Universidad de Cienfuegos "Carlos Rafael Rodríguez"

Facultad de Ingeniería

Carrera de Ingeniería Informática

Trabajo Diploma para optar por el Título de

Ingeniería Informática

¨Modelo para el pronóstico del consumo eléctrico en el Hotel Jagua¨

Autora:

Nimali Shanika Liyanage.

Tutores:

Lic. Maia Viera Cañive.

MsC. Boris Vega Lara.

Cienfuegos, Cuba

Curso 2010 – 2011

Declaración de autoría

Yo, Pethigala Liyanage Nimali Shanika Liyanage, declaro que soy el única autora de este trabajo y autorizo al Departamento de Informática de la Facultad de Ingeniería en la Universidad de Cienfuegos “Carlos Rafael Rodríguez”, para que hagan el uso que estimen pertinente con el trabajo de diploma.

Para que así conste firmo (firmamos) la presente a los 16 días del mes de 02 del 2012.

_____

Autora: Pethigala Liyanage Nimali Shanika Liyanage.

_____

Tutora: Lic. Maia Viera Cañive.

PENSAMIENTO.

“El espíritu le da significado a tu vida, y la posibilidad de su más grande desarrollo. Pero la vida es esencia para el espíritu, ya que su verdad no es nada si no puede vivir.”

Carl Gustav Jung

(26 de julio de 1875 - 6 de junio de 1961)

Médico psiquiatra, Psicólogo y Ensayista suizo, Fundador de la escuela de psicología analítica.

AGRADECIMIENTO.

- Para la realización de este trabajo he contado con el apoyo de muchas personas, a todos ellos quiero gratificar; pero muy en especial agradezco:
- A mi papa por saber sobrepasar todos los momentos difíciles que te ha puesto la vida, siempre has sabido salir adelante apoyándome en todos mis caprichos, gracias por tanto amor y dedicación, lo que soy te lo debo a tí.
- A mi hermano por cuidar de mí y estar siempre a mi lado, sé que puedo contar contigo para lo que necesite y quiero que sepas que yo siempre estaré contigo y que te quiero mucho.
- A mi abuela, por haberme brindado el amor y la comprensión que he necesitado durante todo este tiempo.
- A Daniel por cuidar de mi; gracias por hacerme sentir querida y por cinco años soportarme.
- A Lachy por haber podido contar contigo siempre.
- A las muchachitas del cuarto: querida Beija, Phuong.
- A Somphan, José, Hemantha, Komuni, Amilza, Samantha, Asitha, Jorge Luis(Kiko) por apoyarme cuando necesitaba ayuda.
- A mi tutora Maia, por darme tanto apoyo desde que comencé esta investigación.
- A profesor Boris, profe Ciro por darme tanto a poyo en todo momento.
- A todos los profesores.
- A todos mis amigos y amigas.
- A mi lulu.
- A todos muchas gracias!!!

DEDICATORIA

*A **mi padre** por ser esa persona maravillosa*

que confió en mí,

me señaló el camino y

me dio las fuerzas para recorrerlo,

estando siempre presente a pesar de la distancia.

Resumen.

El objetivo de la investigación es correr los modelos de regresión *Multilayer Perceptron* y Regresión Lineal, para encontrar un modelo que produzca un error menor del 5 %. Se corrieron las técnicas sobre un conjunto de datos formado por valores climáticos, de la comercialización del hotel y del departamento energético. Se aplicaron sobre los datos, técnicas para la sustitución de valores perdidos y técnicas para la selección y extracción de rasgos. Se obtuvieron varios sub-conjuntos con diferentes características, sobre los que se corrieron los modelos antes mencionados. Los resultados obtenidos no fueron los esperados, el error promedio versaba alrededor del 70%. Se puede concluir que a partir de los datos recopilados no se puede alcanzar un buen resultado, recomendando en trabajos posteriores la obtención de mejores datos. Otra conclusión extraída fue que los datos climatológicos son una fuerza importante en el consumo energético del hotel, sin embargo no son una influencia suficiente para explicar todo el consumo.

