En el área de botones se podrá ejecutar comandos que permitan modificar actualizar loa rangos de la predicción, borrar los datos y salir de la interfaz.

El área de selección permite definir loa rangos en que va ha ser analizada toda la información cargada en el área de datos.

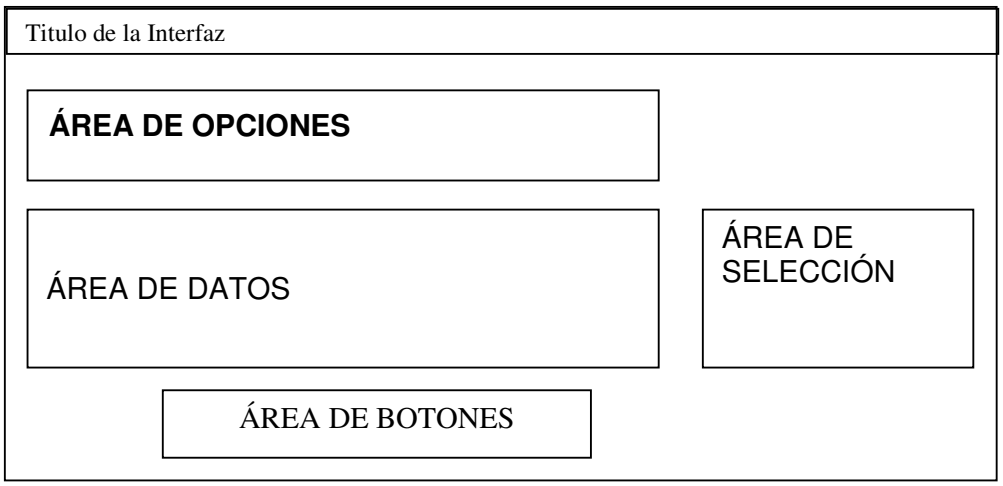


Figura 45 Pantalla de Gestión de Pronósticos
Fuente: Los Autores

3.3.8 IMPLEMENTACIÓN DE LA APLICACIÓN

FASE DE INICIO

En esta fase de inicio de la implementación se toma en cuenta lo que ya se describió sobre la arquitectura candidata en el diseño, así también como el

prototipo de implementación, acerca de la construcción de las interfaces de usuario y la posible navegación a través de la aplicación que ya se realizó en el diseño.

FASE DE ELABORACIÓN

Según ésta fase es posible establecer subsistemas de implementación los cuales son consecuencia de la fase de diseño donde fueron identificados. El modelo de implementación es una secuencia directa del modelo de diseño. Partiendo de esta premisa se define el siguiente subsistema de implementación:

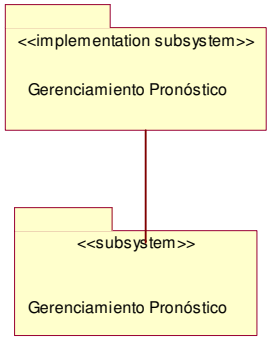


Figura 46 Modelo de Implementación (Subsistema)
Fuente: Los Autores

FASE DE CONSTRUCCIÓN

En ésta fase se obtiene una primera versión operativa inicial de la aplicación luego de algunas iteraciones donde se dan lugar a las pruebas de unidad de los componentes descritos en la fase de diseño. Es decir para esta fase se realizarán pruebas de unidad, se implementarán las clases y subsistemas del modelo de implementación, y de ser necesario se corregirá el diseño y la implementación del componente.

Generalidades

Componente

Equipamiento físico de los elementos de un modelo, que encapsula el código de una aplicación para crear código reutilizable.

Servicios

Es un componente que encierra procesos o algoritmos que desempeñan funciones claras de las aplicaciones a través de la cual se puede acceder al servicio. Los Servicios permiten que las aplicaciones compartan información y que invoquen funciones de otras aplicaciones independientemente del desarrollo de las aplicaciones, el sistema operativo o la plataforma en que se ejecutan y de los dispositivos utilizados para obtener acceso a ellas.

Clase

Es una definición de un objeto en la cual se definen sus atributos o propiedades y sus métodos o funcionalidades.

Subsistema de Implementación Gerenciamiento de Pronósticos

El subsistema de implementación es consecuencia directa del subsistema de diseño, el producto final del la implementación de los componentes depende de la relación entre los mismos.

Los componentes que se utilizan en la aplicación son los de Servicios de Aplicación y la Conexión. En los servicios de la aplicación existen métodos que permiten manipular la información de la base de datos, la conexión permite abrir y cerrar un enlace con la base de datos.

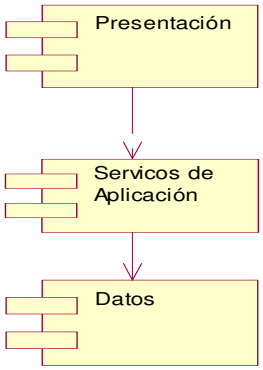


Figura 47 Componentes del subsistema de implementación Gerenciamiento de Pronósticos

Fuente: Los Autores

DESARROLLO DE LA APLICACIÓN

Definición de la Herramienta

De acuerdo a los requisitos establecidos en el análisis, diseño e implementación, se llegó a determinar la necesidad de contar con un software específico para la construcción del Sistema.

Para el Análisis y Diseño:

Herramienta Case

Rational Rose 2002 Enterprise Edition

Para el Desarrollo de la aplicación:

Visual Basic 6.0

Para la Base de Datos:

SQL Server 2000

Implementación del Sistema

La aplicación se encuentra desarrollada de acuerdo al siguiente alcance:

La aplicación está diseñada y construida para que un usuario pueda administrar la predicción de la demanda eléctrica en función de datos históricos y un modelo de aprendizaje.

El modelo de aprendizaje está inmerso en el programa SEE5, por tratarse de una aplicación investigativa, didáctica y científica, se usa una versión de prueba del mismo, por lo que el análisis sobre los datos históricos se basa en 400 registros. De todas formas la aplicación es flexible de tal manera que se pueda trabajar con un programa demostrativo o con un programa comercial.

El código de la aplicación se encuentra en el cd adjunto al proyecto.

3.3.9 PRUEBAS DEL SISTEMA

FASE DE INICIO

La elaboración de un modelo de pruebas, es el objetivo de ésta etapa de desarrollo, el cual debe irse realizando conjuntamente en el Análisis, Diseño e Implementación de la aplicación.

En ésta fase, aún no se realiza algún tipo de prueba porque el prototipo de aplicación que se tiene en ésta fase es de tipo ilustrativo, pero se empieza a trabajar en el modelo tentativo de pruebas.

Para el desarrollo de la aplicación con respecto a pruebas del sistema se considera realizar pruebas que controlen los casos de uso de la aplicación.

FASE DE ELABORACIÓN

Las pruebas aplicadas en esta fase permiten corroborar que los subsistemas funcionen correctamente con la capa de aplicación. Éstas pruebas pueden realizarse con la ejecución de los componentes obtenidos en la fase de implementación.

Las pruebas que se han considerado realizar en la fase de elaboración serán sobre los casos de usos identificados para el desarrollo del sistema.

FASE DE CONSTRUCCIÓN

En ésta fase se ponen a prueba los casos de uso y los procedimientos de prueba, considerados para usados en la aplicación.

Dentro del cumplimiento de pruebas de ésta fase es recomendable revisar el Manual de Usuario e Instalación respectivamente. Lo cual nos permitirá corregir si fuera el caso algún procedimiento o algún componente que no permita su normal secuencia.

A continuación se presenta la elaboración de las pruebas realizadas sobre la aplicación a través de los casos de uso y procedimientos de prueba.

Casos y Procedimientos de Prueba para el caso de uso Gerenciamiento de Pronóstico.

Caso de prueba para el caso de uso Validar Usuario

PRUEBAS DEL SISTEMA
CASO DE PRUEBA: Validar Usuario
DESCRIPCIÓN: Permite realizar el acceso a la aplicación.
ESCENARIO DE PRUEBAS: El usuario con perfil administrador ingresa a la aplicación una vez que tenga conocimiento de un login, password y nombre de servidor.
ENTRADA: Login, password, Nombre de Servidor
PRE CONDICIONES: 1. Debe existir un usuario registrado con su respectivo login y password.
PROCEDIMIENTO: 1. Digita el login. 2. Digita el password. 3. Selecciona el nombre del Servidor. 4. Pulsa el Botón Ingresar.
RESULTADO: Para los valores de entrada el resultado fue lo esperado, el sistema valido todos los campos ingresados permitiendo el acceso a la aplicación del usuario.
OBSERVACIÓN: En caso de no coincidir los campos de login, password y nombre de Servidor con los datos ingresados, se devuelve mensajes de error que advierten ésta situación.

Caso de prueba para el caso de uso Gestionar Pronóstico

PRUEBAS DEL SISTEMA
CASO DE PRUEBA: Gestionar Pronóstico.
DESCRIPCIÓN: Permite preparar a un pronóstico para su uso.
ESCENARIO DE PRUEBAS: El usuario Administrador ingresa al sistema, una vez autorizado el ingreso, debe seleccionar la opción para empezar la gestión de un pronóstico.
ENTRADA: nombre del pronóstico, autor, atributos, datos históricos.

PRE CONDICIONES: 1. Haber pasado el caso de prueba Validar usuario. 2. Conocer los atributos del pronóstico a ser gestionado. 3. Tener los datos históricos.
PROCEDIMIENTO: 1. Ingresa los datos del pronóstico. 2. Selecciona Guardar los datos ingresados del pronóstico. 3. Ingresa los datos de los atributos. 4. Selecciona Guardar los datos ingresados de los atributos. 5. Selecciona el pronóstico, se obtienen datos de los atributos. 6. Se comprueba la correspondencia entre loas atributos del pronóstico con los de los datos históricos, archivos .xls. 7. Se selecciona el botón guardar para que se carguen los datos históricos en el pronóstico con los respectivos atributos.
RESULTADO: Para los valores de entrada el resultado fue lo esperado, el sistema validó todos los campos ingresados permitiendo que los datos sean guardados.
OBSERVACIÓN: De no existir una correspondencia entre los atributos del pronóstico con la de los datos históricos se emite un mensaje de error para que se pueda corregir la correspondencia.

Caso de prueba para el caso de uso Predecir Demanda Eléctrica.

PRUEBAS DEL SISTEMA
CASO DE PRUEBA: Predecir Demanda Eléctrica.
DESCRIPCIÓN: Permite realizar la predicción de demanda eléctrica en función de datos históricos.
ESCENARIO DE PRUEBAS: El usuario Administrador ingresa al sistema, una vez autorizado el ingreso, debe seleccionar la opción de Preparación de Datos.
ENTRADA: Tipo carga de datos, rango para datos de predicción, valores de atributos para predicción en función de resultado.
PRE CONDICIONES: 1. Haber pasado el caso de prueba Validar usuario. 2. Haber definido un pronóstico con sus atributos y sus datos

respectivamente.
PROCEDIMIENTO: 1. Selecciona el tipo de carga de datos, esta selección esta dispuesta para un pronóstico predefinido o una tabla ejemplo que contiene los atributos característicos para una predicción de demanda. 2. Se presiona el botón Ver Datos. 3. Ingresar el rango para datos de predicción. 4. Se obtienen los resultados de la predicción, obtenidos en el caso de prueba Utilizar See5. 5. Se generan las sentencias en función del árbol y o reglas obtenidas. 6. Se ingresan los valores de los atributos que intervienen en el árbol(es) y/o regla(s) obtenidas. 7. Se presiona el botón Predecir, para obtener un rango en que se establece la demanda eléctrica.
RESULTADO: Para los valores de entrada el resultado fue lo esperado, el sistema validó la selección realizada permitiendo el despliegue de la información, además de obtener los resultados de la predicción para su análisis.
OBSERVACIÓN: En caso de no haberse definido ningún pronóstico se puede trabajar con pronóstico característico de la demanda eléctrica definido en atributosbase.

Caso de prueba para el caso de uso Preparar Datos

PRUEBAS DEL SISTEMA
CASO DE PRUEBA: Preparar Datos.
DESCRIPCIÓN: Permite preparar y utilizar los datos en función de generar archivos .names, .data.
ESCENARIO DE PRUEBAS: El usuario Administrador ingresa al sistema, una vez autorizado el ingreso, haber elegido la opción de atributos base.
ENTRADA: Atributos.
PRE CONDICIONES: 1. Haber pasado el caso de prueba Validar usuario. 2. Conocer el pronóstico a ser gestionado.

PROCEDIMIENTO: 1. Selecciona el pronóstico. 2. Selecciona los atributos que participan en la predicción. 3. Elige el botón Preparar grilla. 1. Selecciona el atributo resultado. 2. Se despliega el archivo .names generado 3. Presiona el botón Guardar. 4. Selecciono la ubicación del archivo .names a ser guardado y guardo. 5. Selecciona la manera de tomar la muestra de los datos históricos. 6. Presiona el botón Generar Archivo Data 7. Selecciono la ubicación del archivo .data a ser guardado y guardo.
RESULTADO: Para los valores de entrada el resultado fue lo esperado, el sistema valida todos los campos ingresados permitiendo que los datos sean guardados, todos los archivos fueron generados satisfactoriamente.
OBSERVACIÓN: Los archivos .names, .data generados deben guardarse en el mismo directorio.

Caso de prueba para el caso de uso Utilizar See5

PRUEBAS DEL SISTEMA
CASO DE PRUEBA: Utilizar See5.
DESCRIPCIÓN: Permite utilizar los archivos generados a través del modelo de aprendizaje.
ESCENARIO DE PRUEBAS: El usuario Administrador ingresa al sistema, una vez autorizado el ingreso, haber generado los archivos que van a intervenir en el aprendizaje.
ENTRADA: archivos: names, data.
PRE CONDICIONES: 1. Haber pasado el caso de prueba Validar usuario. 2. Haber generado los archivos names, data.
PROCEDIMIENTO: 1. Presionar el botón See5. 2. Seleccionar la opción de menú Locate Data y ubicar al archivo .data. 3. Presionar la opción Construct Classifier y seleccionar las opciones de

clasificación. 4. Presionar el botón OK. 5. En la opción de clase Ejecutar predicción presionar el botón Archivo de Salida. 6. Se despliega los resultados del modelo de aprendizaje con las condiciones seleccionadas.
RESULTADO: Para los valores de entrada el resultado fue lo esperado, el programa See5 utilizó satisfactoriamente los archivos generados por la aplicación. El archivo entregado como resultado fue analizado satisfactoriamente.
OBSERVACIÓN: El uso del sistema See5 se describen en la etapa de descripción Modelo de Aprendizaje, los resultados obtenidos dependerá de el criterio que se utilice para la construcción del clasificador.

El uso de la aplicación desarrollada, se encuentra descrita explícitamente en el anexo Manual de Usuario, adjunto al cd.

3.4 ANÁLISIS DE RESULTADOS

En esta parte del estudio se presentan los resultados obtenidos para la predicción utilizando el método de aprendizaje de máquina See5. Esta predicción se la realiza para todas las horas del día, incluyendo las 19h30, puesto que esta hora es de vital importancia en el consumo de la demanda eléctrica en el país. Para realizar las predicciones utilizando el modelo de árboles de decisión y empleando el algoritmo See5, se ha empleado el programa See5 de rulequest, en su versión completa.

Además los resultados obtenidos con el modelo de árboles de decisión, se los compara con los resultados que se obtienen con el método que utiliza el CENACE para la predicción. Dicho método, que utiliza actualmente el CENACE, se basa en las metodologías de los modelos estocásticos ARIMA utilizando el método de Box – Jenkins. Y con el fin de evaluar el desempeño de dicho modelo se establece una comparación frente al modelo de series de tiempo estocástico que emplea actualmente el CENACE, para realizar las

predicciones de la demanda eléctrica. Esta comparación se la realiza a través de medidas de error para pronóstico como el error porcentual absoluto medio (MAPE) y U de Theil, aplicados al modelo en estudio y al modelo de series de tiempo.

Para las pruebas realizadas se utilizan los registros mantenidos por el CENACE. Los registros que se utilizan para utilizarlos como la base de aprendizaje constan desde el 1° de Enero del 2003 hasta el 11 de Noviembre del 2005. Todos estos datos forman parte de la base de aprendizaje y son utilizados por la aplicación desarrollada como parte de este trabajo, y su uso se encuentra en el anexo Manual de Usuario, adjunto al cd. En la aplicación desarrollada existe la posibilidad de segmentar los datos, ya sea por clima (seco, lluvioso), dialaborable (si, no), nombre del día (lunes, martes, miércoles, jueves, viernes, sábado, domingo) y por feriado (si, no). También, si es el caso, se puede elegir, no realizar ninguna segmentación.

El objetivo de todo este proceso de segmentación o no, es la generación del archivo de aprendizaje; el cual es utilizado por el programa de rulequest See5 para realizar la clasificación de los casos que constan en el archivo.

Cabe recalcar además, que este archivo es necesario para la creación del árbol de decisión, mediante el cual se clasificarán futuros datos, de los cuales se desee saber su comportamiento, en este caso la Predicción de la Demanda Eléctrica.

Al momento de generar el árbol de decisión, el programa informa el porcentaje de error que ha ocurrido al generar el árbol, este porcentaje se genera debido a que existen casos que no se pueden clasificar y la acumulación de todos estos casos que no se pueden clasificar genera el error antes mencionado.

3.4.1 MEDIDAS DE ERROR PARA LA PREDICCIÓN

Para poder validar de forma consistente los resultados obtenidos respecto con la predicción de la Demanda Eléctrica con el modelo de árboles de decisión, se utilizan diferentes medidas de error, donde cada una evalúa los resultados

desde un punto de vista diferente. Entre los métodos que se utilizan para verificar la consistencia de los resultados obtenidos se utiliza los siguientes:

Comparación gráfica.

Es la más simple de las validaciones, esta dibuja en un mismo gráfico los datos del valor de la demanda reales, los valores pronosticados por las metodologías de predicción: See5 y ARIMA.

En este mismo gráfico se observa además, el error porcentual absoluto de cada valor de la predicción de la demanda eléctrica a cada hora, dando una idea más gráfica del porcentaje de error que cada modelo tiene.

Este porcentaje de error absoluto APE parte se lo calcula a partir de la siguiente fórmula:

$$PE_i = \left(\frac{x_i - f_i}{x_i} \right) * 100\%$$

donde: x_i = Valor Real

f_i = Valor Previsto

Y el valor de APE es igual al valor absoluto de PE, es decir:

$$APE_i = \left| \frac{x_i - f_i}{x_i} \right| * 100\%$$

Error medio porcentual (MPE Mean Percent Error).

$$MPE = \frac{1}{NF} \sum_{i=1}^{NF} \left(\frac{x_i - f_i}{x_i} \right)$$

donde: NF = Número de previsiones

Error medio porcentual absoluto (MAPE Mean Absolute Percent Error).

$$MAPE = \frac{1}{NF} \sum_{i=1}^{NF} APE_i$$

donde: $APE_i = \left| \frac{x_i - f_i}{x_i} \right| * 100\%$.

Uno de los objetivos principales de este estudio consiste en examinar cuan efectivo es el modelo de árboles de decisión para la predicción de la Demanda Eléctrica. Para tal fin se utilizará el criterio del Error Absoluto Medio Relativo (MAPE).

Este criterio se utiliza para examinar cual de los modelos estimados se ajusta mejor a la serie objeto de estudio, es decir una comparación entre el modelo de árboles de decisión y el Arima. El MAPE es un número positivo, el cual no depende de las unidades de medida. Amirkhalkhali sostiene que para efectos de decidir cual de los modelos se ajusta mejor a los datos, se deben comparar sus MAPEs y seleccionar aquel que exhiba el MAPE más bajo, generalmente igual o por debajo del nivel 0.05 ó 5%.

Coeficiente U de Theil.

$$U = \sqrt{\frac{\sum_{i=1}^{NF} (x_i - f_i)^2}{\sum_{i=2}^{NF} (x_i - x_{i-1})^2}}$$

El coeficiente de Desigualdad de Theil, indica que si este valor se acerca a cero el ajuste será perfecto.

Por esta razón se va ha utilizar las medidas, el Error Medio de Predicción (MPE), el Error Medio Absoluto de Predicción (MAPE) y el coeficiente U de Theil. Se debe señalar que valores altos de estos estadísticos indican predicciones malas, mientras que valores cercanos a cero representan buenas predicciones.

3.4.2 DESCRIPCIÓN DE CASOS PARA LAS PREDICCIONES

Para realizar la predicción del valor de la Demanda Eléctrica, se ha visto conveniente dividir las pruebas en los siguientes casos:

- Días normales

- Días no normales (Feriados).
- Días Fines de Semana

Todos estos casos, se los va también a segmentar por la clase del clima al que pertenezcan, es decir en clima seco, lluvioso, y sin distinción del clima, es decir los valores que posean el clima seco o el clima lluvioso.

Todo esto se realiza con el propósito de saber cuanto afecta la segmentación de los datos para generar el aprendizaje. Con esta segmentación se podrá observar cuan efectivo es el árbol generado. Así también se procederá a realizar la preparación del aprendizaje sin ninguna clase de segmentación de los datos.

Días Normales

Un día normal (DN) es aquel en el que no ha sucedido ningún acontecimiento extraño que modifique el consumo de energía eléctrica por parte del usuario final, es decir un día en el que no se ha tenido vacación o el día en que se trabaja normalmente.

Días no Normales (Feriado)

Un día no normal (F), es aquel en el cual ha ocurrido algún acontecimiento que haga que el día tome un comportamiento especial, así podemos clasificar en estos días, aquellas fechas que están consideradas como feriados nacionales por ejemplo: carnaval, Viernes Santo, Año Viejo, entre otros.

Días Fines de Semana

Un día fin de semana (FS) se considera aquellos días Sábados y/o Domingos, los cuales no hayan sido días feriados o en los cuales no coincida con alguna fecha especial considerada como feriado.

El resultado de la Predicción de la demanda eléctrica, cuando se utiliza el modelo de árboles de decisión, nos da una respuesta no numérica, es por esta razón que previamente se ha debido poner en Intervalos de Clase dicha variable a predecir. El procedimiento para realizar los intervalos se muestra en

el ANEXO Cálculo de Intervalos. Para el presente análisis se trabaja con varios intervalos para poder hacer una comparación y verificar el porcentaje de errores generados por cada uno.

El análisis de resultados, se lo realiza para las fechas que se detallan a continuación. Para dichas fechas se cuentan con los registros reales y con los resultados de la predicción utilizando el modelo ARIMA, todos estos registros han sido mantenidos y proporcionados por el CENACE.