Summary.

The objective of the investigation is to run the regression models, Multilayer Perceptron and Lineal Regression, to find a model that produces an error smaller than 5%. The group of data is formed by climatic values, the values of commercialization of the hotel and the values of energy department. Furthermore, applied techniques on the data for the substitution of lost values and for the selection and extraction of features. Several sub-groups were obtained with different characteristics, on those ran the models before mentioned. The obtained results were not the prospective ones, the error average turned around 70%. Therefore, starting from the gathered data you cannot reach a good result, recommending in later works the obtaining of better data. Another extracted conclusion was that the climatological data, although they are an important force in the energy consumption of the hotel, it is not an enough influence to explain the whole consumption.

INDICE DE CONTENIDO.

INDICE DE TABLAS.

INDICE DE FIGURAS.

Introducción.

La energía es uno de los factores principales para la actividad humana, el progreso social y el desarrollo económico. Los diferentes períodos de la evolución han estado caracterizados por el dominio del hombre sobre la energía. La energética actual aún depende de la utilización de los combustibles fósiles (carbón, petróleo y gas natural); estos son agotables, contaminantes y están concentrados en pocas regiones del planeta.

El proceso de globalización económica exige que las empresas redefinan sus estrategias y sus procesos con la finalidad de lograr un uso eficiente de sus recursos y aumento de su productividad, de modo que puedan competir con éxito en el mercado. El incremento de los precios de la energía, así como la incertidumbre de su suministro pone al ahorro energético como una de las estrategias de desarrollo de primera mano para todas las entidades, para poder hacer frente a esta amenaza y así lograr una ventaja competitiva importante.

El sector turístico continúa siendo uno de los motores que mueven la economía cubana. Entre otros recursos, los establecimientos hoteleros utilizan una notable cantidad de energía para suministrar los servicios y el confort que ofrecen a sus clientes. Es por ello que el ahorro de energía se convierte en uno de los compromisos que debe asumir el sector hotelero, donde existe todavía un gran potencial para el ahorro energético.

Los hoteles se caracterizan por desarrollar un uso intensivo de la energía en su actividad diaria. En el negocio hotelero la energía constituye una importante partida de costes. Experiencias internacionales demuestran que una instalación hotelera que funciona eficientemente, desde el punto de vista energético, debe consumir entre 5 y 7% de sus ingresos para cubrir los gastos energéticos, indicador que varía en función del tipo de hotel y la categoría que ellos posean, así como del tipo de servicio que se ha prestar [1]. En Cuba, en las cadenas CUBANACAN, Gran Caribe, Isla Azul y Horizontes, este indicador oscila entre 8 y 16% y puede llegar hasta 20% en hoteles que tienen una infraestructura muy atrasada y bajos niveles de comercialización.

Según los estudios realizados, se relacionan por orden de importancia la siguiente estructura de costos de energéticos: Electricidad (65- 75%), diesel (10-15%), Gas licuado (8- 12 %) y otros hasta un 5 % del costo total.

La energía es el aparato cuyos costos crecen más rápidamente y uno de los pocos costos que pueden ser realmente controlados por expertos en el uso de la energía.

La minería de datos se puede definir como un proceso analítico diseñado para explorar grandes cantidades de datos con el objetivo de detectar patrones de comportamiento consistentes o relaciones entre las diferentes variables para aplicarlos a nuevos conjuntos de datos.

Por este contexto, en el análisis de datos en el proceso del desarrollo de un modelo matemático que pueda predecir el consumo eléctrico del Hotel Jagua, se utilizaron las técnicas de minería de datos de inteligencia artificial.

Como se puede apreciar en la definición anterior, la esencia del problema consiste en "escarbar" en la información almacenada para "descubrir" los elementos de utilidad.[2]

El proceso de minería de datos consta de tres etapas fundamentales:

Exploración de los datos.

Definición de patrones o construcción de modelos.

Validación y verificación.

La minería de datos aún se basa en los principios tradicionales de Análisis exploratorio de Datos, aunque llega más lejos, sobre todo con la incorporación de elementos de inteligencia artificial

Problema de investigación

Ausencia de un modelo matemático para la predicción del consumo eléctrico en el Hotel Jagua, en función de las variables medidas actualmente en el hotel y que su error no exceda el 5%.

Objeto de estudio

El proceso comercial y de control energético del Hotel Jagua.

Campo de acción.

Las técnicas de minería de datos de inteligencia artificial en el pronóstico de consumo eléctrico del Hotel Jagua.

Idea a defender.

Existe un modelo matemático en función de datos comerciales y/o datos climatológicos para el Hotel Jagua, que supera la precisión del 95%, para la predicción del consumo eléctrico mensual en el hotel.

Objetivo General

Obtener modelo matemático capaz de pronosticar el consumo eléctrico del Hotel Jagua, utilizando las técnicas de redes neuronales y regresión lineal, sobre un conjunto de datos formado por datos climatológicos y datos comerciales y energéticos del hotel que supere el 95% de confiabilidad.

Objetivos Específicos

1. Comprender el negocio y los datos.
2. Preparar los datos.
3. Crear un modelo.
4. Evaluar los resultados.

Tareas de Investigación

1.1 Entrevistas y encuentros con los trabajadores de los Departamentos de Comercial y Energético del Hotel Jagua.

1.2 Recolectar todos los datos que puedan aportar estos departamentos y recolectar los datos climatológicos.

1.3 Revisión de los trabajos realizados en este tema por el CEEMA.

1.4 Elaboración del objetivo del negocio a alcanzar con la minería.

1.5 Describir los datos recolectados.

1.6 Explorar y verificar la calidad de los datos.

2.1 Selección de datos.

2.2 Limpiar los datos.

2.3 Construcción de datos.

2.4 Integrar datos.

2.5 Formatear los datos en formato ARFF.

3.1 Documentar la técnica de modelación.

3.2 Construir y correr el modelo.

4.1 Evaluar los resultados.

Distribución del documento

Capítulo I: Descripción de los métodos y herramientas usadas para la solución del problema. Comprensión del negocio.

Capítulo II: Descripción de la aplicación y resultados de las tareas de investigación.

Capítulo III: Análisis de los resultados y propuestas de este trabajo.

CAPITULO 1

Introducción.

En el presente capítulo se describen varias técnicas de selección y de extracción de atributos como: C*fsSubsetEval* y *PrincipalComponents*. Y los métodos de regresión como red neuronal MultilayerPerceptron y *LinearRegression*. El software utilizado para la selección y la extracción de atributos y para la modelación fue WEKA. Además se presentan antecedentes de esta investigación. Luego, una explicación de la situación problemática y la exploración de datos obtenidos del Hotel Jagua.

1. 1 Métodos utilizados:

1.1.1 *Correlation Feature Selection Subset Eval (CfsSubsetEval)*

Este método se utiliza para la selección de atributos. *CfsSubsetEval* es un algoritmo de filtro que ordena los subconjuntos de atributos de acuerdo a una función de evaluación heurística basada en la correlación. El objetivo es encontrar un subconjunto de rasgos con una baja relación entre ellos y una alta relación de cada uno con la clase. Los atributos irrelevantes deben ser ignorados por que ellos tienen baja correlación con la clase. Los atributos redundantes deben ser eliminados por que tienen alta correlación con los atributos restantes. La función de evaluación de los sub conjuntos de atributos de *CfsSubsetEval* es:

$$M_S = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \qquad (1.1)$$

Donde *Ms* es el mérito heurístico de un sub conjunto de atributos *S* que contiene k atributos, $\overline{r_{cf}}$ es la media de las medidas de relación entre cada atributo $(f \in S)$,

y $\overline{r_{ff}}$ es la media de las medidas de relación entre cada uno de los atributos del subconjunto. El numerador de la ecuación se puede pensar como una indicación de cuán predecible de la clase es un conjunto de atributos; y el denominador cuánta redundancia hay entre los atributos. [3]

La búsqueda heurística diseñada puede ser una simple búsqueda ávida (*Greedy*) o el primero mejor (*Best First*). En la búsqueda ávida, el criterio de parada hacia delante es hasta que deje de aumentar la evaluación y hacia atrás mientras que la evaluación no se degrade. En el caso de la búsqueda el primero el mejor, se detiene cuando después de cinco exploraciones no se ha logrado mejora en la evaluación.