Para los días normales, se van a realizar las pruebas para la semana del Lunes 09 de Enero al Viernes 13 de Enero del 2006. Para los días Fines de Semana, se realizan las pruebas para el fin de semana que pertenece a la fecha del 14 y 15 de Enero del 2006, ya estos días son Sábado y Domingo respectivamente. Para los días no normales, las pruebas se realizan para el feriado de Carnaval del año 2006, que es en la fecha Lunes 27 de Febrero y Martes 28 de Febrero de 2006. Además se prevee calcular la predicción para el feriado de Viernes Santo que corresponde a la fecha: 14 de abril de 2006.

Con el uso de la aplicación, y del programa See5 de rulequest, se han elaborado los datos de aprendizaje para días normales (semana 09 Enero – 13 Enero 2006), días fines de semana (14 y 15 Enero 2006) y para los días Feriados (Lunes , Martes de Carnaval y Viernes Santo).

Una vez generados los archivos de aprendizaje (.data), el programa See5 de rulequest, elabora un árbol de decisión, en base a los datos proporcionados como aprendizaje. Cuando se crea el árbol antes mencionado, también se genera un porcentaje de error que representa a los datos que no han podido ser clasificados correctamente. Este porcentaje es importante para darse cuenta que tan efectivo es el árbol generado, y para el presente análisis se lo tendrá en cuenta.

Es así que para cada criterio de utilización de los datos, se ha contabilizado el porcentaje de error en la generación del árbol de decisión.

Como se mencionó anteriormente, se han realizado varias pruebas para varios números de intervalos, para observar el comportamiento del error en la

generación del árbol de decisión. Los números de intervalos para los cuales se han desarrollados las pruebas son: 2, 3, 4, 5, 6, 8, 10, 15, 20, 25, 30, 35, 40, 50, 75 y 100.

A cada uno de estos intervalos se han aplicado diferentes criterios de preparación y elaboración de los casos de aprendizaje. Entre los criterios que se han aplicado constan los siguientes:

- ***Sin segmentación***:

En este caso, se cargan todos los datos que constan en la base, sin distinción de ningún criterio de partición de los datos.

- ***Con segmentación:***

Para este caso, se han tomado varios criterios de partición de datos, como se menciona a continuación:

- ***Segmentación por Día Normal***

En este tipo de segmentación constan todos los días normales, es decir todos los días que son Lunes, Martes, Miércoles, Jueves, Viernes; que son considerados como normales y en los cuales no haya ocurrido ningún evento o feriado que altere el consumo de la demanda eléctrica.

- ***Segmentación por Día No Normal***

Los casos que pertenecen a la segmentación de Día No Normal o Feriado, consta de los días los cuales son considerados como Feriados en los que no se haya laborado normalmente.

- ***Segmentación por Nombre del Día***

En ese criterio de segmentación, los datos se los puede particionar de acuerdo al nombre del Día al que pertenecen, pudiendo tomar valores de Lunes, Martes, Miércoles, Jueves, Viernes. Este criterio de segmentación tiene el objetivo de agrupar los datos que pertenezcan a un solo día. Así por ejemplo se puede agrupar todos los datos que correspondan al día Lunes.

- ***Segmentación por Fin de Semana***

Dentro de la segmentación de Fin de Semana están los días Sábado y Domingo en los cuales no haya ocurrido ningún feriado.

- ***Segmentación por Nombre del Día***

En ese criterio de segmentación, los datos se los puede particionar de acuerdo al nombre del Día al que pertenecen, pudiendo tomar valores de Sábado y Domingo. Este criterio de segmentación tiene el objetivo de agrupar los datos que pertenezcan a un solo día. Así por ejemplo se puede agrupar todos los datos que correspondan al día Sábado.

- ***Segmentación por clima:***

Dentro de este tipo de segmentación, los datos se han dividido en tres tipos de clima:

Clima seco

En la segmentación de los datos por clima seco, están todos los casos que pertenecen al clima seco, es decir todos los registros de la base de datos que tienen un clima seco.

Clima lluvioso

En la segmentación de los datos por clima lluvioso, están todos los casos que pertenecen al clima lluvioso, es decir todos los registros de la base de datos que tienen un clima lluvioso.

Clima seco-lluvioso

En la segmentación de los datos por clima seco-lluvioso, están todos los casos que pertenecen al clima seco-lluvioso, es decir todos los registros de la base de datos que tienen un clima seco-lluvioso.

Una vez detallados los criterios de segmentación, se va a listar aquellas combinaciones utilizadas para las pruebas realizadas.

Para clima Lluvioso:

Clima Lluvioso – Día Normal.-

Todos los casos que son lluviosos y son los días de Lunes a Viernes

Clima Lluvioso – Día Normal – Lunes.-

Todos los casos que son lluviosos y son solo día Lunes.

Clima Lluvioso – Día Normal – Martes.-

Todos los casos que son lluviosos y son solo día Martes.

Clima Lluvioso – Día Normal – Miércoles.-

Todos los casos que son lluviosos y son solo día Miércoles.

Clima Lluvioso – Día Normal – Jueves.-

Todos los casos que son lluviosos y son solo día Jueves.

Clima Lluvioso – Día Normal – Viernes.-

Todos los casos que son lluviosos y son solo día Viernes.

Clima Lluvioso – Fin Semana.-

Todos los casos que son lluviosos y son los días Sábado y Domingo.

Clima Lluvioso – Fin Semana – Sábado.-

Todos los casos que son lluviosos y son solo día Sábado.

Clima Lluvioso – Fin Semana – Domingo.-

Todos los casos que son lluviosos y son solo día Domingo.

Clima Lluvioso – Feriado.

Todos los casos que son lluviosos y pertenecen a algún feriado.

Para clima Seco:

Clima Seco – Día Normal.-

Todos los casos que son secos y son los días de Lunes a Viernes

Clima Seco – Día Normal – Lunes.-

Todos los casos que son secos y son solo día Lunes.

Clima Seco – Día Normal – Martes.-

Todos los casos que son secos y son solo día Martes.

Clima Seco – Día Normal – Miércoles.-

Todos los casos que son secos y son solo día Miércoles.

Clima Seco – Día Normal – Jueves.-

Todos los casos que son secos y son solo día Jueves.

Clima Seco – Día Normal – Viernes.-

Todos los casos que son secos y son solo día Viernes.

Clima Seco – Fin Semana.-

Todos los casos que son secos y son los días Sábado y Domingo.

Clima Seco – Fin Semana – Sábado.-

Todos los casos que son secos y son solo día Sábado.

Clima Seco – Fin Semana – Domingo.-

Todos los casos que son secos y son solo día Domingo.

Clima Seco – Feriado.-

Todos los casos que son secos y pertenecen a algún feriado.

Para Clima Seco-Lluvioso:

Clima Seco-Lluvioso – Día Normal.-

Todos los casos que son secos y son los días de Lunes a Viernes

Clima Seco-Lluvioso – Día Normal – Lunes.-

Todos los casos que son secos y son solo día Lunes.

Clima Seco-Lluvioso – Día Normal – Martes.-

Todos los casos que son secos y son solo día Martes.

Clima Seco-Lluvioso – Día Normal – Miércoles.-

Todos los casos que son secos y son solo día Miércoles.

Clima Seco-Lluvioso – Día Normal – Jueves.-

Todos los casos que son secos y son solo día Jueves.

Clima Seco-Lluvioso – Día Normal – Viernes.-

Todos los casos que son secos y son solo día Viernes.

Clima Seco-Lluvioso – Fin Semana.-

Todos los casos que son secos y son los días Sábado y Domingo.

Clima Seco-Lluvioso – Fin Semana – Sábado.-

Todos los casos que son secos y son solo día Sábado.

Clima Seco-Lluvioso – Fin Semana – Domingo.-

Todos los casos que son secos y son solo día Domingo.

Clima Seco-Lluvioso – Feriado.-

Todos los casos que son secos-lluviosos y pertenecen a algún feriado.

A continuación se muestra una tabla en la que se puede detallar el número de casos para cada criterio de segmentación:

	Clima / Día	Lluvioso		Seco		Seco-Lluvioso	
Día Normal	Lunes	1456	7410	2262	11206	3718	18616
	Martes	1482		2262		3744	
	Miércoles	1534		2210		3744	
	Jueves	1508		2236		3744	
	Viernes	1430		2236		3666	
Fin Semana	Sábado	1300	2626	2080	4108	3380	6734
	Domingo	1326		2028		3354	
Feriado		962		884		1846	

Tabla 20 Número de Casos para cada criterio de segmentación

Fuente: Los Autores

Total de Casos: 27196

Cabe recalcar que en todos los casos, mostrados en la anterior tabla, constan los valores para cada hora del día, desde las 00h00 hasta las 24h00 incluidos también los valores de las 19h30.

Expresado en porcentajes se tiene la siguiente partición de los datos:

	Clima / Día	Lluvioso		Seco		Seco-Lluvioso	
Día Normal	Lunes	5.35%	27.25%	8.32%	41.20%	13.67%	68.45%
	Martes	5.45%		8.32%		13.77%	
	Miércoles	5.64%		8.13%		13.77%	
	Jueves	5.54%		8.22%		13.77%	
	Viernes	5.26%		8.22%		13.48%	
Fin Semana	Sábado	4.78%	9.66%	7.65%	15.11%	12.43%	24.76%
	Domingo	4.88%		7.46%		12.33%	
Feriado		3.54%		3.25%		6.79%	

Tabla 21 Porcentaje de Casos para cada criterio de segmentación

Fuente: Los Autores

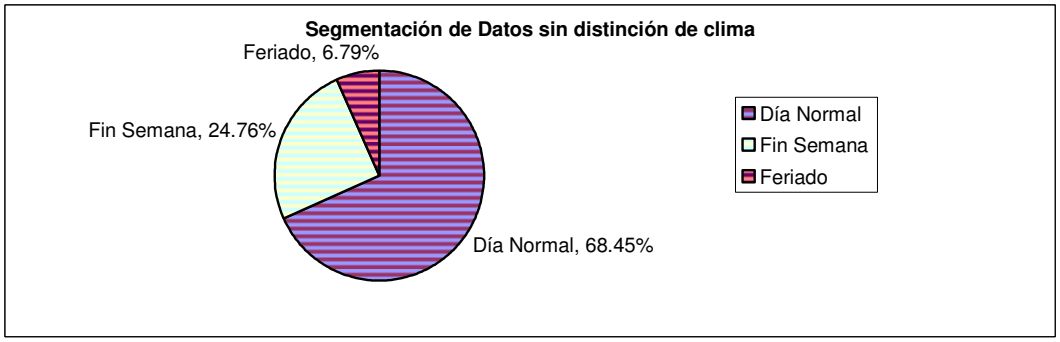


Figura 48 Segmentación de datos

Fuente: Los Autores

De esta forma se puede observar la manera en la que están dispuestos los datos, para tener una idea del porcentaje en que están representados los casos, para cada criterio de partición antes mencionados. De lo que se puede mencionar que el 68.45% de los datos, pertenecen a un clima seco-lluvioso y son días normales, el 24.76% pertenecen a Fin de Semana con clima seco-lluvioso y un 6.79% son casos que pertenecen a Feriado con clima seco-lluvioso.

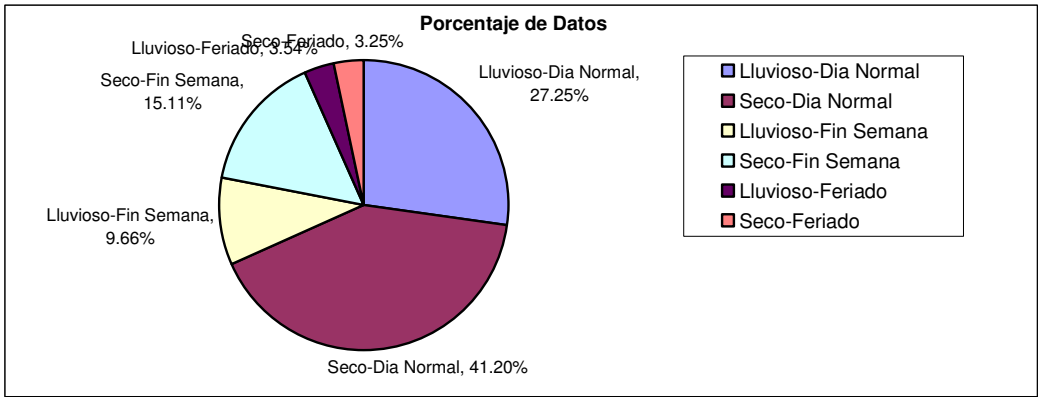


Figura 49 Porcentaje de datos con criterios de segmentación
Fuente: Los Autores

De lo graficado se puede observar que existen más días laborables (días normales) secos que lluviosos, puesto que los días normales secos tienen una representación del 41.20%, mientras que el 27.25% pertenecen a días normales lluviosos.

Con respecto a los fines de semana, existen más fines de semana que pertenecen al clima seco, mientras que los fines de semana que pertenecen al clima seco son el 9.66%. Por último, con relación a los feriados, se observa que existen más feriados (días no normales) que pertenecen al clima lluvioso con un porcentaje de 3.54% respecto a un 3.25% de feriados secos.

Si se particiona los datos de acuerdo al criterio del nombre del día de la semana se obtiene el siguiente gráfico, que muestra el porcentaje de representación de cada día de la semana, dentro de la totalidad de datos:

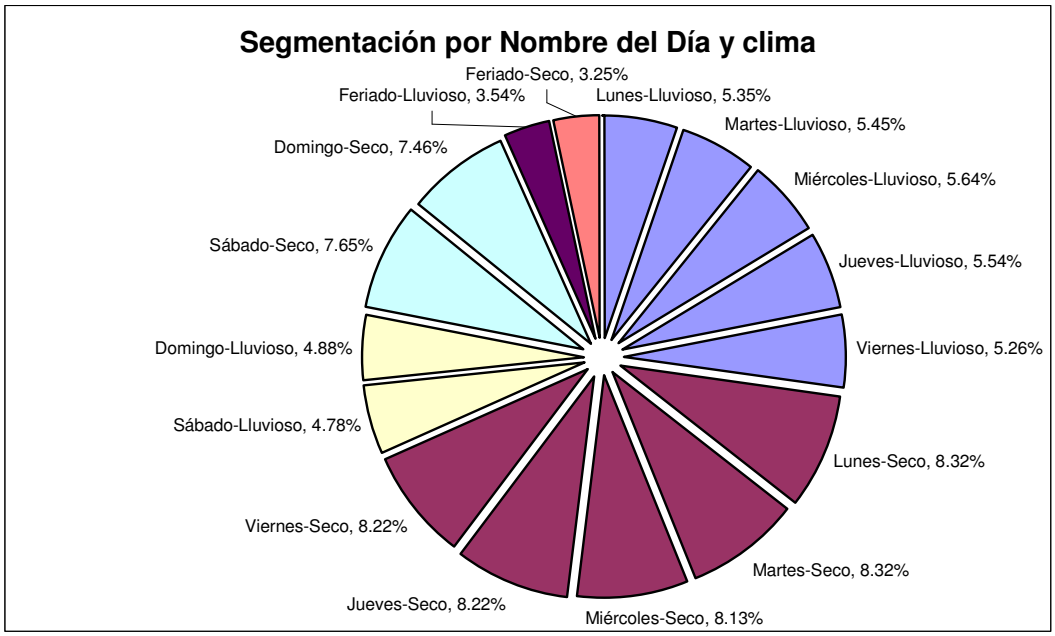


Figura 50 Segmentación por Nombre del día y por clima

Fuente: Los Autores

3.4.3 PORCENTAJES DE ERROR EN LA GENERACIÓN DEL ÁRBOL DE DECISIÓN

Una vez detallados los criterios para las pruebas, y con el conocimiento de que existe un error al generar el árbol de aprendizaje para cada criterio de segmentación antes mencionado, se procede a registrar dichos errores:

Sin segmentación:

Para este criterio, sin segmentación, se toman todos los casos que se encuentran en la base de datos, es decir, toda la información sin partición alguna. Dentro de los registros se encuentran todos los valores para todas las horas de día, de todos los días de los años que consten en el banco de datos.

Así, cuando se genera el archivo de aprendizaje (.data) para posteriormente generar su respectivo árbol de aprendizaje, se obtienen los siguientes porcentajes de error para los intervalos mencionados anteriormente:

	Sin Segmentar
Intervalos	% Error
2	2.7
3	4.8
4	8.1
5	9.5
6	10.4
8	13.5
10	17.1
15	24.5
20	30.7
25	35.1
30	39.4
35	43.9
40	48.5
50	52.5
75	54.6
100	59.8

Tabla 22 Porcentaje Error – sin segmentación

Fuente: Los Autores

Se observa que el menor porcentaje de error es para 2 intervalos (2.7%), además que mientras el número de intervalos aumenta, también crece el porcentaje de error en la generación del árbol de aprendizaje.

Segmentación Clima Lluvioso

Cuando se segmentan los datos por el clima, y en este caso solo aquellos casos que tienen el clima lluvioso, y además para días normales, Fines de Semana y Feriados se obtienen los siguientes porcentajes de error:

	Clima lluvioso		
	Día Normal	Fin Semana	Feriado
Intervalos	% Error	% Error	% Error
2	3.6	1.3	3.8
3	4.1	8	21.3
4	7	8.8	30.9
5	9.5	13.6	31.9
6	10.2	16.8	34.6
8	11.5	18	40.3
10	18	34.3	51.6
15	26.3	51.4	56.9
20	36.2	46.9	55.5
25	41.9	51.8	55.8

30	49	48.8	55.6
35	55.1	53.7	47
40	58.6	49	48.8
50	60.3	51.4	49.5
75	57.3	52.5	43.5
100	51.5	53.7	40.9

Tabla 23 Porcentaje de error Clima lluvioso

Fuente: Los Autores

Analizando los porcentajes obtenidos, se puede observar que, igual que el criterio anterior, a menor número de intervalos al porcentaje de error es menor, ya sea para día normal, Fin de Semana y Feriados; y mientras mayor sea el número de intervalos también será mayor el valor del porcentaje de error, para día normal, fin de semana y feriado.

Segmentación Clima Seco

Al segmentar los datos por clima seco, se obtiene los siguientes valores de porcentaje de error para Día Normal, Fin de Semana y Feriado, cuando se elabora el respectivo árbol de decisión para cada segmentación realizada:

	Clima seco		
	Día Normal	Fin Semana	Feriado
Intervalos	% Error	% Error	% Error
2	3.2	0.6	2.3
3	2.7	5.4	16.2
4	7.1	8.1	21.7
5	7.7	8.2	25.6
6	7.8	8.9	29.2
8	11.4	16.4	30.7
10	13.8	18.8	33
15	20.8	31	46.3
20	27.6	44.1	47.1
25	34.6	51.6	51.2
30	41.3	59.2	48.3
35	46.9	56.9	54
40	51.6	60.7	51.7
50	58.1	57.3	56
75	60.9	56.2	55.5
100	59.7	57.3	51.8

Tabla 24 Porcentaje de error para clima seco

Fuente: Los Autores

Se puede observar que cuando el número de intervalos es el mínimo, el porcentaje de error es el más bajo registrado. Y mientras se aumentan el número de intervalos, también crece el valor del porcentaje de error.

Segmentación Clima Seco-Lluvioso

Al segmentar los datos, por clima seco-lluvioso, es decir particionar los datos sin importar el clima, para los días normales, fines de semana y feriados se obtienen los siguientes porcentajes de error en la generación del árbol de decisión para cada uno:

	Clima seco y lluvioso		
	Día Normal	Fin Semana	Feriado
Intervalos	% Error	% Error	% Error
2	3.3	0.9	2.8
3	3.1	6.5	20.6
4	7.2	9	25.9
5	8.4	9.2	29.8
6	8.9	10.5	31.4
8	11.9	14.7	39.4
10	15.7	17.7	47
15	22.5	27.1	59.3
20	29.1	35.4	54.9
25	33.1	41.9	54.2
30	37.2	43.8	52.2
35	44.4	45.1	48.8
40	48.1	47.5	54.1
50	54.1	49.4	53.7
75	60.5	50.9	52.2
100	61	51.4	50.2

Tabla 25 Porcentaje de Error clima seco-lluvioso

Fuente: Los Autores

De la misma manera, que en casos anteriores, el número de intervalos influye directamente con el valor del porcentaje de error, mientras menor sea el número de intervalos, también menor será el porcentaje de error en la generación del árbol de decisión.