La dirección de búsqueda del algoritmo puede ser hacia delante (*forward*), en la cual el algoritmo comienza sin ningún rasgo y los va adicionando uno a uno. La búsqueda hacia atrás (*backward*) comienza por todo el conjunto de atributos y los va eliminando uno a uno. La dirección de búsqueda puede ser bidireccional. En este caso el algoritmo puede empezar en cualquier punto y buscar en las dos direcciones. En caso de no usar una búsqueda heurística se puede utilizar un método *ranker*. El ordena los atributos por sus evaluaciones individuales, debiendo utilizar conjuntamente un evaluador de atributos como *ReliefF, Entropy, GainRatio*, etc.

CfsSubsetEval puede trabajar con atributos nominales y numéricos. Los atributos numéricos los discretiza utilizando el método de Fayyad e Iraní [4].

1.1.2 Análisis de Componentes Principales (ACP)

Se utiliza en extracción de rasgos. En estadística, el Análisis de Componentes Principales (ACP) es una técnica utilizada para reducir la dimensionalidad de un conjunto de datos. Fue inventado por Karl Pearson en 1901[5]. Intuitivamente la técnica sirve para hallar las causas de la variabilidad de un conjunto de datos y ordenarlos por importancia. El ACP busca la proyección según la cual los datos queden mejor representados en términos de mínimos cuadrados. El ACP se emplea

sobre todo en análisis exploratorio de datos y para construir modelos predictivos. El ACP contempla el cálculo de la descomposición en auto valores[1] de la matriz de covarianza, normalmente tras centrar los datos en la media de cada atributo.

El ACP construye una transformación lineal que escoge un nuevo sistema de coordenadas para el conjunto original de datos en el cual la varianza de mayor tamaño del conjunto de datos es capturada en el primer eje (llamado el Primer Componente Principal), la segunda varianza más grande es el segundo eje, y así sucesivamente. Para construir esta transformación lineal debe construirse primero la matriz de covarianza o matriz de coeficientes de correlación. Debido a la simetría de esta matriz existe una base completa de vectores propios[2] de la misma. La transformación que lleva de las antiguas coordenadas a las coordenadas de la nueva base es precisamente la transformación lineal necesaria para reducir la dimensionalidad de datos. Además las coordenadas en la nueva base dan la composición en factores subyacentes de los datos iniciales.

Una de las ventajas del ACP para reducir la dimensionalidad de un grupo de datos, es que retiene aquellas características del conjunto de datos que contribuyen más a su varianza, manteniendo un orden de bajo nivel de los componentes principales e ignorando los de alto nivel. El objetivo es que esos componentes de bajo orden a veces contienen el aspecto "más importante" de esa información. Esta técnica permite descubrir relaciones lineales entre los atributos.

El ACP tiene la desventaja de que transforma completamente el espacio del dominio de atributos, haciendo de difícil comprensión para el ingeniero de datos en los resultados.

[1] El **autovalo**r o **valor propio** de un vector propio es el factor de escala por el que ha sido multiplicado.

[2] Los **vectores propios**, **autovectores** o **eigenvectores** de un operador lineal son los vectores no nulos que, cuando son transformados por el operador, dan lugar a un múltiplo escalar de sí mismos, con lo que no cambian su dirección.