En resumen, apilando todos los resultados obtenidos, es decir para clima lluvioso, clima seco, clima seco-lluvioso, y los datos sin segmentar, obtenemos la siguiente tabla:

	Sin Segmentar	Clima lluvioso			Clima seco			Clima seco y lluvioso		
		DN	FS	F	DN	FS	F	DN	FS	F
Intervalos	% Err	% Err	% Err	% Err	% Err	% Err	% Err	% Err	% Err	% Err
2	2.7	3.6	1.3	3.8	3.2	0.6	2.3	3.3	0.9	2.8
3	4.8	4.1	8	21.3	2.7	5.4	16.2	3.1	6.5	20.6
4	8.1	7	8.8	30.9	7.1	8.1	21.7	7.2	9	25.9
5	9.5	9.5	13.6	31.9	7.7	8.2	25.6	8.4	9.2	29.8
6	10.4	10.2	16.8	34.6	7.8	8.9	29.2	8.9	10.5	31.4
8	13.5	11.5	18	40.3	11.4	16.4	30.7	11.9	14.7	39.4
10	17.1	18	34.3	51.6	13.8	18.8	33	15.7	17.7	47
15	24.5	26.3	51.4	56.9	20.8	31	46.3	22.5	27.1	59.3
20	30.7	36.2	46.9	55.5	27.6	44.1	47.1	29.1	35.4	54.9
25	35.1	41.9	51.8	55.8	34.6	51.6	51.2	33.1	41.9	54.2
30	39.4	49	48.8	55.6	41.3	59.2	48.3	37.2	43.8	52.2
35	43.9	55.1	53.7	47	46.9	56.9	54	44.4	45.1	48.8
40	48.5	58.6	49	48.8	51.6	60.7	51.7	48.1	47.5	54.1
50	52.5	60.3	51.4	49.5	58.1	57.3	56	54.1	49.4	53.7
75	54.6	57.3	52.5	43.5	60.9	56.2	55.5	60.5	50.9	52.2
100	59.8	51.5	53.7	40.9	59.7	57.3	51.8	61	51.4	50.2

Tabla 26 Resumen de porcentaje de Error en general

Fuente: Los Autores

DN = Día Normal

FS = Fin de Semana

F = Feriado

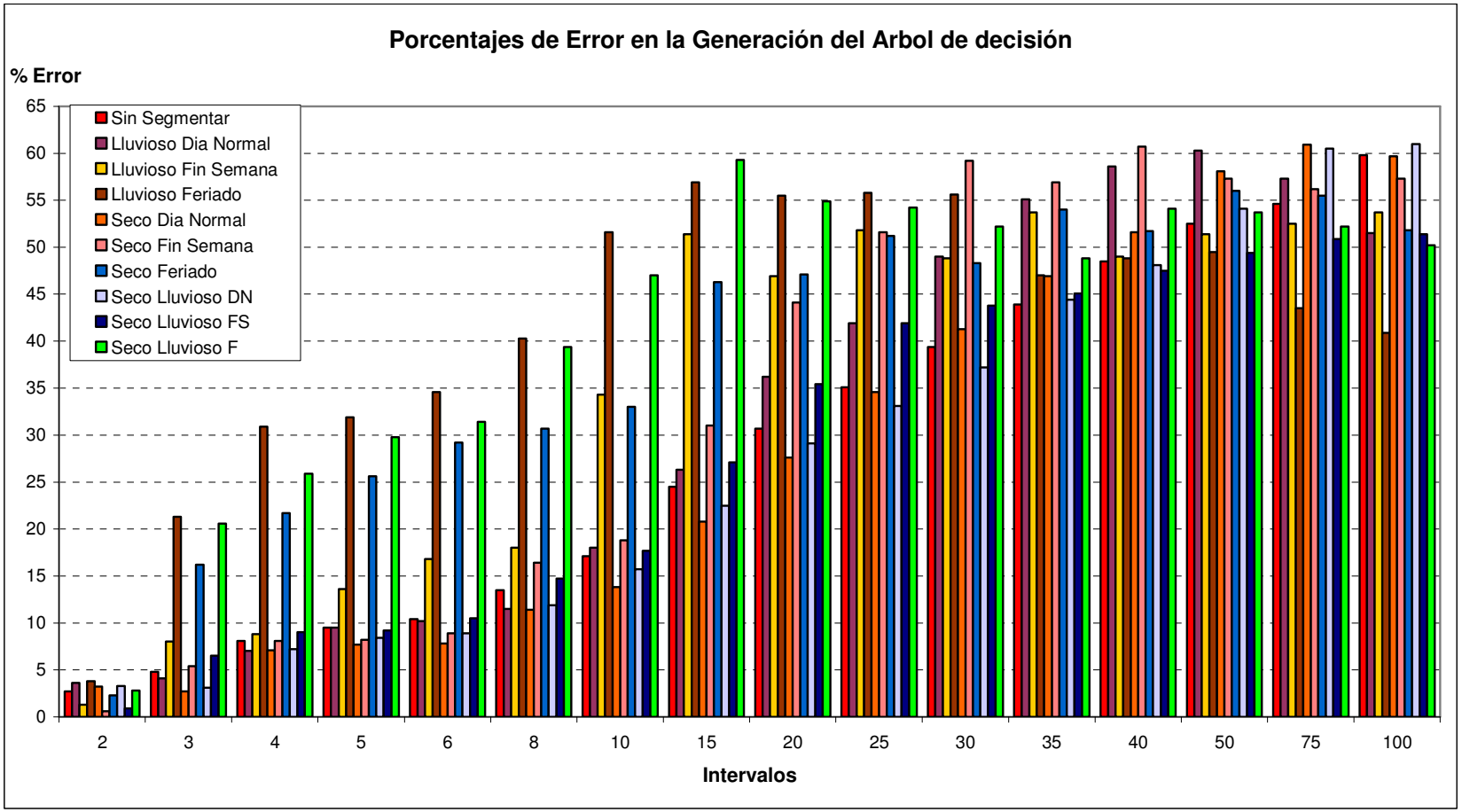


Figura 51 Porcentajes de error en la generación del Árbol de decisión

Fuente: Los Autores

Analizando de forma general los resultados de valores de porcentajes de error en la generación de los árboles de decisión obtenidos, se puede observar muy claramente que a menor número de intervalos este porcentaje de error es más bajo, esto es así puesto que en todos los casos cuando el número de intervalos es 2, el porcentaje de error para cada partición de datos es menor en comparación con su respectivo valor al número de intervalo superior. Cuando el número de intervalos crece, crece también el porcentaje de error al aumentar el número de intervalos.

Esto se debe a que mientras más intervalos tiene la variable a predecir, más difícil es la clasificación de los casos, por parte del algoritmo See5, y además se hace más complejo generar el árbol de decisión, puesto que existen más intervalos.

Al realizar el análisis para el menor número de intervalos, es decir 2 intervalos, se observa que el porcentaje de error, al no segmentar los datos, es 2.7%. Mientras que segmentando los datos por fin de semana ya sea por clima lluvioso, clima seco y sin segmentación del clima los valores son 1.3%, 0.6% y 0.9%, resultando estos valores menores que el porcentaje de error sin segmentar los datos.

Para 3 intervalos, se observa que cuando se segmenta el clima o sin segmentar el clima, para los días normales; los porcentajes de error disminuyen respecto al porcentaje de error sin segmentar los datos. Mientras que al segmentar por clima o sin segmentación del clima, para los fines de semana el porcentaje de error es mayor comparado con el porcentaje de error sin segmentar; esto mismo ocurre para los días feriados, siendo estos porcentajes considerablemente mucho mayores.

Toda esta regla ocurre para los demás intervalos, es decir, que cuando se segmenta, por clima seco o lluvioso o seco-lluvioso, para los días normales; estos porcentajes son menores comparados con los porcentajes de error cuando no se segmentan los datos; a excepción de los valores de intervalos (10,15) cuando se nota claramente, que para estos casos el porcentaje de los días normales es mayor que cuando no se segmentan los datos.

Además, el porcentaje de error cuando se segmenta por clima o no segmenta por clima, para los fines de semana, es mayor que el valor del porcentaje de error cuando no se realiza ninguna segmentación. Lo mismo ocurre para los días feriados, que el porcentaje de error es mayor, comparado cuando no se segmentan los datos, y en este caso se nota que este porcentaje es mucho mayor en comparación del porcentaje de error al no segmentar los datos.

Una vez presentados los porcentajes de error para los días normales (lunes – viernes), fines de semana (Sábado y Domingo), a continuación se presenta de una manera más detallada los porcentajes de error, para cada día de la semana, es decir, que para un día normal, se presenta de forma general, y además se presenta el error para el día Lunes, Martes, Miércoles, Jueves, Viernes; así como también para un fin de semana se presenta el error para todo el fin semana tomando en cuenta los dos días del fin de semana, y también se presenta el error para cada día del fin de semana, es decir el error para el Sábado y para el Domingo. De esta forma se pude observar el comportamiento del porcentaje del error y saber si afecta o no la segmentación de los datos en el porcentaje de error en la generación del árbol de decisión.

En la siguiente tabla se presentan los porcentajes de error que se obtuvieron al generar el árbol de decisión para cada criterio de segmentación de los datos, es decir se presenta, para cada intervalo el porcentaje cuando no se segmentan los datos, cuando se segmentan por clima y solo por día normal, fin de semana y feriado, y además para cada día de la semana ya sea segmentado por clima lluvioso, seco o seco-lluvioso:

Intervalos	Sin Segmentar		Clima Día	Lluvioso		Seco		Seco-Lluvioso	
	% Err			% Err	% Err	% Err	% Err	% Err	% Err
2	2.7	DN	Lunes	4	3.6	3.5	3.2	4.7	3.3
			Martes	3.4		3.4		4.4	
			Miércoles	3.7		4.5		5.2	
			Jueves	5.2		3.4		4.2	
			Viernes	4.3		4.7		4.2	
		FS	Sábado	1.4	1.3	0.7	0.6	1.1	0.9
			Domingo	0.8		0.4		0.6	
		F		3.8		2.3		2.8	
3	4.8	DN	Lunes	3.3	4.1	2.9	2.7	3.9	3.1
			Martes	3.8		2.6		3.7	

			Miércoles	3.8		2.2		3.3		
			Jueves	4.5		2.4		3.6		
			Viernes	5.9		2.1		3.8		
		FS	Sábado	9.8	8	5.9	5.4	8.2	6.5	
			Domingo	6.5		4.8		5.4		
		F		21.3		16.2		20.6		
4	8.1	DN	Lunes	9	7	7.9	7.1	8.9	7.2	
			Martes	6.6		9.7		8		
			Miércoles	8.3		10.5		9.6		
			Jueves	9.4		8.8		8.8		
			Viernes	9.7		10.5		8.3		
		FS	Sábado	7.7	8.8	7.5	8.1	7.7	9	
			Domingo	10.3		10		10.9		
		F		30.9		21.7		25.9		
5	9.5	DN	Lunes	10.3	9.5	10.3	7.7	10	8.4	
			Martes	11.8		8.1		10.4		
			Miércoles	10.4		8.1		9.7		
			Jueves	12.9		8.5		11		
			Viernes	15.7		9.5		12.3		
		FS	Sábado	9.2	13.6	11.1	8.2	10.4	9.2	
			Domingo	13.3		8.7		9.4		
		F		31.9		25.6		29.8		
6	10.4	DN	Lunes	13	10.2	14.2	7.8	11.5	8.9	
			Martes	10.5		11.9		9.7		
			Miércoles	10.1		11.4		11.4		
			Jueves	13.3		11.9		11		
			Viernes	14.7		13.1		13.9		
		FS	Sábado	15.2	16.8	9.8	8.9	11.9	10.5	
			Domingo	12.1		10.8		11.6		
		F		34.6		29.2		31.4		
8	13.5	DN	Lunes	16.1	11.5	17.9	11.4	14.2	11.9	
			Martes	14.1		14.3		14.3		
			Miércoles	17.1		18.1		15		
			Jueves	19.2		17.3		16.7		
			Viernes	20.3		19.3		17.8		
		FS	Sábado	17.3	18	16.3	16.4	16.5	14.7	
			Domingo	19		15.5		17.4		
		F		40.3		30.7		39.4		
10	17.1	DN	Lunes	21.8	18	23.1	13.8	18.6	15.7	
			Martes	27.7		18.6		17.4		
			Miércoles	27.9		22.2		20.3		
			Jueves	38.8		21.4		22.4		
			Viernes	39.9		22.9		22.7		
		FS	Sábado	33.1	34.3	20.7	18.8	19.5	17.7	
			Domingo	33.1		18.2		19.4		
		F		51.6		33		47		
15	24.5	DN	Lunes	46.4	26.3	43.3	20.8	27.1	22.5	
			Martes	50.7		37		23.9		
			Miércoles	52.5		73		27.9		
			Jueves	49.4		32.8		29.9		
			Viernes	58		38.2		28.5		

		FS	Sábado	57.2	51.4	33.7	31	27.4	27.1
			Domingo	52.7		32		32	
		F		56.9		46.3		59.3	
20	30.7	DN	Lunes	58.2	36.2	56.6	27.6	34.5	29.1
			Martes	65.8		54.4		35.7	
			Miércoles	61.5		52.1		38.7	
			Jueves	58.8		54.4		39.4	
			Viernes	61.7		57.8		36.6	
		FS	Sábado	64.5	46.9	54.9	44.1	35.1	35.4
			Domingo	60.6		45.7		42.1	
		F		55.5		47.1		54.9	
25	35.1	DN	Lunes	62.3	41.9	66	34.6	47.5	33.1
			Martes	67		64.5		41.6	
			Miércoles	63.8		66.3		43.6	
			Jueves	54.9		65.8		47.8	
			Viernes	57.7		72.1		50.6	
		FS	Sábado	56.6	51.8	70.5	51.6	46.9	41.9
			Domingo	58.5		61.4		42.8	
		F		55.8		51.2		54.2	
30	39.4	DN	Lunes	59.9	49	72.4	41.3	46.3	37.2
			Martes	65.8		74.3		48.6	
			Miércoles	57.6		75.2		49.5	
			Jueves	47.1		72.9		53.3	
			Viernes	56.8		74		48.9	
		FS	Sábado	57.3	48.8	75.9	59.2	54.6	43.8
			Domingo	63.4		71.9		50	
		F		55.6		48.3		52.2	
35	43.9	DN	Lunes	61.4	55.1	74.5	46.9	48.4	44.4
			Martes	57.8		77.3		46.8	
			Miércoles	59.6		76.3		50.3	
			Jueves	43.2		76.7		50.1	
			Viernes	56.4		76.4		51.7	
		FS	Sábado	47.7	53.7	77.9	56.9	48	45.1
			Domingo	57.2		75.7		54.6	
		F		47		54		48.8	
40	48.5	DN	Lunes	54.2	58.6	76.1	51.6	51.4	48.1
			Martes	58.3		78.4		50.1	
			Miércoles	53.3		79.6		54.5	
			Jueves	47.6		79		52.7	
			Viernes	49.5		76.7		48.9	
		FS	Sábado	41.6	49	78.7	60.7	49.7	47.5
			Domingo	55.4		76.5		52.7	
		F		48.8		51.7		54.1	
50	52.5	DN	Lunes	50.3	60.3	74.8	58.1	52.8	54.1
			Martes	56.3		77.4		52	
			Miércoles	56.3		78.6		55.2	
			Jueves	49.1		77.4		54	
			Viernes	55.3		74.4		53.3	
		FS	Sábado	39.8	51.4	77.5	57.3	53.6	49.4
			Domingo	46		77.5		53.2	
		F		49.5		56		53.7	
75	54.6	DN	Lunes	50.1	57.3	67.2	60.9	53.4	60.5

			Martes	48.4		74.7		55.5	
			Miércoles	53		75		55.4	
			Jueves	46.8		70.8		56	
			Viernes	52.1		67.9		54.4	
		FS	Sábado	45	52.5	73.4	56.2	54.2	50.9
			Domingo	46		74.2		54.6	
		F		43.5		55.5		52.2	
100	59.8	DN	Lunes	45.1	51.5	66.5	59.7	55.3	61
			Martes	46.5		68.2		55.5	
			Miércoles	45.9		65.1		57.2	
			Jueves	46.8		66.1		58	
			Viernes	43.1		68.6		56.3	
		FS	Sábado	41.9	53.7	63.7	57.3	55.7	51.4
			Domingo	42.2		68.3		56.3	
		F		40.9		51.8		50.2	

Tabla 27 Porcentaje de Error de Lunes a Domingo

Fuente: Los Autores

Esta tabla muestra en forma general y resumida todos los porcentajes de error generados durante la creación de los respectivos árboles de decisión, para diferente número de Intervalos de clase de la variable a predecir. Todos los archivos de resultados obtenidos se encuentran como anexo en el cd.

Dentro de los resultados de los porcentajes de error que se obtuvieron se puede notar que mientras menor sea el número de intervalos de clase de la variable a predecir, menor también será el porcentaje de error en la generación del árbol de decisión. Esta regla se cumple para todos los criterios de segmentación y para cuando no se segmentan los datos. De esta manera, se observa que cuando se segmentan los datos para que consten solo aquellos que cumplan la condición de clima lluvioso, y sean días laborables, y generamos el árbol de decisión para 2 intervalos, el porcentaje de error de ese árbol es 3.6%, mientras que para árboles con las mismas condiciones pero para 3 intervalos, el porcentaje de error es 4.1%, y al generar los sucesivos árboles para intervalos siguientes el porcentaje de error aumenta. Lo mismo ocurre para los otros criterios de segmentación, es decir cuando segmentamos para fin de semana y feriados, ya sea para clima seco o con clima seco-lluvioso.

Se observa que los valores del porcentaje de error en la generación del árbol de decisión, que pertenecen a la segmentación Día normal ya sea para

cualquier partición de clima, son menores al porcentaje cuando no se segmentan los datos, esto se cumple para los intervalos 3, 4, 5, 6 y 8. Puesto que para los intervalos 10,15, 20, solo son menores los valores de porcentajes que pertenecen a la segmentación de clima seco y clima seco-lluvioso. Cuando el número de intervalos es 30, los valores de porcentajes son menores cuando son segmentados por clima seco-lluvioso.

Los valores de porcentajes de error para los fines de semana, con partición de clima seco, lluvioso y seco-lluvioso solo son menores cuando el número de intervalos es 2 y 100. Cuando el intervalo es 4 y 6, el porcentaje de error solo es menor cuando se segmenta por clima seco. Si el número de intervalos es 5 y 30, en cambio, los porcentajes que son menores son los que pertenecen a la segmentación de clima seco y seco-lluvioso. Para el resto de intervalos el porcentaje cuando se segmenta por fin de semana es mayor al porcentaje cuando no se segmentan los datos.

En cambio, cuando se habla de días feriados, el porcentaje es siempre mayor comparado al porcentaje cuando no se segmentan los datos, a excepción cuando el número de intervalos es 2 y 100 con segmentación de clima seco, como se observa en el gráfico a continuación:

Además, los datos se pueden particionar por el nombre del día al que pertenecen. Esta segmentación con la representación del porcentaje error en la generación del árbol de decisión como se muestra en la figura 53.

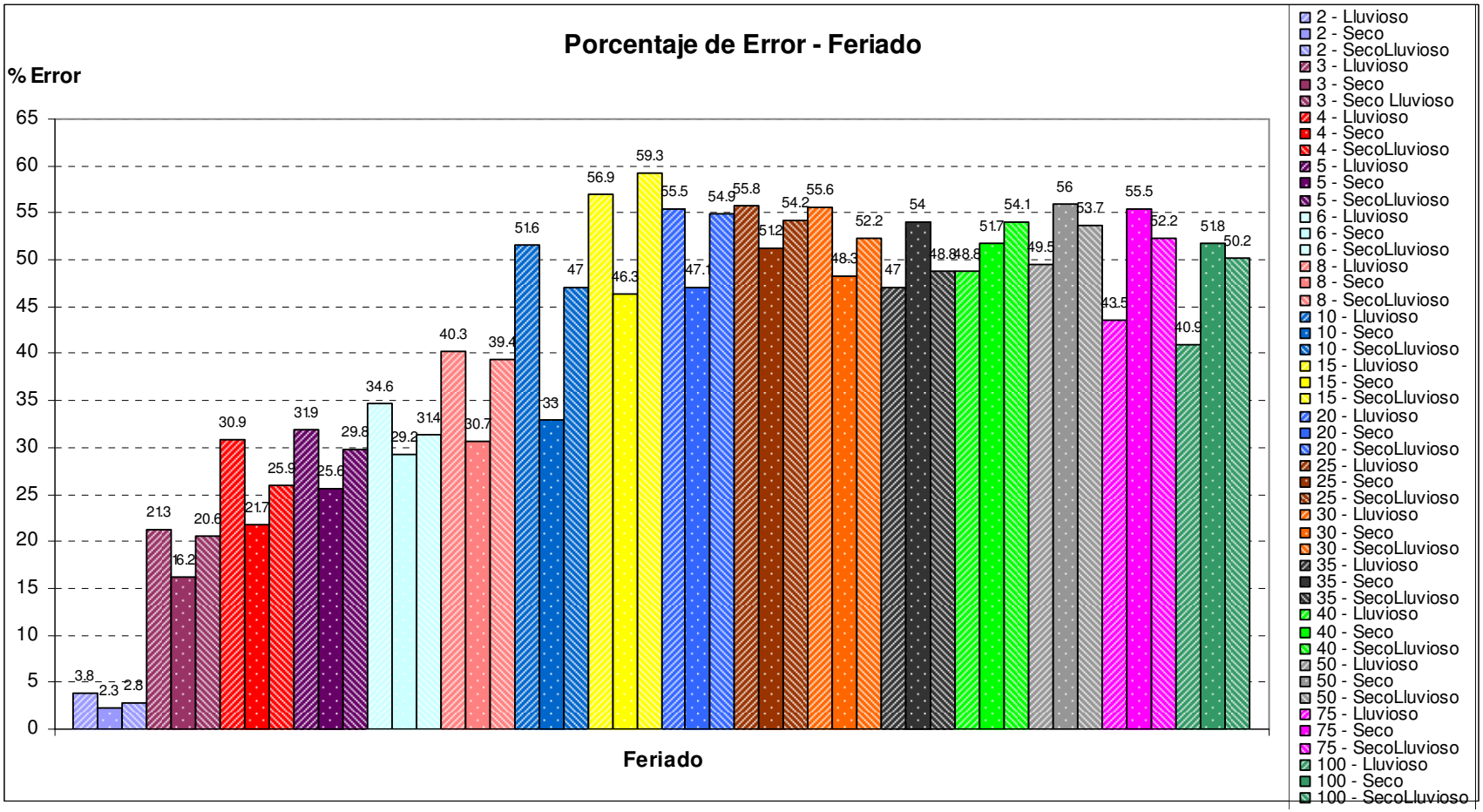


Figura 52 Porcentajes de error - Feriado

Fuente: Los Autores

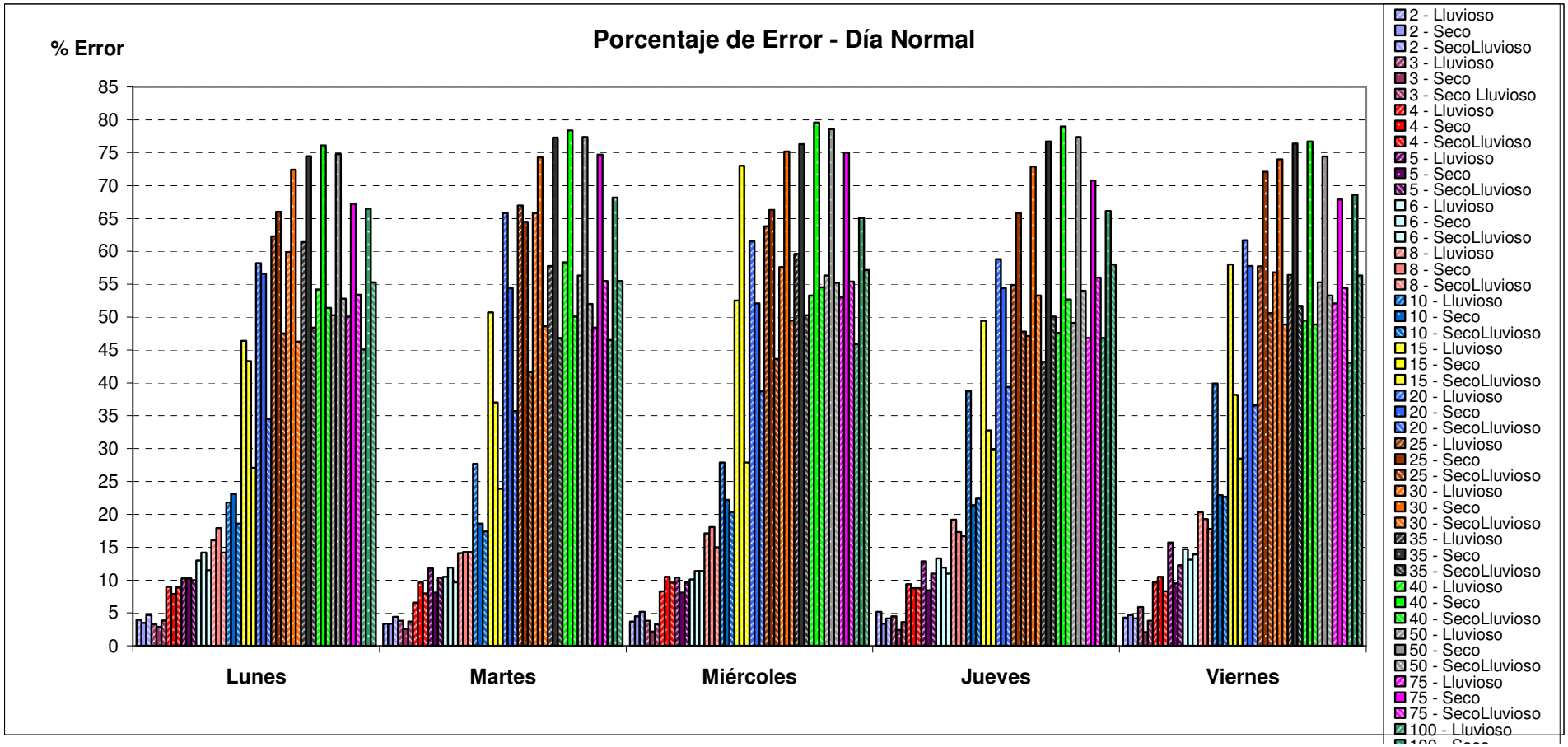


Figura 53 Porcentajes de error – Días normales

Fuente: Los Autores

Cuando se particiona los datos por el nombre del día para los días normales, es decir cuando se generan árboles diferentes para los días Lunes, Martes, Miércoles, Jueves y Viernes que son días laborables, se observa que solo cuando el número de intervalos es 3 el porcentaje de error es menor que el porcentaje de error al no particionar los datos, todo esto para cualquier clima. Cuando el número de intervalos es 4, el porcentaje de error solo es menor al porcentaje cuando no se segmentan los datos, para el día Martes con clima lluvioso y con clima seco-lluvioso, y para el Lunes con clima seco. Para un número de 5 intervalos, el porcentaje es menor solo para los días Martes a Viernes para un clima seco, y para 6 intervalos, el error solo es menor para el día Miércoles con clima lluvioso y para el Martes con clima seco-lluvioso. El resto de valores de porcentajes de error cuando se segmenta por el nombre del día normal, es mayor al valor cuando no existe segmentación alguna, a excepción cuando el número de intervalos es 100, donde para los días normales el porcentaje de error es menor cuando no se segmentan los datos.

Si se particiona los datos por el nombre del día para los fines de semana, es decir para los días Sábados y Domingos, se observa que cuando el número de intervalos es 2, estos porcentajes de error son menores que cuando no se segmentan los datos, para los dos días mencionados. Cuando el número de intervalos es 4, el error es menor solo para el día sábado con cualquier clima. Cuando el intervalo corresponde al valor de 5, el error es menor para el día Sábado con clima lluvioso y para el Domingo con clima seco y clima seco-lluvioso, como se muestra en el siguiente gráfico:

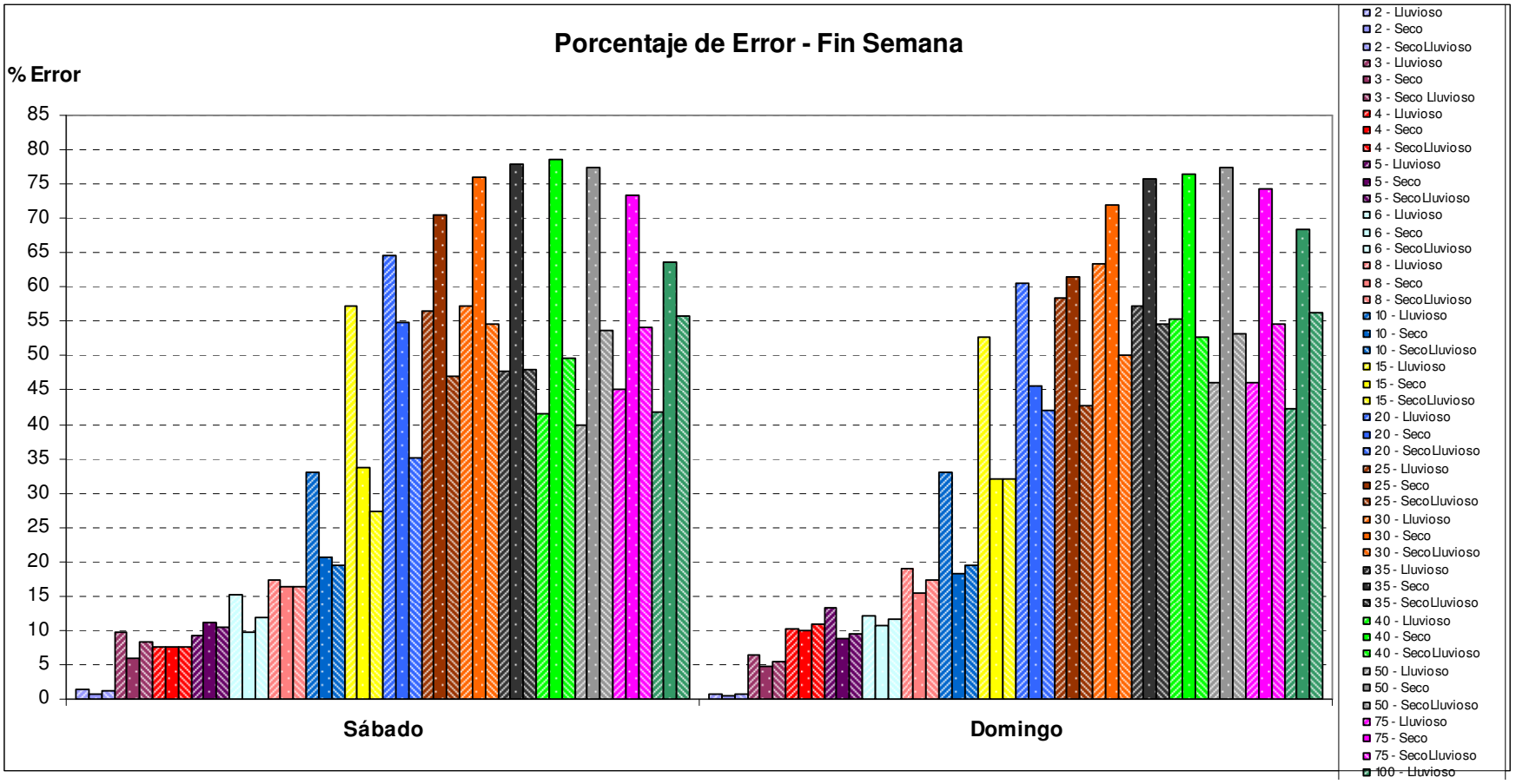


Figura 54 Porcentajes de error – Fines de semana

Fuente: Los Autores

Por lo expuesto anteriormente, se observa que la segmentación de los datos ayuda a disminuir el porcentaje de error en la generación de los árboles de decisión, solo para ciertos casos, no para todos. Así, cuando segmentamos los datos por el nombre del día, existen pocos valores que son menores al valor cuando no se segmentan los datos. Existen más valores menores cuando se particiona solo por día normal y fin de semana, sin especificar que se particione también por el nombre del día. Esto no significa que cuando se particionan los datos sale siempre un porcentaje de error menor comparado al error que da cuando no se segmentan los datos.

3.4.4 RESULTADOS DE LAS PREDICCIONES UTILIZANDO EL MODELO SEE5

Una vez realizada la explicación sobre los porcentajes de error que se generan cuando el programa See5 elabora los árboles de decisión, es turno de pasar a detallar los errores obtenidos en las predicciones realizadas.

Como se mencionó anteriormente, las pruebas de predicción para días normales se realizarán para la semana del 09 de Enero del 2006 al 13 de Enero del 2006; para fin de semana en las fechas 14 de Enero del 2006 y 15 de Enero del 2006; y para feriados para Carnaval (27 y 28 de Febrero del 2006), Viernes Santo (14 de Abril del 2006). Se han escogido estas fechas puesto que se cuenta con datos reales y datos de la predicción con el Método ARIMA, para estos días mencionados y para poder comparar los resultados de la predicción utilizando el modelo de árboles de decisión.

Además, como el resultado de la predicción utilizando el modelo de árboles de decisión, es un valor que no es numérico, para poder comparar con los resultados del modelo ARIMA, se ha optado por la opción de corresponder ese valor no numérico con un valor numérico, y poder disponer de ese valor, graficarlo y compararlo.

Esta correspondencia se realiza con la marca de la clase a la que pertenece. Ya que el valor no numérico, original, que presenta como resultado el método See5, tiene un valor máximo y un valor mínimo que son los límites del intervalo;

el valor que se representa como se dijo, es la marca de clase que es valor promedio entre los valores máximo y mínimo del intervalo.

Se han realizado predicciones para los diferentes intervalos (2,3,4,5,6,8,10,15,20,25,30,35,40,50,75,100); los resultados de las comparaciones con el modelo ARIMA que se presentan en el Anexo Resultados de Predicción incluido en el cd.

3.4.4.1 Comparación de Medidas de Error para la predicción See5 - ARIMA

En la siguiente tabla se presentan los resultados obtenidos de las medidas de error de pronóstico, que se generaron en las predicciones realizadas para la semana del 09 de enero al 15 de enero del 2006, y para los feriados de Lunes y Martes de Carnaval, y para Viernes Santo; para los diferentes intervalos, tanto del modelo See5, como del modelo ARIMA:

Porcentajes de Error MAPE See5 – ARIMA

Modelo	Interv	Lun	Mar	Mier	Jue	Vier	Sáb	Dom	Lun Car	Mar Car	Vier San
See 5	2	11.81	12.12	11.93	12.32	12.42	16.4	10.57	8.36	9.31	9.22
	3	9.83	10.91	11.22	11.67	11.36	8.14	13.62	14.41	14.59	14.94
	4	5.61	4.74	4.95	4.84	4.38	7.76	6.25	8.51	8.41	8.42
	5	6.42	6.34	6.35	7.15	7.24	3.96	6.05	5.98	8.81	5.03
	6	2.99	4.17	4.48	5.35	5.03	7.2	5.69	7.01	7.42	7.58
	8	4.17	3.69	3.83	4.24	3.97	4.11	4.37	5.38	3.69	6.6
	10	3.32	4.07	4.26	4.62	4.53	3.9	3.81	3.76	8.26	9.76
	15	2.74	3.05	3.12	4.16	4.2	3.66	2.82	6.27	8.89	7.43
	20	1.63	2.6	3.37	4.18	4.26	3.54	3.83	4.95	9.55	9.31
	25	1.89	2.32	2.15	3.75	3.76	3.24	4.13	5.12	11.79	10.35
	30	2.02	2.36	2.46	3.75	4.27	3.13	3.46	10.09	9.55	10.15
	35	1.48	1.86	2.25	3.76	3.95	4.17	4.85	14.63	11.1	8.5
	40	1.85	2.41	2.4	3.83	4.12	4.74	5.53	6.37	11.1	8.5
	50	2.59	2.65	2.93	4.71	4.87	11.64	7.46	7.79	13.29	11
	75	9.29	8.18	9.3	11.29	10.9	11.12	11.2	10.68	9.26	8.68
	100	6.31	5.75	6.95	11.31	9.19	10.73	9.65	2.91	10.09	9.07
ARIMA		1.55	2.9	3.06	0.78	1.96	1.21	1.12	20.41	3.2	19.59

Tabla 28 MAPE See5 – ARIMA

Fuente: Los Autores

De lo mostrado en la tabla anterior se obtienen los siguientes gráficos para la medida de error MAPE, para los días normales Lunes 09, Martes 10, Miércoles 11, Jueves 12 y Viernes 13 de Enero del 2006:

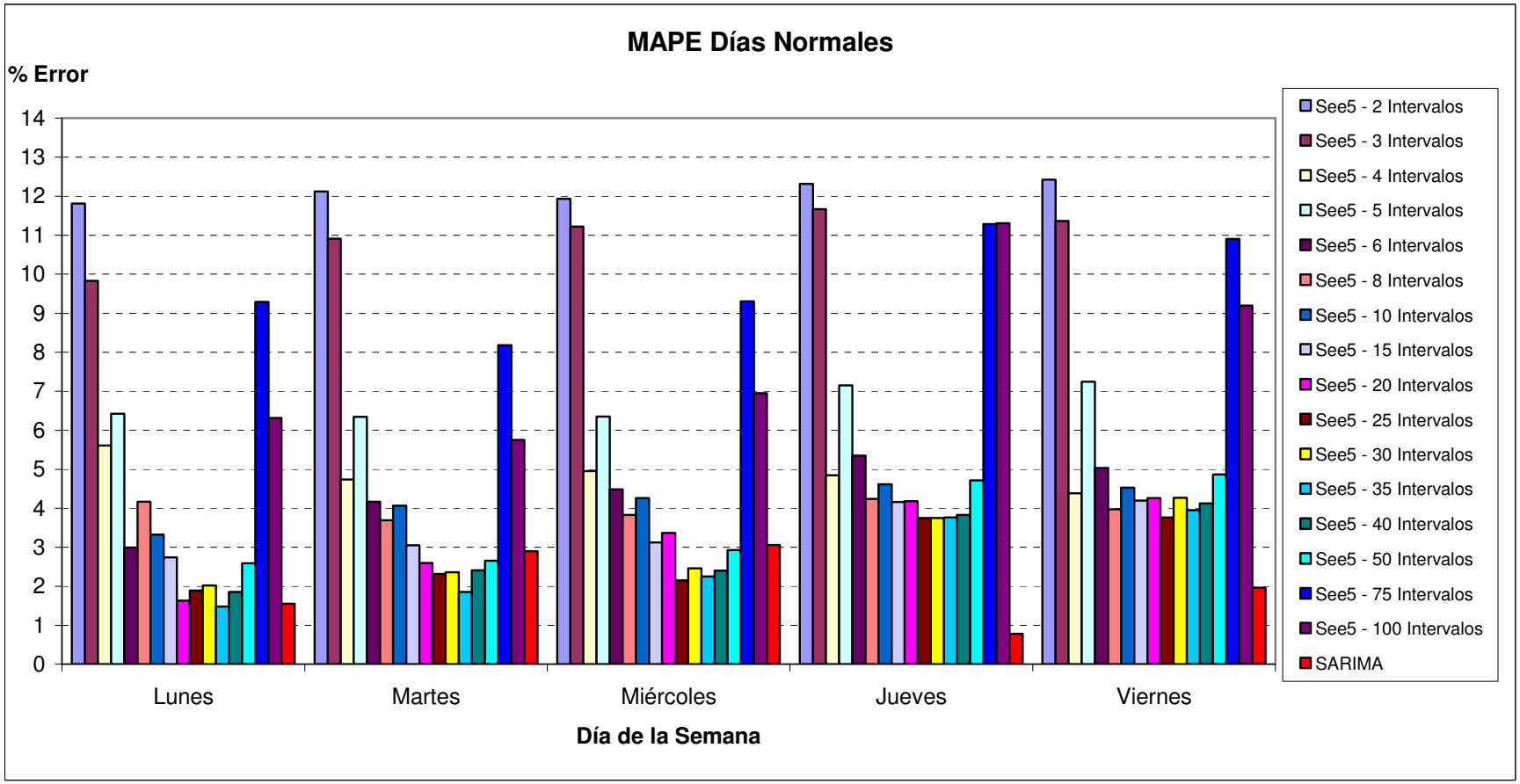


Figura 55 MAPE Días Normales

Fuente: Los Autores

Se observa que solo el MAPE cuando se utiliza el modelo de árboles de decisión con 20 intervalos es menor que el MAPE utilizando el modelo ARIMA para el día Lunes y Martes, además cuando el número de intervalos es 30 y 50, sus respectivos MAPES son menores comparados con el del Modelo ARIMA, para los días Martes y Miércoles. Para cuando el número de intervalos es 35 se observa que su respectiva MAPE es menor que la MAPE del modelo ARIMA para el día Lunes, Martes y Miércoles, como se muestra en las figuras 56, 57 y 58 respectivamente, viendo que para este caso se obtienen los mejores resultados cuando se aplica el modelo de árboles de decisión, comparado cuando se utilizan otro número de intervalos en la variable a predecir. Y para todos los demás casos el MAPE del modelo ARIMA es el menor, para los días Normales de la semana del 09 de Enero del 2006 al 13 de Enero del 2006.

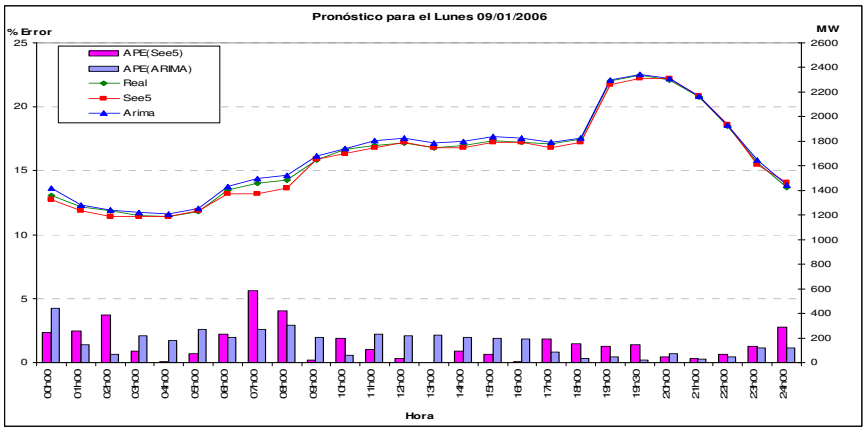


Figura 56 Pronóstico para el Lunes 09/01/2006

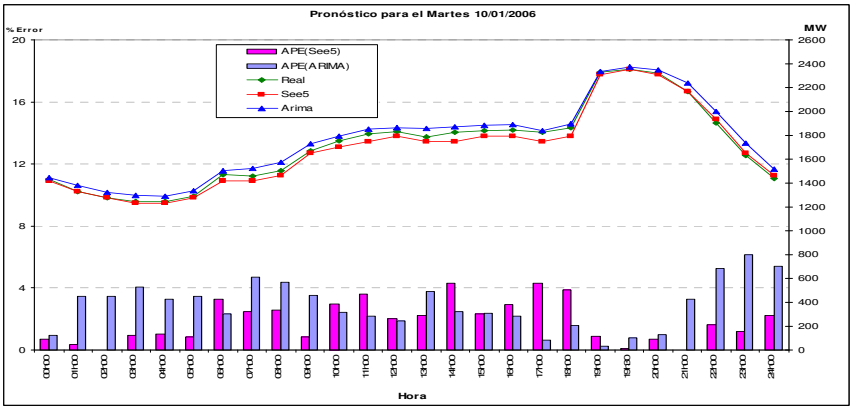


Figura 57 Pronóstico para el Martes 10/01/2006

Fuente: Los Autores

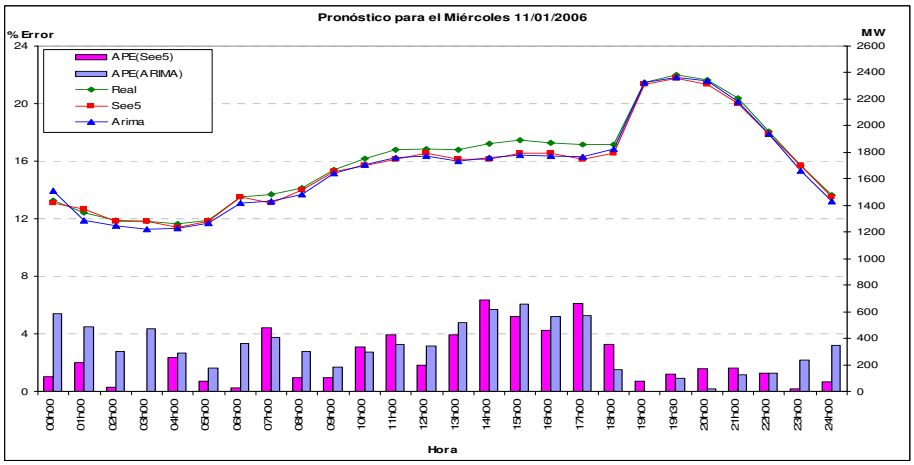


Figura 58 Pronóstico para el Miércoles 11/01/2006

Fuente: Los Autores

El gráfico de la medida de error MAPE para los días Sábado 14 y Domingo 15 de Enero del 2006, se muestra a continuación:

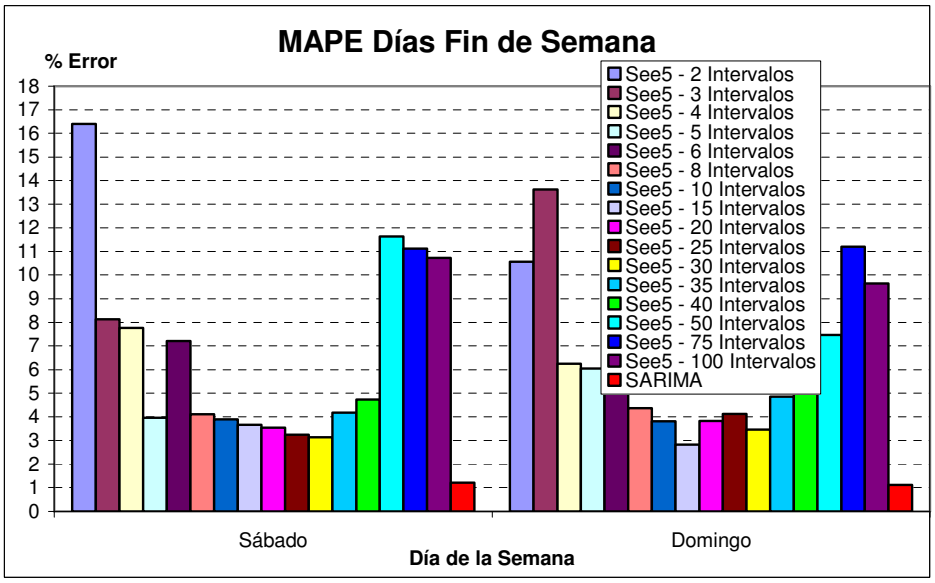


Figura 59 MAPE Días fin de semana

Fuente: Los Autores

Ningún MAPE utilizando la predicción con el modelo de árboles de decisión, es menor que el MAPE empleando el modelo ARIMA; para los días Fines de semana (Sábado y Domingo).

El gráfico de la medida de error MAPE para los días Feriados Lunes y Martes de Carnaval del 2006 y Viernes Santo del 2006, se muestra a continuación:

¡Error! No se pueden crear objetos modificando códigos de campo.