1.1.3 *MultilayerPerceptron* (MLP)

El *MultilayerPerceptron* es una red neuronal artificial (RNA) formada por múltiples capas. Esto le permite resolver problemas que no son linealmente separables (también se pueden resolver problemas que son linealmente separables). El perceptrón multicapa puede ser totalmente o localmente conectado. En el primer caso cada salida de una neurona de la capa "i" es entrada de todas las neuronas de la capa "i+1", mientras que en el segundo cada neurona de la capa "i" es entrada de una serie de neuronas (región) de la capa "i+1".

Las capas pueden clasificarse en tres tipos:

Capa de entrada: Compuesta por aquellas neuronas que introducen los patrones de entrada en la red. En estas neuronas no se produce procesamiento.

Capas ocultas: Formadas por aquellas neuronas cuyas entradas proceden de capas anteriores y cuyas salidas pasan a neuronas de capas posteriores.

Capa de salida: Constituida por neuronas cuyos valores de salida se corresponden con las salidas de toda la red.

Una buena opción para entrenar los MLP es utilizar el algoritmo de propagación de los errores hacia atrás (*Backpropagation*). Las funciones de transferencia de las neuronas han de ser derivables. Este pertenece a la categoría de supervisado, pues requiere conocer las salidas correctas para cada ejemplo de entrada. La red *Multi Layer Perceptron* tiene algunas limitaciones como es; si la red se entrena mal o de manera insuficiente, las salidas pueden ser imprecisas. Además la existencia de mínimos locales en la función de error dificulta considerablemente el entrenamiento, pues una vez alcanzado un mínimo el entrenamiento se detiene aunque no se haya alcanzado la tasa de convergencia fijada.

Aplicando *red de retropropagación*[3] a numerosos problemas, se comprobó experimentalmente que éste era capaz de abordar problemas de clasificación de gran envergadura, de una manera eficaz y relativamente simple.[6]

1.1.4 Regresión Lineal.

Cuando la salida o la clase es numérica, y todos los atributos son numéricos, la regresión lineal es una técnica natural para considerar. Es un método principal en estadística. La idea es expresar la clase como una combinación lineal de los atributos. Por supuesto, los modelos lineales sufren por la desventaja de la linealidad. Si los datos muestran una dependencia no lineal, una recta sería un modelo extremadamente ineficiente.

1.2 Herramientas utilizadas:

1.2.1 *Waikato Environment for Knowledge Analysis* (WEKA)

El WEKA es una herramienta de s*oftware* libre. Ella como el código muy utilizada por los investigadores del mundo para hacer estudios de minería de datos.

WEKA fue desarrollado en la Universidad de Waikato (Nueva Zelanda) en 1997. Esta escrito con el lenguaje Java. Contiene una interfaz gráfica de usuario (GUI) para interactuar con los archivos de datos y producir resultados visuales. El WEKA también se puede utilizar dentro de otras aplicaciones como cualquier biblioteca. WEKA es una colección de algoritmos de aprendizaje que se utiliza en la rama de minería de datos. Contiene herramientas de pre procesamiento de datos, clasificación, regresión, *clustering*, reglas de asociación y visualización. Los algoritmos de WEKA pueden ser aplicados directamente a un conjunto de datos o llamados desde programas que usan códigos Java.

Esquemas de aprendizaje y herramientas incluidas en WEKA[7].

Classifiers (cubre la clasificación y regresión supervisada)

[3] En muchas ocasiones al conjunto arquitectura de MLP más aprendizaje (propagación de los errores hacia atrás) se le denomina *red de retropropagación.*

Clusterers (aprendizaje no supervisado)

Associations

Attribute Selection (se encuentran los métodos para la selección y extracción de rasgos)

Preprocessing Filters (procesamiento de datos supervisados y no supervisados)

Para aplicar las técnicas de minería de datos a los conjuntos de datos utilizados en este trabajo se utilizó el *software* WEKA versión 3-7-4.

1.2.2 MATLAB

MATLAB (abreviatura de *MATrix LABoratory*, "laboratorio de matrices") es un *software* matemático que ofrece un entorno de desarrollo integrado con un lenguaje de programación propio. Está disponible para las plataformas Unix, Windows y Apple Mac OS X.