Figura 60 MAPE Días Feriados

Fuente: Los Autores

En este caso se observa que para el feriado de Lunes de Carnaval, la predicción que tuvo un resultado aceptable es cuando se utilizan 100 intervalos en la variable a predecir empleando el modelo de árboles de decisión, puesto que su MAPE es la menor comparada con los demás casos, como se muestra en la figura 61.

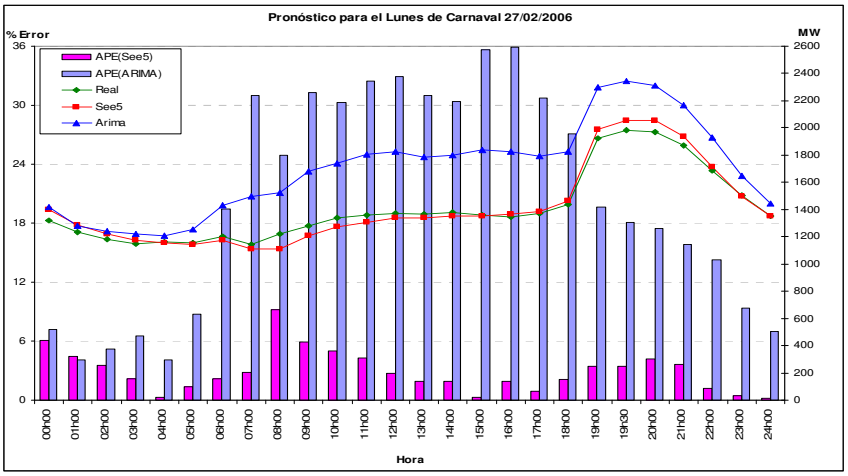


Figura 61 Pronóstico para el lunes de Carnaval

Fuente: Los Autores

Para el Martes de Carnaval la predicción que más se aproxima a lo real es cuando se utiliza el modelo ARIMA, ya que su MAPE es la menor comparada con las MAPEs de las demás predicciones, el resultado más aceptable con el modelo de árboles de decisión se obtuvo con 8 intervalos en la variable a predecir, como se muestra en la figura 62. Y para el feriado de Viernes Santo, cuando se utiliza el modelo de árboles de decisión con 5 intervalos en la variable a predecir, se nota que su MAPE es la menor para este feriado, por ende es la que más se acerca al valor real de la demanda para este día, como se muestra en la figura 63.

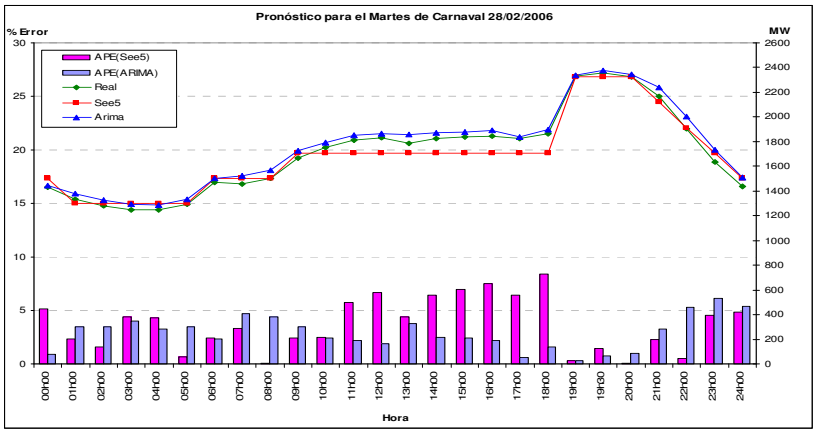


Figura 62 Pronóstico para el Martes de Carnaval

Fuente: Los Autores

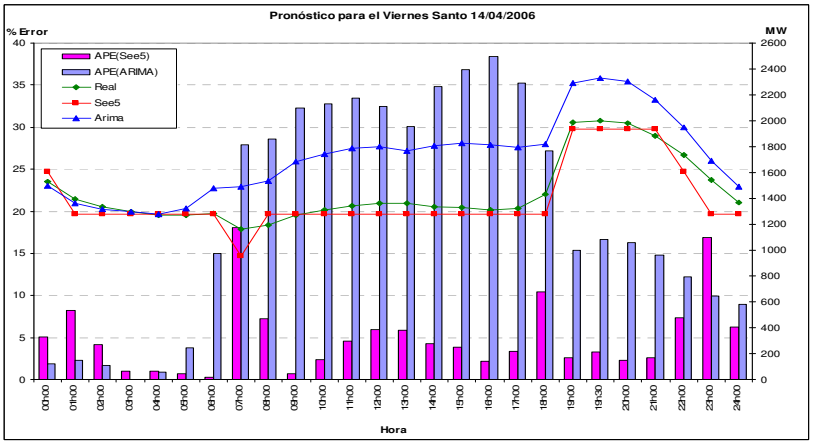


Figura 63 Pronóstico para el Viernes Santo

Fuente: Los Autores

Los resultados obtenidos para las predicciones, para los días antes mencionados, se encuentran detallados en el Anexo Resultados del modelo See5. Para la medida de Error para pronósticos, U de Theil se obtuvieron los siguientes resultados, tanto para el modelo de árboles de decisión con sus diferentes intervalos, como para el modelo ARIMA:

Mod	Interv	Lun	Mar	Miér	Jue	Vier	Sáb	Dom	Lun Carn	Mar Carn	Vier San
See 5	2	1.622	1.616	1.628	1.701	1.670	2.123	1.341	1.215	1.144	1.175
	3	1.186	1.354	1.428	1.546	1.488	1.167	1.549	1.802	1.571	1.629
	4	0.743	0.647	0.673	0.703	0.623	1.055	0.897	1.138	1.026	1.106
	5	0.839	0.914	0.977	1.083	1.067	0.628	0.842	0.727	0.977	0.694
	6	0.461	0.569	0.641	0.762	0.721	0.907	0.703	1.000	0.926	0.942
	8	0.514	0.548	0.629	0.725	0.675	0.601	0.579	0.718	0.548	0.920

	10	0.455	0.577	0.644	0.739	0.740	0.540	0.535	0.682	1.028	1.254
	15	0.361	0.437	0.498	0.659	0.631	0.539	0.437	0.952	1.055	0.987
	20	0.233	0.387	0.495	0.626	0.605	0.523	0.506	0.866	1.208	1.147
	25	0.263	0.326	0.387	0.543	0.536	0.458	0.530	0.804	1.339	1.231
	30	0.280	0.334	0.394	0.574	0.591	0.471	0.458	1.224	1.208	1.221
	35	0.212	0.284	0.380	0.565	0.558	0.590	0.626	1.871	1.334	1.063
	40	0.264	0.353	0.426	0.575	0.584	0.671	0.669	1.070	1.334	1.063
	50	0.348	0.370	0.479	0.673	0.673	1.647	0.996	1.143	1.530	1.330
	75	1.142	1.067	1.226	1.571	1.433	1.343	1.277	1.295	1.069	1.016
	100	0.999	0.879	1.163	1.694	1.388	1.316	1.341	0.418	1.144	1.058
ARIMA		0.196	0.374	0.424	0.114	0.280	0.168	0.158	2.697	0.349	2.410

Tabla 29 U Theil See5 – ARIMA

Fuente: Los Autores

De los resultados obtenidos en la tabla anterior, derivan los siguientes gráficos:

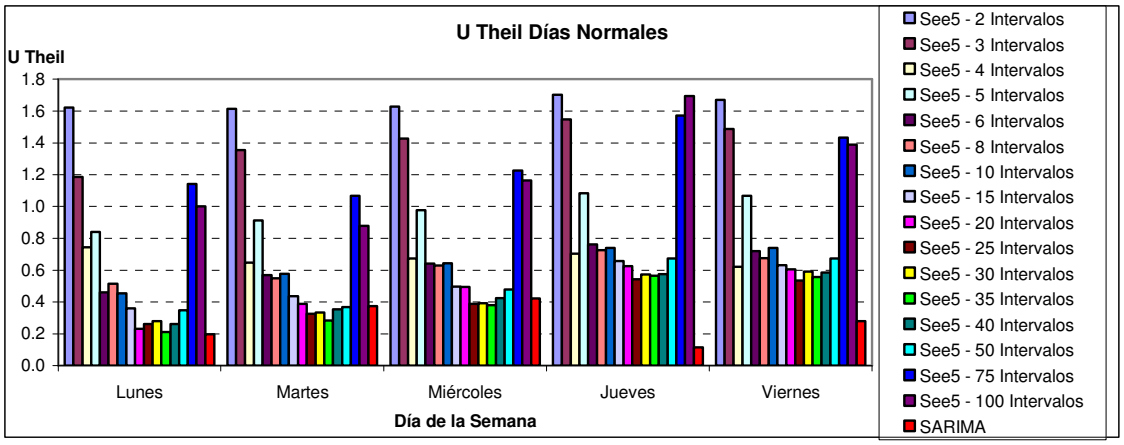


Figura 64 U Theil Días Normales

Fuente: Los Autores

Cuando se emplea el modelo ARIMA, se observa que el valor de U de Theil es el que más se aproxima a cero, por ende los resultados de esta predicción son los que más se asemejan a los valores reales de la demanda eléctrica, para los días de la semana del 09 de enero al 13 de enero del 2006. A excepción cuando el número de intervalos es 30 para los días Martes y Miércoles, además cuando se utilizan 50 intervalos para el día Martes. También se observa que cuando se utiliza un número de intervalos igual a 35 en la variable a predecir, su respectivo MAPE para el día Lunes, Martes y Miércoles es menor comparado con la MAPE del modelo ARIMA, siendo este el mejor resultado que se obtiene el la predicción de la demanda eléctrica para los días normales.

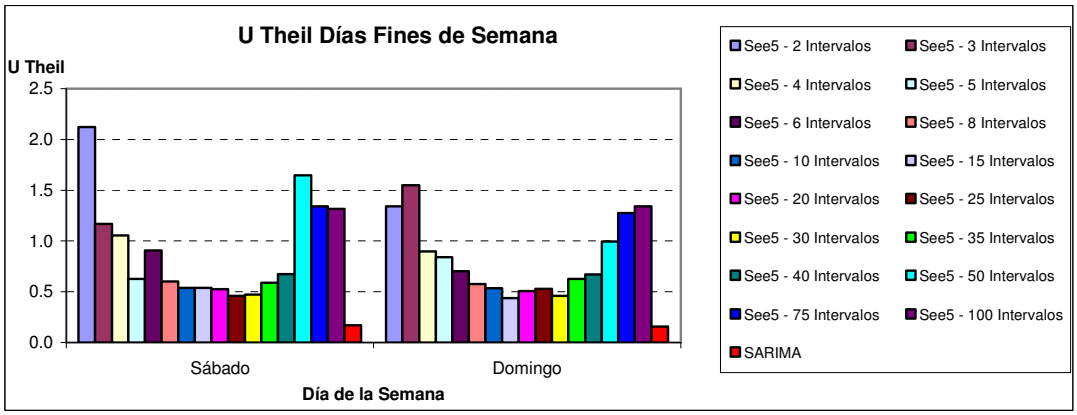


Figura 65 U Theil Días Fines de semana

Fuente: Los Autores

Igual que en el caso anterior, se puede observar que la predicción utilizando el modelo ARIMA se aproxima más a los valores reales, puesto que sus valores de U de theil son más cercanos a cero, para ambos días; Sábado 14 de Enero del 2006 y Domingo 15 de Enero del 2006.

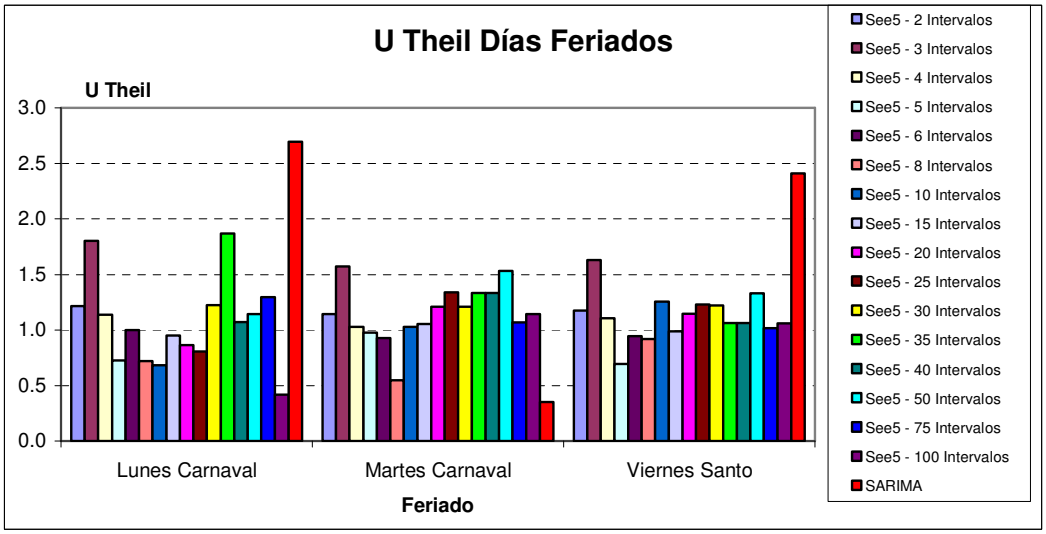


Figura 66 U Theil Días Feriados

Fuente: Los Autores

De este gráfico se puede observar que para el feriado de Martes de Carnaval del 2006, la predicción utilizando el modelo ARIMA es mejor que la utilizada con el modelo de árboles de decisión, puesto que el valor de su U de Theil es la menor. Para el feriado de Lunes de Carnaval, se observa que cuando se emplea el modelo de árboles de decisión con 100 intervalos en su variable a

predecir, los valores predichos se aproximan más a los valores reales puesto que su U de Theil es la menor. Por último, cuando se analizan los resultados obtenidos para el feriado de Viernes Santo, se observa que los valores obtenidos de predicción cuando se utiliza el modelo de árboles de decisión con 5 intervalos en la variable a predecir, es la mejor comparada con las demás para este caso, puesto que su valor de U de Theil es el más bajo.

De acuerdo con los resultados obtenidos para la predicción de la demanda eléctrica utilizando el modelo de aprendizaje de máquina See5, los mismos que se los comparó con los valores reales de la demanda eléctrica y con los resultados de la predicción del modelo ARIMA, para todas las horas del día, de las fechas seleccionadas y anteriormente mencionadas; cuando se utiliza el método de aprendizaje de máquina; los valores de error (MAPE, U Theil) en la predicción disminuyen a medida que las predicciones se generen con mayor número de Intervalos en la variable a predecir. Además, como ya se mencionó anteriormente, que el valor del porcentaje de error en la generación de los árboles de decisión aumenta conforme también aumenta el número de intervalos en la variable a predecir; se puede decir que cuando se requiere mayor precisión en los valores de la demanda eléctrica a predecir, se pierde confiabilidad en los resultados, puesto que aumenta el porcentaje de error del árbol de decisión generado para ese número de intervalo.

En comparación con los resultados del modelo ARIMA proporcionados por el CENACE, con los resultados del modelo de árboles de decisión, se observa que los primeros se aproximan más a los valores reales de la demanda eléctrica para las fechas analizadas, puesto que las medidas de error de las mismas son menores en su gran mayoría a las medidas de error generadas cuando se aplica el modelo de árboles de decisión para la predicción de la demanda eléctrica, de las mencionadas fechas.

CAPITULO 4.
CONCLUSIONES Y RECOMENDACIONES

4.1 CONCLUSIONES

- Debido a la necesidad de nuestro país por un servicio de electricidad que sea eficiente y continuo es importante tener estudios y modelos alternativos que predigan el comportamiento del valor de la demanda eléctrica, para que de ésta manera ayudar a los entes de generación, transmisión y distribución de energía eléctrica, en la toma de decisiones y evitar los temidos racionamientos de energía.
- Los modelos de aprendizaje de máquina por su flexibilidad pueden ser aplicados en diferentes áreas que tienen relación con la vida diaria de las personas, ofreciendo una manera diferente y novedosa de analizar su información, además de obtener resultados alternativos.
- Se ha realizado la descripción de algunos modelos de aprendizaje que pueden ser utilizados para la predicción de acuerdo a los requerimientos que tenga cada caso de estudio.
- El modelo de aprendizaje de máquina que se ha aplicado en el presente trabajo, ofrece una ventaja sobre las técnicas estadísticas puesto que los resultados que se obtienen pueden ser fácilmente interpretados por el usuario final.
- El algoritmo See5 basado en aprendizaje inductivo, admite dentro de su estudio tanto valores cualitativos como cuantitativos siendo esto la ventaja en casos que requieran éste tipo de atributos.
- La aplicación desarrollada ayuda a preparar los datos para la generación de archivos que son utilizados por el modelo de aprendizaje See5.
- La metodología del Proceso Unificado de Desarrollo, ha servido para cubrir el modelo orientado a objetos, para nuestro caso de estudio, con lo cual se ve que no es necesario realizar un sistema extenso para poder aplicar ésta metodología.

- La metodología que se utilizó fue adaptada a la necesidad del caso de estudio ya que se realizaron los modelos necesarios para cubrir las etapas de desarrollo de la aplicación.
- En el presente trabajo se han obtenido resultados expresados en árboles de decisión en base a datos históricos proporcionados por el CENACE para la predicción de la demanda eléctrica en el Ecuador.
- Cuando se obtiene un árbol de decisión con mucha extensión se dificulta su interpretación, esto se produce si los rangos de predicción son numerosos.
- El error de clasificación del modelo se obtiene de la proporción existente entre la cantidad de errores sobre la cantidad total de instancias clasificadas es decir muestra cuantos de ellos no pertenecen a la clase predicha.
- El error de clasificación del modelo árboles decisión tiene un comportamiento directamente proporcional con respecto al números de intervalos del Atributo de predicción es decir a menor numero de intervalos menor error de clasificación o su vez mayor número de intervalos se obtiene mayor error de clasificación.
- Es importante definir la trascendencia de los resultados, ya que se logro determinar que mientras que el error obtenido en la construcción del árbol disminuye, el error de predicción aumenta y viceversa, es decir tiene un comportamiento inversamente proporcional.
- Se han obtenido resultados de predicciones para días específicos, los mismos que han sido comparados con otros resultados del método utilizado por el CENACE en la actualidad.
- De los resultados obtenidos descritos en el capitulo3, se puede observar que el porcentaje de error, las ramas de los árboles de decisión y el número de reglas aumentan cuando se incrementan los rangos a predecir.
- Las medidas de error del modelo ARIMA poseen un menor valor en comparación del modelo de árboles de decisión con el algoritmo See5 para los casos de los días Lunes a Domingos normales, con excepción

de ciertos casos; el modelo usado en el caso de estudio es mejor para días feriados.

- Al comparar los resultados del modelo aplicado en este estudio con el modelo que usa actualmente el CENACE, los resultados obtenidos demuestran que para días feriados el modelo de árboles de decisión obtiene resultados de predicción que se aproximan más a los valores reales. En cambio para el resto de casos como son días normales y fines de semana los resultados que se obtienen cuando se utiliza el modelo de árboles de decisión difieren en un pequeño porcentaje comparados con los resultados del modelo ARIMA, siendo mejores los valores de resultado del modelo ARIMA, a excepción de pocos casos.
- Cuando se utiliza el modelo de árboles de decisión con 35 intervalos en la variable a predecir, se obtuvieron los mejores resultados en las predicciones realizadas, puesto que para el día Lunes, Martes y Miércoles se obtuvieron mejores resultados en las medidas de error para la predicción comparadas con otros números de intervalos e inclusive comparado con los resultados del modelo ARIMA.

4.2 RECOMENDACIONES

- La identificación y sobre todo el registro de nuevos factores que influyan directa o indirectamente en la demanda eléctrica es importante ya que va ha permitir tener mayor fuente de datos para un mejor análisis utilizando el modelo aplicado.
- Es importante considerar las características de los atributos del caso de aplicación de aprendizaje de máquina cuando se quiera aplicar un modelo de aprendizaje.
- Para que el porcentaje de error, en la generación de árboles de decisión y en las reglas de aprendizaje, disminuyan se recomienda seleccionar un número pequeño de rangos.
- Se recomienda utilizar el modelo de aprendizaje basado en el algoritmo See5, por su facilidad de uso y obtención de resultados, además de las ventajas que tiene ante sus predecesores.

- Es recomendable usar el modelo de árboles de decisión See5, para días feriados puesto que se aplica de mejor manera a los días atípicos.
- Si se desea aplicar una metodología de desarrollo orientada a objetos a una aplicación que no sea tan extensa, se puede utilizar el Proceso Unificado de Software PUD, ajustando ésta metodología a los modelos y requerimientos mínimos de la aplicación.

REFERENCIAS BIBLIOGRAFICAS

Libros

[1] MITCHELL, Tom, MACHINE LEARNING, Primera Edición, Editorial McGraw-Hill. United Sates. 1997.

[2] WINSTON, Patrick, INTELIGENCIA ARTIFICIAL, Segunda Edición, Editorial Addison-Wesley. United States. 1994.

[3] RUSSELL, Stuart & NORVING, Peter, ARTIFICIAL INTELLIGENCE A MODERN APPROACH, Primera edición, Editorial Prentice Hall, USA. 1995.

[4] QUINLAN , J.R. INDUCTION OF DECISION TREES. EN MACHINE LEARNING, Morgan Kaufmann, 1990.

[5] SANCHEZ MONTAÑEZ, Manuel. TEORIA Y APLICACIONES A PROBLEMAS DE PREDICCIÓN, 2005.

Internet

[6] Corporación CENACE

http://www.cenace.org.ec/, 2006.

[7] TRANSELECTRIC S.A.

http://www.transelectric.com.ec/, 2006.

[8] CONSEJO NACIONAL DE ELECTRICIDAD

http://www.transelectric.com.ec/, 2006.

[9] Árboles de Clasificación

http://www-etsi2.ugr.es/depar/ccia/rf/www/tema3_00-01_www/node25.html, 2000.

[10] See5: An Informal Tutorial

http://www.rulequest.com/see5-win.html, 2005.

[12] CONCEPTOS TEÓRICOS

http://www.cs.us.es/~delia/sia/html98-99/pag-alumnos/web2/con_teor.html, 2002.

[13] Árboles de Decisión

http://www.fing.edu.uy/inco/cursos/aprendaut/transp/arboles06.pdf, 2004.

[14] Introducción a UML

http://www.programacion.com/tutorial/uml/, 2005.

[15] Desarrollo Orientado a Objetos com UML

http://www.clikear.com/manuales/uml/index.asp, 2004

ANEXO. Glosario de Términos.

GLOSARIO	
Término	**Descripción**
ARIMA	Modelo Estadístico, Autorregresive Integrated Moving Average
ENERGÍA	La energía eléctrica, es un concepto asociado al tiempo y a la potencia nominal de una determinada carga eléctrica
DEMANDA MÁXIMA	La demanda máxima representa para un instante dado, la máxima coincidencia de cargas eléctricas operando al mismo tiempo, es decir, la demanda máxima corresponde a un valor instantáneo en el tiempo.
MEM	Mercado Eléctrico Mayorista
POTENCIA ELÉCTRICA	Es el producto de la diferencia de potencial entre los terminales y la intensidad de corriente que pasa a través del dispositivo.
SARIMA	Modelo Estadístico, Seasonal Autoregressive Integrated Moving Average
SNI	Sistema Nacional Interconectado
SPSS	Paquete de Software Estadístico.
RN	Redes Neuronales

ANEXO. Árboles De Decisión

INTRODUCCIÓN

Los árboles de decisión es uno de los métodos de aprendizaje inductivo supervisado no paramétrico más utilizado. Como forma de representación del conocimiento, los árboles de clasificación destacan por su sencillez. Su dominio de aplicación no está restringido a un ámbito concreto sino que pueden utilizarse en diversas áreas: diagnóstico médico, juegos, predicción, control de calidad, etc.

DEFINICIÓN

Un árbol de decisión es un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas.

ELEMENTOS

Un árbol de decisión tiene los siguientes elementos:

Raíz: se trata del nodo inicial en el que se encuentran todas las observaciones (datos).
Nodo: grupo de observaciones que cumplen unas condiciones determinadas.
Nodo hijo: los nodos resultantes de dividir un nodo superior
Rama: cada uno de los diferentes caminos que unen los nodos padres con los nodos hijos.
Nodo hoja: nodo sin hijos.

Un árbol de clasificación es una forma de representar el conocimiento obtenido en el proceso de aprendizaje inductivo. Puede verse como la estructura resultante de la partición recursiva del espacio de representación a partir del conjunto (numeroso) de prototipos. Esta partición recursiva se traduce en una organización jerárquica del espacio de representación que puede modelarse mediante una estructura de tipo árbol. Cada nodo interior contiene una pregunta sobre un atributo concreto (con un hijo por cada posible respuesta) y cada nodo hoja se refiere a una decisión (clasificación).

La clasificación de patrones se realiza en base a una serie de preguntas sobre los valores de sus atributos, empezado por el nodo raíz y siguiendo el camino determinado por las respuestas a las preguntas de los nodos internos, hasta llegar a un nodo hoja. La etiqueta asignada a esta hoja es la que se asignará al patrón a clasificar.

CONSTRUCCIÓN

Los árboles TDIDT, (ID3, C4.5 y C5), se construyen con el método de Hunt. Se parte de un conjunto T de datos de entrenamiento. Dadas las clases {C1, C2,. . ., Ck}, existen tres posibilidades:

1. T contiene uno o más casos, todos pertenecientes a un única clase Cj: El árbol de decisión para T es una hoja identificando la clase Cj.

2. T no contiene ningún caso: El árbol de decisión es una hoja, pero la clase asociada debe ser determinada por información que no pertenece a T. Por ejemplo, una hoja puede escogerse de acuerdo a conocimientos de base del dominio, como ser la clase mayoritaria.

3. T contiene casos pertenecientes a varias clases: Se refina T en subconjuntos de casos que tiendan hacia una colección de casos de una única clase. Se elige una prueba basada en un único atributo, que tiene uno o más resultados, mutuamente excluyentes {O1, O2,. . ., On}. T se particiona en los subconjuntos

T1, T2,. . ., Tn donde Ti contiene todos los casos de T que tienen el resultado Oi para la prueba elegida. El árbol de decisión para T consiste en un nodo de decisión identificando la prueba, con una rama para cada resultado posible. El mecanismo de construcción del árbol se aplica recursivamente a cada subconjunto de datos de entrenamientos, para que la i-ésima rama lleve al árbol de decisión construido por el subconjunto Ti de datos de entrenamiento.

Cálculo de la Ganancia de Información

Cuando los casos en un conjunto T contiene ejemplos pertenecientes a distintas clases, se realiza una prueba sobre los distintos atributos y se realiza una partición según el "mejor" atributo. Para encontrar el "mejor" atributo, se utiliza la teoría de la información, que sostiene que la información se maximiza cuando la entropía se minimiza.

Supongamos que tenemos ejemplos positivos y negativos. En este contexto la entropía de un subconjunto S_i, H(S_i), puede calcularse como:

$$H(S_i) = -p_i^+ \log p_i^+ - p_i^- \log p_i^-$$

Donde p_i^+ es la probabilidad de que un ejemplo tomado al azar de S_i sea positivo.
Esta probabilidad puede calcularse como:

$$p_i^+ = \frac{n_i^+}{n_i^+ + n_i^-}$$

Si el atributo at divide el conjunto S en los subconjuntos S_i, i = 1,2, ... , n, entonces, la entropía total del sistema de subconjuntos será:

$$H(S, at) = \sum_{i=1}^{n} P(S_i) \cdot H(S_i)$$

Donde $H(S_i)$ es la entropía del subconjunto S_i y $P(S_i)$ es la probabilidad de que un ejemplo pertenezca a (S_i). Puede calcularse, utilizando los tamaños relativos de los subconjuntos, como:

$$P(S_i) = \frac{|S_i|}{|S|}$$

La ganancia en información puede calcularse como la disminución en entropía. Es decir:

$$I(S, at) = H(S) - H(S, at)$$

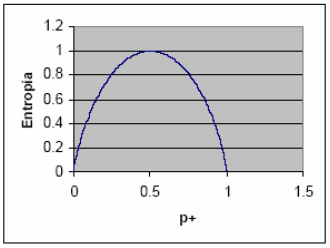
Donde $H(S)$ es el valor de la entropía a priori, antes de realizar la subdivisión, y

$H(S, at)$, es el valor de la entropía del sistema de subconjuntos generados por la partición según at.

El uso de la entropía para evaluar el mejor atributo no es el único método existente o utilizado en Aprendizaje Automático. Sin embargo, es el utilizado por Quinlan al desarrollar el ID3 , C4.5 y C.5.

Entropía

Es la cantidad de información que se espera observar cuando un evento ocurre según una distribución de probabilidades. Mide la incertidumbre dada una distribución de probabilidades. Si tomamos un conjunto con elementos positivos y negativos, la entropía variará entre 0 y 1.



Es 0 si todos los ejemplos pertenecen a la misma clase, y 1 cuando hay igual número de ejemplos positivos y negativos en el conjunto de datos. Cuanto tenemos c clases posibles, el valor máximo de la entropía será log2c.

La probabilidad de que un ejemplo tomado al azar pertenezca a la clase i y se calcule en base a la frecuencia de los datos de dicha clase en los datos de entrenamiento, se representa en la siguiente fórmula:

$$H(S_i) = \sum_{i=1}^{n} - p_i \log p_i$$

Proporción de ganancia

Favorece a los atributos que tienen muchos valores frente a los que tienen pocos valores. Si se tiene un conjunto de registros con fecha y se particiona según el campo fecha, se obtendrá un árbol perfecto, pero que no servirá para clasificar casos futuros, dato el gran tamaño del mismo. Para resolver esta situación se divide a los datos de entrenamiento en conjuntos pequeños, con lo cual tendrá una alta ganancia de información.

Una alternativa para dividir a los datos es la ganancia de información. Esta medida penaliza a los atributos como fecha al incorporar el término de información de la división:

$$I\_división(X) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} * \log_2\left(\frac{|T_i|}{|T|}\right)$$

La información de la división no es otra cosa que la entropía del conjunto con respecto al atributo i. Se define, entonces, a la proporción de ganancia como:

$$proporción\_de\_ganancia(X) = \frac{I(T,X)}{I\_división(X)}$$

La información de la división penalizará a aquellos atributos con muchos valores uniformemente distribuidos. Si tenemos n datos separados perfectamente por un atributo, la información de la división para ese caso será $\log_2 n$. En cambio, un atributo que divide a los ejemplos en dos mitades, tendrá una información de la división de 1.
Cuando la información de la división es cercana a cero, pueden aplicarse varias heurísticas. Puede utilizarse la ganancia como medida y utilizar la proporción de ganancia sólo para los atributos que estén sobre el promedio.

Datos Numéricos

Cuando los árboles de decisión se generan con atributos discretos, la partición del conjunto según el valor de un atributo es simple. Por ejemplo, agrupamos todos los animales que tengan pico, siendo tiene_pico un atributo y sus posibles valores si y no. Cuando los atributos son continuos, no es tan fácil. Por ejemplo, si queremos partir los días de un mes en función a la cantidad de lluvia caída, es casi imposible que encontremos dos días con exactamente la misma cantidad de precipitaciones caídas. Para ello se aplica binarización.
Este método consiste en formar dos rangos de valores de acuerdo al valor de un atributo, que pueden tomarse como simbólicos. Por ejemplo, si en un día hubo 100ml de lluvia, pueden crearse los intervalos [0,100) y [100, +α) y el cálculo de la entropía se realiza como si los dos intervalos fueran los dos valores simbólicos que puede tomar el atributo.

Poda de los árboles generados

Hay varias razones para podar los árboles generados por los métodos de TDIDT, la sobregeneralización, evaluación de atributos poco importantes o

significativos; y el gran tamaño del árbol. Ejemplos con ruido, atributos no relevantes, deben podarse ya que sólo agregan niveles en el árbol y no contribuyen a la ganancia de información. Si el árbol es demasiado grande, se dificulta la interpretación, con lo cual hubiera sido lo mismo utilizar un método de caja negra.

Existen dos enfoques para podar árboles: la prepoda (preprunning), detiene el crecimiento del árbol cuando la ganancia de información producida al dividir un conjunto no supera un umbral determinado y la post-poda (postprunning), se aplica sobre algunas ramas una vez que se ha terminado.

La prepoda, no pierde tiempo en construir una estructura que luego será simplificada en el árbol final, busca la mejor manera de partir el subconjunto y evaluar la partición desde el punto de vista estadístico mediante la teoría de la ganancia de información, reducción de errores, etc. Si esta evaluación es menor que un límite predeterminado, la división se descarta y el árbol para el subconjunto es simplemente la hoja más apropiada. Tiene la desventaja de que no es fácil detener un particionamiento en el momento adecuado, un límite muy alto puede terminar con la partición antes de que los beneficios de particiones subsiguientes parezcan evidentes, mientras que un límite demasiado bajo resulta en una simplificación demasiado leve.

La post-poda, es utilizada por el ID3 y el C4.5. Una vez construido el árbol se procede a su simplificación según los criterios propios de cada uno de los algoritmos.

El Principio de Longitud de Descripción Mínima

El fin de los sistemas de aprendizaje es aprender una “teoría” (árboles o reglas de decisión, por ejemplo) del dominio de los ejemplos, predictiva en el sentido de que es capaz de predecir la clase de nuevos instancias.

El Principio de Longitud de Descripción Mínima (MDL) sostiene que la mejor teoría es aquella que minimiza el tamaño y la cantidad de información necesaria para especificar las excepciones. El MDL provee una forma de medir la desempeño de los algoritmos basándose en los datos de entrenamiento únicamente. Supongamos que un sistema de aprendizaje genera una teoría T, basada en un conjunto de entrenamiento E, y requiere una cierta cantidad de

bits L[T] para codificar la teoría. Dada la teoría, el conjunto de entrenamiento puede codificarse en una cantidad L[E/T] de bits. L[E/T] está dada por la función de ganancia de información sumando todos los miembros del conjunto de entrenamiento. La longitud de descripción total de la teoría es L[E]+L[E/T]. El principio MDL recomienda la teoría T que minimiza esta suma.

Funciones alternativas

La entropía no es la única alternativa para elegir el "mejor" atributo en la partición de datos al momento de construir un árbol de decisión según el método de divide y reinarás, Existen otras medidas alternativas. Una de ellas es la función de pérdida cuadrática: dada una instancia con k clases posibles a la que puede pertenecer, el sistema aprendiz devuelve un vector de probabilidades p1, p2, ...,pk de las clases de la instancia. Es decir, p_i indica la probabilidad que tiene la instancia de pertenecer a la clase i. Con lo cual, los elementos del vector suman 1.
El resultado verdadero de la clasificación de la instancia será una de las clases posibles, entonces, si lo expresamos en un vector a1, a2, ...,ak donde ai=1 si el elemento es de clase i y es 0 en caso contrario.
Entonces, utilizamos la siguiente función para evaluar la pérdida de información según cada atributo:

Ejemplo

$$\sum_{j=1}^{k}(p_j - a_j)^2 = 1 + 2p_i + \sum_{j=1}^{k} p_j^2$$

Atributos Desconocidos

El método de Hunt, considera los resultados de todas las pruebas para todos los casos conocidos. Pero cuando los datos están incompletos, podemos tomar dos caminos posibles:

Descartar una proporción importante de los datos por incompletos y declarar algunos casos como inclasificables, o adaptar los algoritmos para poder trabajar con valores de atributos faltantes. La primera opción es inaceptable. Para la segunda opción, hay tres cuestiones importantes que deben ser tenidas en cuenta:

1. Selección de una prueba en la cual la partición del conjunto de entrenamiento se realiza en base a un criterio heurístico como ser la ganancia o la proporción de ganancia. Si dos pruebas distintas utilizan atributos con distinta cantidad de valores desconocidos, ¿cómo debe tenerse esto en cuenta al medir su importancia relativa?
2. Una vez que una prueba ha sido seleccionada, los casos de entrenamiento con valores desconocidos para los atributos relevantes no pueden ser asociados con una respuesta particular de la prueba, y, por lo tanto, no pueden asignarse a un subconjunto {Ti}.

3. Cuando el árbol de decisión se utiliza para clasificar un caso nuevo.

Transformación a Reglas de Decisión

Los árboles de decisión demasiado grandes son difíciles de entender porque cada nodo debe ser interpretado dentro del contexto fijado por las ramas anteriores. Cada prueba tiene sentido si se analiza junto con los resultados de las pruebas previas. Cada prueba tiene un contexto único. Puede ser muy difícil comprender un árbol en el cual el contexto cambia demasiado seguido al recorrerlo. Además, la estructura puede hacer que un concepto en particular quede fragmentado, lo cual hace que el árbol sea aún más difícil de entender.
Existen dos maneras de solucionar estos problemas: definir nuevos atributos que estén relacionados con las tareas o cambiar de método de representación, por ejemplo, a reglas de decisión.

En cualquier árbol de decisión, las condiciones que deben satisfacerse cuando un caso se clasifica por una hoja pueden encontrarse analizando los resultados de las pruebas en el camino recorrido desde la raíz. Es más, si el camino fuese transformado directamente en una regla de producción, dicha regla podría ser expresada como una conjunción de todas las condiciones que deben ser satisfechas para llegar a la hoja. Consecuentemente, todos los antecedentes de las reglas generadas de esta manera serían mutuamente excluyentes y exhaustivos. Al hablar de reglas de decisión o de producción nos referimos a una estructura de la forma:

$$Si\ atributo_1 = valor_x\ \ y\ atributo_2 = valor_y ... y\ atributo_n = valor_z\ \ Entonces\ clase_k$$

UTILIDAD

Los árboles de decisión se adaptan especialmente bien a ciertos tipos de problemas. Básicamente, los casos para los que son apropiados son aquellos en los que:

Los ejemplos pueden ser descritos como pares valor-atributo.

La función objetivo toma valores discretos.

Podemos tomar hipótesis con disyunciones.

Posible existencia de ruido en el conjunto de entrenamiento.

Los valores de algunos atributos en los ejemplos del conjunto de entrenamiento pueden ser desconocidos.

VENTAJAS Y DESVENTAJAS DEL USO DE ÁRBOLES DE DECISIÓN

Ventajas:

Puede ser aplicado a cualquier tipo de variables predictoras: continuas y categóricas

La regla se asignación son simples y legibles, por tanto la interpretación de resultados es directa e intuitiva.

No tiene problema de trabajar con datos perdidos.

Hace automáticamente selección de variables.

Es invariante a transformaciones de las variables predictoras.

Es computacionalmente rápido.

Es robusta frente a datos atípicos u observaciones mal etiquetadas.

Es válida sea cual fuera la naturaleza de las variables explicativas: continuas, binarias nominales u ordinales.

Es una técnica no parámetrica que tiene en cuenta las interacciones que pueden existir entre los datos.

Desventajas:

El proceso de selección de variables es sesgado hacia las variables con más valores diferentes.

La superficie de predicción no es muy suave, ya que son conjuntos de planos.

Las reglas de asignación son bastantes sensibles a pequeñas perturbaciones en los datos (inestabilidad).

Dificultad para elegir el árbol óptimo.

Los árboles de clasificación requieren un gran número de datos para asegurarse que la cantidad de las observaciones de los nodos hoja es significativa.

Algoritmo ID3

El ID3 (Induction Decision Trees), es un sistema de aprendizaje supervisado que construye árboles de decisión a partir de un conjunto de ejemplos. Estos ejemplos son tuplas compuestas por varios atributos y una única clase. El dominio de cada atributo de estas tuplas está limitado a un conjunto de valores.

Características

Desarrollado por J.Ross Quinlan en los años 80.

Pertenece a la familia TDIDT (Top-Down Induction of Decision Trees).

Maneja reglas sin incertidumbre.

Casos repetidos: Los múltiples casos repetidos no afectan al resultado.

No maneja ejemplos contradictorios. No distingue ejemplos contradictorios. El algoritmo no los va a distinguir

Objetivo: construir un árbol de decisión que explique cada instancia de la secuencia de entrada de la manera más compacta posible, según los criterios de coste y bondad. En cada momento elige el mejor atributo dependiendo de una determinada heurística.

Inconveniente: favorece indirectamente a aquellos atributos con muchos valores, los cuales no tienen que ser los más útiles.

En el ID3 se maneja una post-poda, así: si de un nodo nacen muchas ramas, las cuales terminan todas en la misma clase, entonces se reemplaza dicho nodo por una hoja con la clase común. En caso contrario, se analizan todos los nodos hijos.

Para pasar a reglas de decisión, el ID3 recorre el árbol desde la raíz hasta las hojas y genera una regla por cada camino recorrido.

El antecedente de cada regla estará compuesto por la conjunción de las pruebas de valor de cada nodo visitado, y la clase será la correspondiente a la hoja.

El recorrido del árbol se basa en el recorrido de pre-orden (de raíz a hojas, de izquierda a derecha).

Limitaciones al ID3

El ID3 puede aplicarse a cualquier conjunto de datos, siempre y cuando los atributos sean discretos. Este sistema no cuenta con la facilidad de trabajar con atributos continuos ya que analiza la entropía sobre cada uno de los valores de un atributo, por lo tanto, tomaría cada valor de un atributo continuo individualmente en el cálculo de la entropía, lo cual no es útil en muchos de los dominios.

Cuando se trabaja con atributos continuos generalmente se piensa en rangos de valores y no en valores particulares.

Pseudo código del Algoritmo IDE3

```
Función ID3
(R: conjunto de atributos no clasificadores,
C: atributo clasificador,
S: conjunto de entrenamiento) devuelve un árbol de decisión;

Comienzo

    Si S está vacío,
        Devolver un único nodo con Valor Falla;
    Si todos los registros de S tienen el mismo valor para el atributo clasificador,
        Devolver un único nodo con dicho valor;
    Si R está vacío,
        Devolver un único nodo con el valor más frecuente del atributo
        clasificador en los registros de S [Nota: habrá errores, es decir, registros
        que no estarán bien clasificados en este caso];
    Si R no está vacío,
        D ← atributo con mayor Ganancia (D,S) entre los atributos de R;
        Sean {dj | j=1,2,...., m} los valores del atributo D;
```

Sean {d_j | j=1,2,...., m} los subconjuntos de S correspondientes a los valores de d_j respectivamente;
Devolver un árbol con la raíz nombrada como D y con los arcos nombrados d_1, d_2,...., d_m que van respectivamente a los árboles.
ID3(R-{D}, C, Sl), ID3(R-{D}, C, S2), .., ID3(R-{D}, C, Sm);

Fin

Clasificación del mejor atributo

La clasificación del mejor atributo se realiza en función de los factores que se menciona a continuación:

Entropía:

Entropía de un conjunto de ejemplos D (respuesta de una clasificación):

$$Ent(D) = -\frac{|P|}{|D|} \cdot \log_2 \frac{|P|}{|D|} - \frac{|N|}{|D|} \cdot \log_2 \frac{|N|}{|D|}$$

donde P y N son, respuestas, los subconjuntos de ejemplos positivos, negativos y/o cualquier tipo de D.

Notación: $Ent([p+,n-])$, donde p = |P| y n= |N|

Intuición:

Mide la ausencia de homogeneidad de la clasificación

Teoría de la Información:

Cantidad media de información (en bits) necesaria para codificar la clasificación de un ejemplo de D.

Ejemplos

$$Ent([9+,5-]) = -\frac{9}{14}\cdot \log_2 \frac{9}{14} - \frac{5}{14}\cdot \log_2 \frac{5}{14} = 0{,}94$$

$Ent([k+,k-]) = 1$ (ausencia total de homogeneidad)

$Ent([p+,0]) = Ent([0,n-]) = 0$ (homogeneidad total)

Ganancia de información:

Preferimos nodos con menos entropía (árboles pequeños)

Entropía esperada después de usar un atributo A en el árbol:

$$\sum_{v \in Valores(A)} \frac{|D_v|}{|D|} \cdot Ent(D_v)$$

donde D_v es el subconjunto de ejemplos de D con valor del atributo A igual a v

Ganancia de información esperada después de usar un atributo A:

$$Ganancia\ (D,A) = Ent(D) - \sum_{v \in Valores(A)} \frac{|D_v|}{|D|} \cdot Ent(D_v)$$

En el algoritmo ID3, en cada nodo se usa el atributo con mayor ganancia de información (considerando los ejemplos correspondientes al nodo).

Medida del rendimiento del aprendizaje

La medida del rendimiento de aprendizaje en el IDE3 esta basado en los siguientes aspectos:

Conjunto de entrenamiento y conjunto de prueba

El algoritmo ID3 aprende con el conjunto de entrenamiento, cuya medida de rendimiento será la proporción de ejemplos bien clasificados en el conjunto de prueba.

Repetición de este proceso

Curva de aprendizaje

La estratificación, permite a cada clase correctamente representada en el entrenamiento y en la prueba

- Validación cruzada

  Dividir en k partes, y hace k aprendizajes, cada uno de ellos tomando como prueba una de las partes y entrenamiento el resto. Finalmente hacer la media de los rendimientos.
  En la práctica: validación cruzada, con k = 10 y estratificación

Sobreajuste y ruido

Una hipótesis $h \in H$ sobreajusta los ejemplos de entrenamiento si existe $h \in H$ que se ajusta peor que h a los ejemplos pero actúa mejor sobre la distribución completa de instancias.

Ruido: ejemplos incorrectamente clasificados. Causa sobreajuste.

Ejemplo: supongamos que por error, se incluye el ejemplo

< Verde, Redondo, Pequeño > como ejemplo positivo

El árbol aprendido en este caso sería (sobrejustado a los datos):

Otras causas de sobreajuste

Atributos que en los ejemplos presentan una aparente regularidad pero que no son relevantes en realidad
Conjuntos de entrenamiento pequeños

Maneras de evitar el sobreajuste

Parar el desarrollo del árbol antes de que se ajuste perfectamente a todos los datos
Podar el árbol a posteriori

Poda a posteriori, dos aproximaciones

Transformación a reglas, podado de las condiciones de las reglas
Realizar podas directamente en el árbol
Las podas se producen siempre que reduzcan el error sobre un conjunto de prueba

Poda de árboles

Un algoritmo de poda para reducir el error

1. Dividir el conjunto de ejemplos en Entrenamiento y Prueba
2. Árbol = árbol obtenido por ID3 usando Entrenamiento

3. Medida = proporción de ejemplos en Prueba correctamente clasificados por Árbol

Continuar = True

4. Mientras Continuar:

* Por cada nodo interior N de Árbol:
  - Podar temporalmente Árbol en el nodo N y sustituirlo por una hoja etiquetada con la clasificación mayoritaria en ese nodo
  - Medir la proporción de ejemplos correctamente clasificados en el conjunto de prueba.
* Sea K el nodo cuya poda produce mejor rendimiento
* Si este rendimiento es mejor que Medida, entonces Árbol = resultado de podar permanentemente Árbol en K
* Si no, Continuar = Falso

5. Devolver Árbol

Algoritmo C4.5

El C4.5 se basa en el ID3, por lo tanto, la estructura principal de ambos métodos es la misma. Ambos construyen un árbol de decisión mediante el algoritmo "divide y vencerás" y evalúa la información en cada caso utilizando los criterios de entropía y ganancia o proporción de ganancia, según sea el caso.

A diferencia del ID3, en cada nodo, el algoritmo C4.5 decide qué prueba escoge para dividir los datos. Los tres tipos de pruebas posibles propuestas por el C4.5 son:

1. La prueba "estándar" para las variables discretas, con un resultado y una rama para cada valor posible de la variable.

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

Características

Básicamente el núcleo del algoritmo es el mismo que en el ID3. Sin embargo, se han incorporado algunas características interesantes que se presentan a continuación:

- Pertenece a la familia TDIDT (Top Down Induction Decision Trees).
- Incorporación de atributos tanto discretos como continuos.
- Utilización de la ganancia proporcional (gain ratio), en caso de que la ganancia no sea conveniente como heurística en la tarea concreta que se esté estudiando.
- Método probabilístico para solucionar el problema de los atributos con valor desconocido.

El C4.5 genera un árbol de decisión a partir de los datos mediante particiones realizadas recursivamente, según la estrategia de profundidad-primero (depth-first). Antes de cada partición de datos, el algoritmo considera todas las pruebas posibles que pueden dividir el conjunto de datos y selecciona la prueba que resulta en la mayor ganancia de información o en la mayor proporción de ganancia de información. Para cada atributo discreto, se considera una prueba con n resultados, siendo n el número de valores posibles que puede tomar el atributo. Para cada atributo continuo se realiza una prueba binaria sobre cada uno de los valores que toma el atributo en los datos.

Para el caso de atributos desconocidos, el algoritmo C4.5 asume que todos los resultados de pruebas desconocidos se distribuyen probabilísticamente según la frecuencia relativa de los valores conocidos.

Un caso con un valor desconocido se divide en fragmentos cuyos pesos son proporcionales a dichas frecuencias relativas. El resultado es que un caso puede seguir múltiples caminos en el árbol. Esto se aplica cuando los casos de entrenamiento se dividen durante la construcción del árbol, como cuando el árbol se utiliza para clasificar casos.

Con respecto a la poda que se realiza a los árboles de decisión que se genera con C4.5, utiliza una post-poda que consiste en remover retrospectivamente alguna parte de la estructura construida por el particionamiento recursivo.

Así se tiene que se procesa los datos de entrenamiento libremente, y el árbol sobre-ajustado producido es podado después.

Estimación de la Proporción de Errores para los Árboles de Decisión

Después de la poda, las hojas de los árboles de decisión generados por el C4.5 tendrán dos números asociados:

N: cantidad de casos de entrenamiento cubiertos por la hoja.

E: cantidad de errores si un conjunto de N nuevos casos fueran clasificados por el árbol.

La suma de los errores predichos en las hojas, dividido para el número de casos de entrenamiento, es un estimador inmediato del error de un árbol podado sobre nuevos casos.

La versión superior al C4.5 es el See5, es un sistema que extrae patrones informativos desde los datos (Quinlan, Ross. Internet:1), utilizando para el efecto las técnicas de árboles de decisión.

Pseudo Código

Función C4.5
(R: conjunto de atributos no clasificadores,
C: atributo clasificador,
S: conjunto de entrenamiento) devuelve un árbol de decisión;

Comienzo

Si S está vacío,
Devolver un único nodo con Valor Falla;
Si todos los registros de S tienen el mismo valor para el atributo clasificador,
Devolver un único nodo con dicho valor;
Si R está vacío,
Devolver un único nodo con el valor más frecuente del atributo clasificador en los registros de S [Nota: habrá errores, es decir, registros que no estarán bien clasificados en este caso];
Si R no está vacío,
D ← atributo con mayor Proporción de Ganancia(D,S) entre los atributos de R;
Sean $\{d_j | j=1,2,...., m\}$ los valores del atributo D;
Sean $\{d_j | j=1,2,...., m\}$ los subconjuntos de S correspondientes a los valores de d_j respectivamente;
Devolver un árbol con la raíz nombrada como D y con los arcos nombrados d_1, d_2,....,d_m, que van respectivamente a los árboles C4.5(R-{D}, C, Sl), C4.5(R-{D}, C, S2), C4.5(R-{D}, C, Sm);

Fin

See5/C5.0

La versión superior al C4.5 es el C5, es un sistema que extrae patrones informativos desde los datos, utilizando para el efecto las técnicas de árboles de decisión.

El C5 incorpora una facilidad para extraer ejemplos en forma aleatoria de un conjunto de datos, construye el clasificador desde el ejemplo, y prueba el clasificador desde una colección disjunta de casos.

El algoritmo C5 permite combinar dos tipos de poda: prepoda y post poda. El proceso de prepoda se realiza durante la construcción del árbol impidiendo la ramificación de nodos que contengan un número de ejemplos inferior a una cierta constante. Además, se implementa también un método de postpoda del árbol ajustado inicialmente que consiste en sustituir una rama del árbol por una hoja, en función de una tasa de error prevista o estimada. Consideremos que existe una hoja que cubre N casos clasificando incorrectamente E de ellos, lo que se puede interpretar suponiendo que existe una variable aleatoria que

sigue una distribución binomial en la que el experimento se repite N veces obteniendo E errores. A partir de esto se estima la probabilidad de error P_e, que será la tasa de error prevista o estimada. Para ello se realiza una estimación de un intervalo de confianza para la probabilidad de error de la variable binomial y se toma como P_e el límite superior de ese intervalo (será una estimación pesimista). Entonces, para una hoja que cubra N casos, el número de errores previstos será N - P_e. Si en lugar de una hoja tenemos una rama el número de errores previstos será la suma de los de cada una de sus hojas. De este modo, una rama será sustituida por una hoja (es decir, la rama será podada) cuando el número de errores previstos de ésta sea menor que el de aquélla.

Funcionamiento

La misión del algoritmo es la elaboración de un árbol de decisión bajo las siguientes premisas:

1) Cada nodo corresponde a un atributo y cada rama al valor posible de ese atributo. Una hoja del árbol especifica el valor esperado de la decisión de acuerdo con los ejemplos dados. La explicación de una determinada decisión viene dada por la trayectoria desde la raíz a la hoja representativa de esa decisión.

2) A cada nodo se le asocia aquel atributo más informativo que aún no haya sido considerado en la trayectoria desde la raíz.

3) Para medir el nivel informativo de un atributo se emplea el concepto de entropía. Cuanto menor sea el valor de la entropía, menor será la incertidumbre y más útil será el atributo para la clasificación.

Como se explica en Quinlan, en las versiones iniciales de este algoritmo se usaba un criterio denominado ganancia para elegir el atributo (variable) en base al cual hacer cada partición de las que forman el árbol. Muy brevemente, la idea es la siguiente: la información que contiene un mensaje depende de su probabilidad y puede ser medida en bits como menos el logaritmo en base 2 de

esa probabilidad. Por ejemplo, si tenemos 8 mensajes equiprobables, la información contenida en cualquiera de ellos $-\log_2\left(\frac{1}{8}\right)$ o 3 bits.

Imaginemos que seleccionamos aleatoriamente un caso de entre un conjunto T de casos, sabiendo que pertenece a alguna clase C_j. La probabilidad de este mensaje es $\frac{freq(Cj,T)}{|T|}$, y la información que contiene es $-\log_2\left(\frac{freq(Cj,T)}{|T|}\right)$ bits, donde |T| denota el número de casos u observaciones que contiene el conjunto T. Por tanto, la cantidad media de información necesaria para identificar la clase de entre k clases posibles a la que pertenece un caso en el conjunto T (o entropía del conjunto T) viene dada por la expresión:

$$info(T) = -\sum_{j=1}^{k} \frac{freq(C_j,T)}{|T|} \times \log_2\left(\frac{freq(C_j,T)}{|T|}\right) \text{ bits.}$$

Si ahora dividimos este conjunto de acuerdo con los valores que tome un atributo X, entonces para clasificar cada caso necesitaremos una cantidad de información menor que la anterior. Esta cantidad vendrá dada por:

$$info_X(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \times info(T_i),$$

Donde T_i es cada una de las particiones hechas considerando el atributo X y T_i el número de observaciones que contiene cada una de dichas particiones. La magnitud

$$gain(X) = info(T) - info_X(T)$$

mide la cantidad de información que se gana dividiendo el conjunto de datos T de acuerdo con el atributo X. Entonces, el criterio de ganancia selecciona para hacer la partición aquel atributo para el cual se maximiza la ganancia de información.

Posteriormente, Quinlan observó que este criterio favorecía a aquellos atributos

con un número mayor de valores posibles, con lo que no clasificaba correctamente los ejemplos dados cuando alguno de los atributos era una variable continua. Para evitar este sesgo que favorece a los atributos con muchos valores posibles tomó la magnitud:

$$split\ info(X) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} \times \log_2\left(\frac{|T_i|}{|T|}\right),$$

que representa la entropía del conjunto T cuando es dividido de acuerdo con los valores que toma el atributo X. Esta entropía será tanto mayor cuanto más elevado sea el número de dichos valores. De este modo, puede ser utilizado como divisor de gain (X) para corregir los elevados valores que esta magnitud tomará para aquellos atributos que adopten un mayor número de valores posibles. Entonces, el atributo elegido para la partición será aquél para el cual el ratio de ganancia sea mayor, definiéndose esta medida como:

$$gain\ ratio(X) = \frac{gain(X)}{split\ info(X)}$$

El árbol construido por aplicación reiterada de este criterio consta del mínimo número de atributos (variables) que se requieren para la clasificación correcta

de los ejemplos dados, con lo que es claro el gran poder explicativo de esta técnica. También se pueden elaborar, a partir del árbol, una serie de reglas de clasificación más simples y fácilmente interpretables, que definen las características que más diferencian a las distintas categorías establecidas inicialmente.

Para interpretar correctamente la información facilitada, ha de tenerse en cuenta lo siguiente:

Cada regla predice la pertenencia a una de las agrupaciones.

La aplicación de una regla requiere necesariamente el cumplimiento de la condición o condiciones establecidas en la misma.

Para las opciones que no cumplan ninguna de las reglas está establecida una clasificación por defecto, también calculada a través del algoritmo, para lo cual sigue el criterio de minimizar los errores de clasificación.

See5, tiene la posibilidad de construir los clasificadores a ser aplicados a la generación del árbol de decisión. Entre las principales opciones de clasificación tenemos:

Boosting

La idea de seleccionar ésta opción de clasificador es generar varios clasificadores (árboles de decisión ó conjunto de reglas de decisión) en vez de solo uno, Cuando un nuevo caso es clasificado, cada clasificador vota por la clase predicha y los votos son contados para determinar la clase final.

Cuando se van generando los clasificadores uno tras otro, el segundo clasificador será diferente del primero y arrojará otros resultados y otros errores, este proceso continuará hasta que el más reciente clasificador sea extremadamente preciso o extremadamente impreciso. Este es un procedimiento que incrementa la habilidad de generalización del algoritmo.

Este mecanismo denominado “boosting” reduce la tasa de error para los casos de prueba en alrededor del 25%.

Sort by Utility

Ésta opción permite ordenar las reglas por clase y por nivel de confianza , la regla que más reduce la tasa de error aparece primero y la regla que menos contribuye aparece al final.

Winnow

Ésta opción permite observar los atributos utilizados dentro de la clasificación del árbol de decisión, puesto que por lo general no se utilizan todos los atributos.

Croos Validate

Ésta opción permite determinar con anterioridad una cierta cantidad particiones de los datos. En el caso en que la posibilidad de que los datos utilizados para entrenamiento y prueba no sean representativos de los datos sobre los que se utilizará el modelo posteriormente. Así, si se utiliza tres, es decir, los datos se dividen al azar en tres particiones de aproximadamente la misma cantidad, y cada una a su turno se utiliza para prueba mientras que las otras dos se utilizan para entrenamiento. Por lo tanto, se utiliza un tercio para prueba y dos tercios para entrenamiento, y se repite el procedimiento tres veces. Las tres proporciones de error obtenidas se promedian para llegar a una proporción de error general. Este procedimiento conocido como validación cruzada de tres pliegues, puede trabajar con datos estratificados, en cuyo caso sería validación cruzada de tres pliegues estratificada.

Características que deben cumplir los datos para aplicar el See5

El aprendizaje por medio de árboles de decisión debe cumplir las siguientes características de entrada.

Representación por pares de atributo-valor: Las instancias deben representarse por un conjunto fijo de atributos y los valores respectivos; éste último puede describirse mediante valores disjuntos ó también puede tener valores numéricos. Cada atributo puede ser discreto o numérico, pero los atributos utilizados para describir un caso no pueden variar de un caso a otro. Esto

restringe los dominios de aplicación en los cuales los objetos tienen inherentemente atributos variables. El hecho de que los atributos no puedan variar de un caso a otro, no restringe aquellos casos en los cuales los valores de algunos atributos son desconocidos. Ejemplo:

Atributo: Temperatura

Valores de Atributo: Alta, Media, Baja.

Clases predefinidas: las categorías a las cuales se asignan los casos deben estar establecidas de antemano. Esto significa que los algoritmos se aplican sobre un conjunto de datos de entrenamiento previamente clasificados, del tipo {valor_atributo1, valor_atributo2, ..., valor_atributon, clasek}.

Clases discretas y disjuntas: las clases a las cuales se asignan los casos deben ser totalmente disjuntas: un caso pertenece o no pertenece a una clase, pero no puede pertenecer a dos clases a la vez. Además, deben existir muchos más casos que clases para que el modelo generado sea válido en el dominio analizado. Por otro lado, dado la naturaleza de los árboles de decisión, las clases deben ser discretas o discretizarse en caso de ser continuas.

Datos suficientes: los patrones dados por la generalización inductiva no serán válidos si no se los pueden distinguir de las casualidades. Como ésta diferenciación se basa generalmente en pruebas, deben existir casos suficientes para que dichas pruebas sean efectivas. La cantidad de datos requeridos está afectada por factores como la cantidad de propiedades y clases, y la complejidad del modelo de clasificación; a medida que estos se incrementan, se necesitan más datos para construir un modelo confiable.

Los datos de entrenamiento pueden contener errores: según Mitchell, los métodos de aprendizaje utilizando árboles de decisión son robustos frente a los errores, tanto en los valores de las clases como en los ejemplos de entrenamiento

Los datos de entrenamiento pueden contener valores de atributos faltantes: Los métodos que utilizan árboles de decisión pueden utilizarse aún cuando no se conocen todos los valores de todos los atributos de los datos de entrenamiento. El tratamiento de valores faltantes varía de un algoritmo a otro, éste tratamiento ha sido explicado para los algoritmos propuestos en el presente trabajo.

Modelos lógicos generados: los programas sólo construyen clasificadores que pueden ser expresados como árboles de decisión o como un conjunto de reglas de producción. Estos modelos restringen las descripciones de clases a una expresión lógica cuyas primitivas son afirmaciones acerca de los valores de atributos particulares. La expresión lógica representada por un árbol de decisión es una disyunción de conjunciones.

ENTRADA Y SALIDA DE DATOS DE C5

C5 recibe las siguientes entradas:

Archivo .names: Enumeración de las posibles clases, y nombre y tipo de los atributos.

Existen dos importantes clases de atributos:

Atributos explícitamente definidos, que son aquellos que están dados directamente en el dato.

Atributos implícitamente definidos, son aquellos que están definidos por una fórmula.

Éste archivo debe estar bajo el siguiente formato:

Opc1_ atrib_result, opción2_ atrib_result, opción_n_ atrib_result. (Aquí se define las opciones del atributo resultado)

Nombre_atributo1: opc1_atributo1, opc2_atributo1, opc_n_atributo1.

Nombre_atributo2: opc1_atributo2, opc2_atributo2, opc_n_atributo1.

Nombre_atributo_n: opc1_atributo_n, opc2_atributo_n, opc_n_atributo_n.

Archivo .data: Conjunto de entrenamiento (conjunto de casos clasificados).

Éste archivo debe estar bajo el siguiente formato:

Val_opc1_atributo1, val_opc1_atributo2, …., val_Opc1_ atrib_result.

Val_opc2_atributo1, val_opc2_atributo2, …., val_Opc2_ atrib_result.

Val_opc_n_atributo1, val_opc_n_atributo2, …., val_Opc_n_ atrib_result.

Observaciones:

Cada valor de las opciones de los respectivos atributos deberán estar separados con comas y al terminar, la respectiva fila de valores deberá estar terminada con punto.

El último valor de la fila, corresponde al valor de la opción del atributo que es tomado como resultado.

C5 genera la siguiente salida:

Árbol de decisión.

Reglas de aprendizaje.

Para cada nodo hoja, el número de casos del conjunto de entrenamiento T que clasifica correctamente e incorrectamente.

Evaluación del árbol de decisión sobre T: tamaño (número de nodos, tanto internos como hojas), número de errores de clasificación y tasa de errores aparente (en %).

Evaluación del árbol de decisión podado: tamaño, número de errores de clasificación sobre T, tasa de errores aparente (en %) sobre T, tasa de errores pesimista (en %) estimada probabilísticamente.

Opcionalmente, el árbol de decisión podado.

ANEXO. Cálculo de los Intervalos de la variable Predicción

Intervalos de Clase

Definimos como intervalos de clase a las subdivisiones o intervalos en que se divide el dominio o campo de variabilidad de la variable, de modo tal que cada intervalo esté compuesto por tramos del recorrido de la misma.

Rango o recorrido (R)

Es la diferencia entre el valor más alto y el más bajo observado

R = xmáx - xmin

Límites de Clase

Llamamos **límites de clase** a los valores que definen los extremos de un intervalo.
Por ejemplo: el intervalo 0 a 10 años, tiene como límites a los valores 0 y 10.

Número de clases o intervalos. (K)

Se dispone de varias reglas para plantearlas, pero en general, el sentido común basta para tomar esa decisión y también los lineamientos siguientes ayudan:

a. Para la mayor parte de las aplicaciones; la experiencia ha mostrado que entre seis y catorce clases es un número adecuado para proporcionar información suficiente sin excederse en detalles.

b. El número de clases deberá ser suficiente para mostrar la forma de distribución pero no excesivo para registrar demasiadas fluctuaciones menores.

Generalmente se aconseja que las tablas de frecuencias tengan entre 6 y 14 intervalos de clase, de modo que no haya tantos como para que no sea manejable la tabla, ni tan pocos como para que la amplitud sea tan grande que haga perder información.

Cuando los datos se agrupan en intervalos, el problema fundamental es pensar en una amplitud adecuada para los mismos. La amplitud de los mismos está vinculada estrechamente con la cantidad de intervalos considerados.

Amplitud del Intervalo (A)

La amplitud del Intervalo se obtiene dividiendo el Rango (*R*) entre la Amplitud del intervalo (*K*), es decir:

$$A = \frac{R}{K}$$

Puede ocurrir que se necesite la información agrupada en intervalos con una amplitud determinada; en este caso, conociendo la amplitud, se divide el rango y se obtiene la cantidad de intervalos.

Marca de Clase

Cada intervalo tendrá también lo que se llama **marca de clase**, que es el punto medio del mismo.

$$X_j = \frac{L_i + L_s}{2}$$

FRECUENCIAS

Frecuencia absoluta

Llamaremos así al número de repeticiones que presenta una observación. Se representa por n_i.

Frecuencia relativa

Es la frecuencia absoluta dividida por el número total de datos, se suele expresar en tanto por uno, siendo su valor **-iésimo** $\boxed{f_i = \frac{n_i}{n}}$**.**

La suma de todas las frecuencias relativas, siempre debe ser igual a la unidad.

Frecuencia absoluta acumulada

Es la suma de los distintos valores de la frecuencia absoluta tomando como referencia un individuo dado. La última frecuencia absoluta acumulada es igual al nº de casos:

$N_1 = n_1$

$N_2 = n_1 + n_2$

$N_n = n_1 + n_2 + ... + n_{n-1} + n_n = n$

Frecuencia relativa acumulada

Es el resultado de dividir cada frecuencia absoluta acumulada por el número total de datos, se la suele representar con la notación: F_i

De igual forma, también se puede definir a partir de la frecuencia relativa, como suma de los distintos valores de la frecuencia relativa, tomando como referencia un individuo dado. La última frecuencia relativa acumulada es igual a la unidad.

Cálculo de los Intervalos de Clase

Debido a que el Valor de la Demanda Eléctrica (valorDemanda), es una variable continua, además de que toma muchos valores diferentes, se ve la necesidad de agrupar estos datos en intervalos, para obtener un resumen efectivo de la información original, así como también para establecer los rangos para su posterior uso, y para que de esta manera se pueda utilizar de manera más clara los datos almacenados en la Base de Datos. Toda esta información de intervalos de clase se la presenta en las tablas de frecuencias para datos agrupados.

Hay que tomar en cuenta que al formar las clases no debemos dejar fuera a ningún valor de la variable. Así, de esta manera se procede a formar los diferentes intervalos para esta variable.

Cálculo del Rango (*R***)**

Para el cálculo del Rango R, se debe conocer el valor máximo de la demanda y el valor mínimo de ésta.

$$R = Máximo - Mínimo$$

Aplicando la fórmula para los valores de la demanda se tiene:

$$R = 2424.1 - 790.2 = 1633.9$$

Cálculo del número de Clases o Intervalos (*K* **)**

Pueden calcular el número de clases mediante el método:

K = *Ln*(*n*) / *Ln*(2), en donde *K* es el número de clases (intervalos);
Ln , es el logaritmo neperiano y *n* es el número de datos por agrupar

Este es solo un método para el cálculo de los intervalos. Para este estudio se trabajarán con diferente número de intervalos, para observar el efecto que tiene el seleccionar cada uno de estos. Específicamente se trabajarán con 2, 3, 4, 5, 6, 8, 10 y 15 intervalos, para observar el efecto dentro de las predicciones realizadas.
A continuación se detalla un ejemplo para calcular los intervalos en la variable a predecir y para demostrar la manera de cómo se obtienen los intervalos:

Si se aplica la fórmula para calcular el número de intervalos tenemos:

$$K = \frac{Ln(n)}{Ln(2)}$$

En el caso de estudio tenemos 27196 valores de demanda eléctrica, distribuidos todos estos para las diferentes horas del día, es decir desde las 00h00 hasta las 24h00 incluyendo las 19h30. Es por esta razón que el número de 27196 datos se lo debe dividir para 26, puesto que este es el número de horas registradas. Es resultado de esta división es de 1046.

Este número representa el número de datos a agrupar, y aplicando la fórmula:

$$K = \frac{Ln(1046)}{Ln(2)} = 10.03$$

Al redondear este valor se tiene el valor de 10.

Entonces K = 10, el número de intervalos para este caso es de 10 intervalos.

Para la mayor parte de las aplicaciones; la experiencia ha mostrado que entre seis y catorce clases es un número adecuado para proporcionar información suficiente sin excederse en detalles, por tal razón se puede escoger el número de intervalos para trabajar.

Cálculo de la Amplitud de Intervalo (*A***).**

La amplitud del Intervalo se obtiene dividiendo el Rango (*R*) entre la Amplitud del intervalo (*K*), es decir:

$$A = \frac{R}{K}$$

Aplicando la fórmula al caso de estudio se tiene:

$$A = \frac{1633.9}{10} = 163.39$$

Se redondea este número para saber la amplitud y tenemos:

$$A = 163.4$$

Construcción de los Intervalos de Clase

El primer intervalo resulta de la suma de la amplitud del intervalo al valor mínimo de todos los datos:

Primer Intervalo: 790.2 – 953.6.

De esta manera obtenemos el resto de intervalos:

Intervalo	Valor Inicial	Valor Final
1	790.2	953.6
2	953.6	1117
3	1117	1280.4
4	1280.4	1443.8
5	1443.8	1607.2
6	1607.2	1770.6
7	1770.6	1934
8	1934	2097.4
9	2097.4	2260.8
10	2260.8	2424.2

Cálculo de las Marcas de clase o Puntos Medios de clase (X_j).

Se calcula las marcas de cada clase con la siguiente fórmula:

$$X_j = \frac{L_i + L_s}{2}$$

Por ejemplo para la primera clase se tiene lo siguiente:

$$X_j = \frac{790.2 + 953.6}{2} = 871.9$$

Las marcas de clase para los intervalos calculados son las siguientes:

Intervalo		Marca de Clase
Valor Inicial	Valor Final	X_j
790.2	953.6	871.9
953.6	1117	1035.3
1117	1280.4	1198.7
1280.4	1443.8	1362.1
1443.8	1607.2	1525.5
1607.2	1770.6	1688.9
1770.6	1934	1852.3
1934	2097.4	2015.7
2097.4	2260.8	2179.1
2260.8	2424.2	2342.5

Cálculo de las Frecuencias Absolutas : (n_i).

La frecuencia absoluta es el número total de veces que se repite cierto dato. Para esto se ve en la base de datos el número de veces que se repite el valor de la demanda para los intervalos obtenidos.

Intervalo	Marca de Clase X_j	Frecuencia Absoluta n_i
790.2 - 953.6	871.9	67
953.6 - 1117	1035.3	2658
1117 - 1280.4	1198.7	6197
1280.4 -1443.8	1362.1	4718
1443.8 - 1607.2	1525.5	4769
1607.2 - 1770.6	1688.9	3818
1770.6 - 1934	1852.3	1769
1934 - 2097.4	2015.7	1594
2097.4 - 2260.8	2179.1	1150
2260.8 - 2424.2	2342.5	456

Cálculo de las Frecuencias Relativas (f_i).

Las frecuencias relativas (f_i) se calculan dividiendo cada una de las frecuencias absolutas (n_i) entre el número total de observaciones, (N), pudiéndose expresar en forma de una fracción o de un porcentaje.

Para este caso, hasta el momento tenemos que N = 27196 registros. Y se obtiene la siguiente tabla de frecuencias Relativas:

Intervalos	Marca de clase X_j	Frecuencia Absoluta n_i	Frecuencia Relativa f_i
790.2 - 953.6	871.9	67	0.25
953.6 - 1117	1035.3	2658	9.77
1117 - 1280.4	1198.7	6197	22.79
1280.4 -1443.8	1362.1	4718	17.35
1443.8 - 1607.2	1525.5	4769	17.54
1607.2 - 1770.6	1688.9	3818	14.04
1770.6 - 1934	1852.3	1769	6.50
1934 - 2097.4	2015.7	1594	5.86
2097.4 - 2260.8	2179.1	1150	4.23
2260.8 - 2424.2	2342.5	456	1.68

Acumular las Frecuencias Absolutas (N_i), Calcular las Frecuencias Relativas Acumuladas (F_i).

La Tabla con las frecuencias absolutas acumuladas (N_i) y las frecuencias Relativas Acumuladas (F_i) es la siguiente:

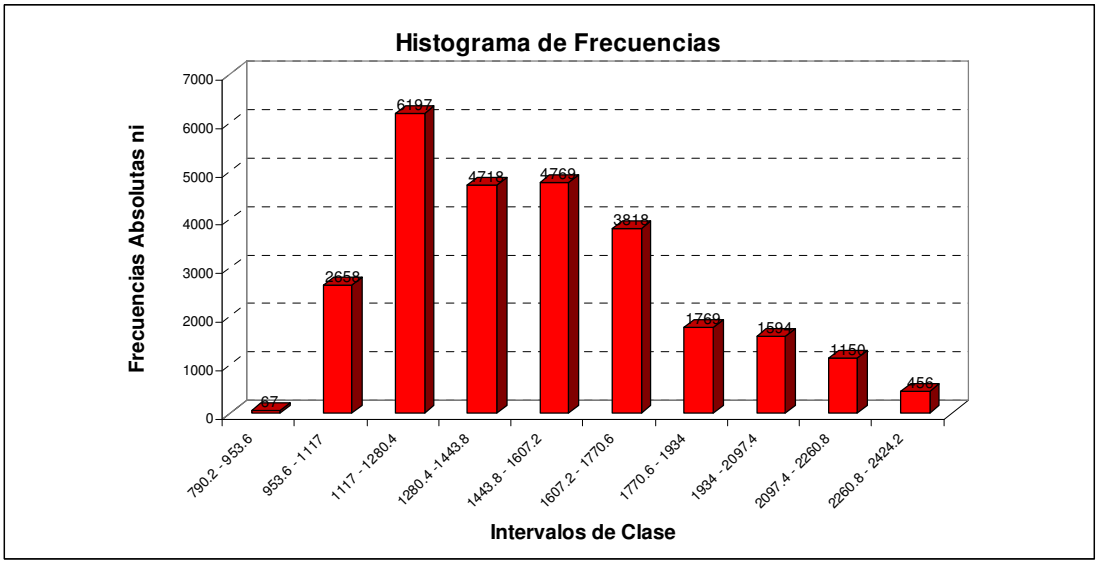
Intervalos	Marca de clase X_j	Frecuencia Absoluta n_i	Frecuencia Relativa f_i	Frecuencia Absoluta Acumulada N_i	Frecuencia Relativa Acumulada F_i
790.2 - 953.6	871.9	67	0.25	67	0.25
953.6 - 1117	1035.3	2658	9.77	2725	10.02
1117 - 1280.4	1198.7	6197	22.79	8922	32.81
1280.4 -1443.8	1362.1	4718	17.35	13640	50.15
1443.8 - 1607.2	1525.5	4769	17.54	18409	67.69
1607.2 - 1770.6	1688.9	3818	14.04	22227	81.73
1770.6 - 1934	1852.3	1769	6.50	23996	88.23
1934 - 2097.4	2015.7	1594	5.86	25590	94.09
2097.4 - 2260.8	2179.1	1150	4.23	26740	98.32
2260.8 - 2424.2	2342.5	456	1.68	27196	100.00

A continuación se presentan los diferentes gráficos:

Histograma de Frecuencias

Un histograma de frecuencias es un gráfico que se forma levantando rectángulos sobre cada uno de los Límites de cada intervalo, con una altura equivalente a la frecuencia absoluta de cada clase.

El histograma se utiliza para representar datos que corresponden a los valores de una variable cuantitativa continua.

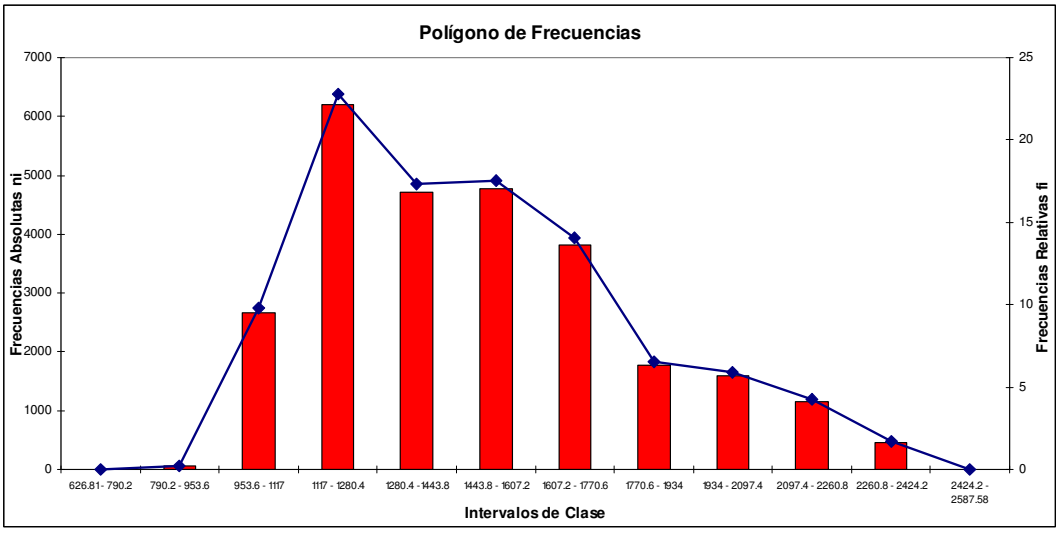


Histograma y Polígono de Frecuencias

Un polígono de frecuencias es sólo una línea que conecta los Puntos Medios de todas las barras de un histograma.

En el polígono de frecuencia como en el histograma, el valor de la variable aparece en el eje horizontal y la frecuencia absoluta o relativa en el eje vertical.

La diferencia con respecto al histograma es que el polígono sólo toma en consideración los ***Puntos medios de clase*** como representativo de cada clase o intervalo.



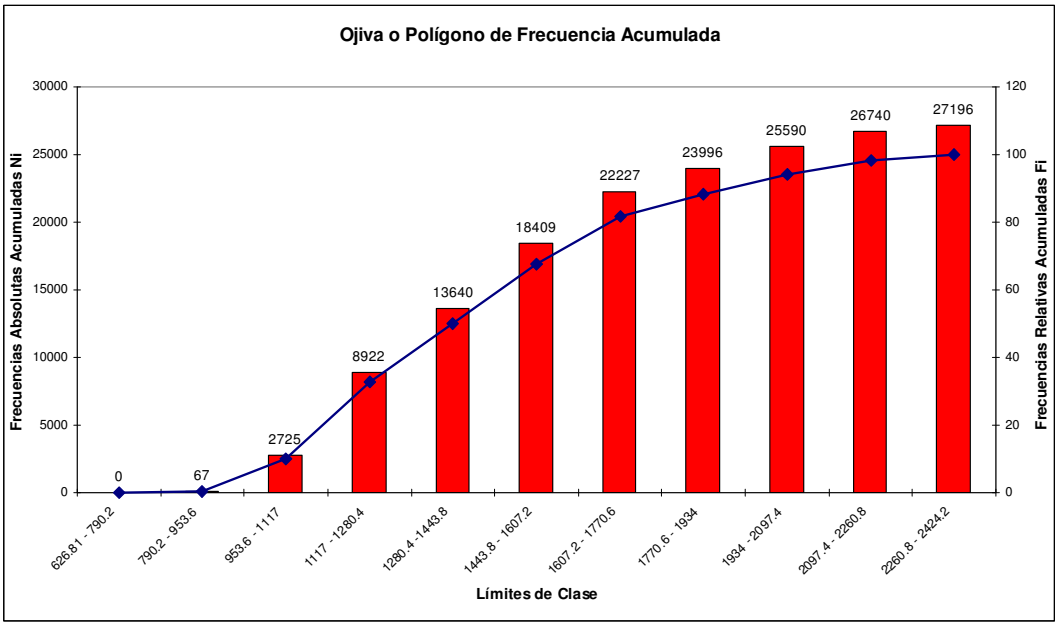
Ojiva o Polígono de Frecuencia Acumulada

La gráfica de una distribución de frecuencias acumuladas se conoce como *ojiva*

Una distribución de frecuencias acumuladas nos permite ver cuántas observaciones son ***menores o igual a un valor específico***, en lugar de hacer un mero registro del número de elementos que hay dentro de los intervalos.

El intervalo (o el límite superior del intervalo) aparece en el eje horizontal y la frecuencia absoluta acumulada o relativa acumulada en el eje vertical.

Esta gráfica facilita la comparación dos grupos de datos de forma visual y de manera mucho más efectiva que el polígono de frecuencia, puesto que permite comparar los porcentajes acumulados de dos distribuciones con respecto al mismo intervalo.


Ojiva o Polígono de Frecuencia Acumulada
Frecuencias Absolutas Acumuladas Ni
Frecuencias Relativas Acumuladas Fi
Límites de Clase
30000
25000
20000
15000
10000
5000
0
120
100
80
60
40
20
0
0
67
2725
8922
13640
18409
22227
23996
25590
26740
27196
626.81 - 790.2
790.2 - 953.6
953.6 - 1117
1117 - 1280.4
1280.4 -1443.8
1443.8 - 1607.2
1607.2 - 1770.6
1770.6 - 1934
1934 - 2097.4
2097.4 - 2260.8
2260.8 - 2424.2

ANEXO. Medidas De Error Para La Predicción

En general, la evaluación de una predicción o de un pronóstico en un determinado periodo se utiliza para medir el grado de bondad de la predicción realizada sobre la variable de interés en comparación con el valor realmente observado para tal variable en dicho período. Es frecuente hablar de la *exactitud* de una predicción como sinónimo de evaluación de la misma. Cuando una predicción muestra menor error predictivo que otra se acostumbra a decir que es *más exacto,* permitiéndose esta licencia en el empleo del término *exactitud en* el contexto de la predicción.

Es conveniente tener en cuenta que, en cualquier modelo, su eventual buen grado de ajuste respecto a periodos históricos no garantiza su buena *capacidad predictiva* respecto al futuro. Para obtener la predicción puntual de la variable dependiente en un periodo futuro se debe conocer el valor de las variables independientes en dicho período. Para evaluar, además, la capacidad predictiva del modelo se ha de conocer también el valor observado de la variable dependiente en tal período

La cuestión de cual es la mejor forma de medir la exactitud de una predicción ha sido ampliamente debatida en la literatura especializada. Puesto que el medio más objetivo de evaluar una predicción es medir el valor realmente observado de la variable de interés en el periodo pronosticado, la forma más simplista de expresar la exactitud de una predicción es mediante el denominado *error de predicción* correspondiente al período f, o diferencia entre Y_t, el valor observado por la variable considerada y F_t su valor predicho, $e_f = Y_t - F_t$.
Cuando no se considera la diferencia, sino el cociente, entre las dos magnitudes del segundo miembro de esta última expresión y se multiplica tal cociente por 100, se tiene lo que habitualmente se conoce como *porcentaje de consecución* en el mundo empresarial, ámbito en el que esta última expresión se utiliza más en relación con objetivos financieros o de ventas que como medida de la exactitud de una predicción.

Para evaluar un pronóstico, además de la expresión del error de predicción, se utilizan otros parámetros a los que se suele denominar *medidas del error de predicción.*

Cuando se desea tener en cuenta el número de períodos de datos, n, que se han utilizado en la predicción se suele utilizar el *error medio (EM)* de la predicción definido como

$$EM = \frac{1}{n}\left\{\sum_{t=1}^{n}(Y_t - F_t)\right\}$$

El inconveniente principal del parámetro *EM* radica en que su valor puede ser pequeño a pesar de existir grandes desviaciones del valor previsto respecto al realmente observado, debido a la posible compensación de las diferencias positivas entre ambos conceptos en algunos períodos con las negativas en otros dando, de esta forma, una errónea imagen global de buen ajuste del modelo a los datos reales. Para evitar esta compensación de errores de distinto signo es frecuente utilizar el parámetro *DMA (Desviación Media Absoluta) o MAD (Mean Absolute Deviation)* definido como el valor medio de los valores absolutos de tales desviaciones. A este parámetro también se le suele denominar *EAM (Error Absoluto Medió)* o *MAE (Mean Absolute Error)* y se utiliza para evaluar la bondad del ajuste.

$$EAM = DMA = \frac{1}{n}\left\{\sum_{t=1}^{n}|Y_t - F_t|\right\}$$

Otra medida muy utilizada del error de la predicción es el denominado Error *Cuadrático Medio, ECM o MSE (Mean Squared Error)* cuya expresión involucra una función cuadrática de pérdida

$$ECM = \frac{\sum_{t=1}^{n}(Y_t - F_t)^2}{n}$$

A fin de que la medida del error quede expresada en las mismas unidades de medida que la variable objeto de pronóstico, a veces se utiliza la raíz cuadrada del parámetro anterior a la que se denomina *RECM (Raiz del Error Cuadrático Medio)* que suele aparecer en la literatura especializada como *RMS o RMSE (Root Mean Squared Error)* y que se utiliza para medir la bondad del ajuste dentro de la muestra. Según lo anterior su expresión es:

$$RECM = \sqrt{ECM}$$

A excepción del *porcentaje de consecución,* las medidas de error comentadas hasta ahora presentan el inconveniente de que cualquier cambio de escala de la variable objeto de estudio afecta a la magnitud de tales medidas. Es lógico, por consiguiente, que hayan surgido medidas del error de predicción que traten de evitar tal inconveniente, relativizando las magnitudes. Entre tales medidas adimensionales figuran el *Error Relativo o Porcentaje de Error (PE, Percentage Error),* el *Porcentaje Medio de Error (MPE, Mean Percentage Error)* y posiblemente la más utilizada de todas ellas, el *EAMP (Error Absoluto Medio del Porcentaje de Error)* o, como se le suele designar en la literatura especializada, *MAPE (Mean Absolute Percentage Error)* y que se utiliza para evaluar tanto la bondad del ajuste dentro de la muestra como la exactitud de la predicción en los periodos extra muestrales. Estos tres parámetros se definen respectivamente como sigue:

$$PE_t = \left(\frac{Y_t - F_t}{Y_t} \right) \times 100$$

$$MPE = \frac{1}{n} \sum_{t=1}^{n} PE_t$$

$$MAPE = \frac{1}{n} \sum_{t=1}^{n} \left| PE_t \right|$$

Con excepción del *ECM* y del *RECM,* cuyas funciones cuadráticas de pérdida aumentan (penalizan) los errores de predicción de gran magnitud, las demás medidas de exactitud de la predicción anteriores ponderan con igual peso a

tocios los errores. El *U* o *Coeficiente de Desigualdad de Theil* -una de las medidas de error de la predicción más utilizadas en los estudios econométricos- representa, en opinión de Makridakis, S., Wheelright, S.C. y Hyndman, R.J.(1998), un compromiso entre las medidas absolutas y relativas del error de predicción.

$$U = \sqrt{\frac{\sum_{i=1}^{NF}(x_i - f_i)^2}{\sum_{i=2}^{NF}(x_i - x_{i-1})^2}}$$

El *U de Theil* es adimensional, contiene una función cuadrática de pérdida a fin de penalizar los grandes errores de predicción, indicando la mayor proximidad a cero una mejor capacidad predictiva del modelo. El cero representa el ajuste perfecto y el valor uno indica.

CONTENIDO DEL CD E INDICE DE ANEXOS

El Cd adjunto al presente trabajo, incluye la documentación, manuales, instaladores, fuentes de la aplicación desarrollada y los resultados obtenidos. Además los anexos que complementan el trabajo realizado.

- DATOS

- Atributos Base

  Esta carpeta contiene los datos necesarios para cargar a la base de datos correspondientes a los atributos base necesarios para la aplicación, puesto que estos datos corresponden a la base del conocimiento del algoritmo See5.

- Históricos

  Esta carpeta contiene los datos necesarios para cargar a la base de datos correspondientes a los datos históricos registrados por el CENACE necesarios para graficar la curva de la demanda eléctrica y además para poder comparar luego con los resultados obtenidos con el modelo de árboles de decisión.

- PrediccionArima

  Esta carpeta contiene los datos necesarios para cargar a la base de datos correspondientes a los resultados registrados por el CENACE correspondientes al modelo ARIMA.

- DOCUMENTOS

ANEXOS

Este directorio contiene los respectivos anexos complementarios del presente trabajo.

MANUALES

Este directorio contiene el manual de instalación y el manual de usuario de la aplicación desarrollada.

- INSTALADORES

Aplicación_Prediccion

Este directorio contiene el respectivo instalador de la aplicación desarrollada.

BDD

Este directorio contiene el instalador para la creación de la base de datos con los datos necesarios para que funcione la aplicación.

See5

Este directorio contiene el instalador del programa See5 de rulequest .

- RESULTADOS

Contiene los archivos de los resultados obtenidos cuando se emplea el modelo de árboles de decisión, para diferentes intervalos.

- SCRIPTS

Base de Datos

Este directorio contiene el archivo en el cual se crea la base de datos necesaria para el funcionamiento de la aplicación.

Código Fuente

Este directorio contiene todo el código fuente de la aplicación desarrollada.

Indice de Archivos en Anexos