Se usa en el campo de las matemáticas, la computación, el desarrollo de algoritmos, adquisición de datos, modelación y simulación, análisis de datos de exploración y visualización, gráficas científicas y desarrollo de aplicaciones entre otros.

Una de las características más especiales de MATLAB son los *toolboxes*. Estos contienen las herramientas que facilitan la resolución de problemas complejos. En este trabajo se hizo uso del *toolbox* de redes neuronales.

1.3 Comprensión del negocio.

Es de interés del Centro de Estudios de Energía y Medio Ambiente (CEEMA) de la Universidad de Cienfuegos, desde hace varios años, hacer una caracterización energética de los hoteles cienfuegueros y desarrollar herramientas que permitan un mejor control de los consumos energéticos en estos inmuebles. En particular este trabajo toma como objeto de estudio el Hotel Jagua, para explorar su comportamiento energético a partir de datos históricos.

La estructura administrativa del hotel está conformada por departamentos, estando involucrados en esta investigación el Departamento de Energía y el Departamento de Reserva. El Departamento de Energía controla los consumos energéticos del hotel y se encarga de proponer las medidas de ahorro. El Departamento de Reserva maneja la información referente a los huéspedes.

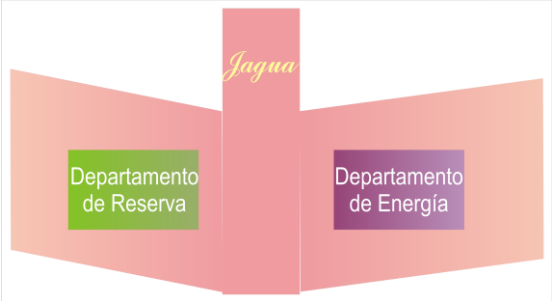


Figura 1.1: Departamentos administrativos del Hotel Jagua involucrados en la investigación.

El Departamento de Energía realiza lecturas del metro contador en tres ocasiones al día, en horario de la madrugada, en el día y en el horario pico[4]. Para el mes el hotel contrata una cuota de consumo, establecida de ante mano, en un plan anual para cada uno de los meses del año. La tarifa vigente para los hoteles contempla que si el inmueble supera el consumo contratado es penalizado.

La experiencia de los operadores y trabajadores del Departamento de Energía les permite valorar sin acabar el mes como van según el plan establecido y prever posibles incumplimientos. El análisis previsor de los energéticos se basa en el análisis de los números registrados por el contador, la época del año y los acumulados históricos para la fecha, no tienen acceso a otros datos del hotel como número de huéspedes previstos para ese mes o datos reales de cómo se comportarán las temperaturas, pero saben por los años de trabajo, que en los meses mas fríos los consumos suelen ser inferiores a la media y que en verano a

[4] El horario de la madrugada comprende las horas de 21:00 a 6:00, el horario del día 6:00 a 17:00 y el resto de las horas es el horario pico.

pesar de ser temporada baja los consumos se disparan, debido a las altas temperaturas y el sistema de climatización.

Sería superficial dejar el proceso actual de control energético del hotel sólo en la experiencia. Existen indicadores internacionales, para valorar aspectos específicos de los hoteles, uno de ellos es el indicador energético de Habitaciones – Días - Ocupadas (I-HDO)[5]; el control y evaluación de las medidas tomadas por el Departamento de Energía se realiza a través de este estándar. Este indicador, según estudios realizados por el propio CEEMA en el Hotel Jagua [8],[9],[10], tiene una pérdida considerable en su estimación y valoración del consumo energético.

Las sucesivas investigaciones han conducido a que se desee por el centro de estudios revisar un grupo más amplio de factores hoteleros y climatológicos que pudieran estar influyendo en el consumo energético del hotel y brindar un indicador más confiable, basado en otros factores, que se ajuste más a los consumos reales.

Con la obtención de un indicador más fiable y ajustado a las características del hotel y la obtención de una herramienta que permita su cálculo fácilmente se desea poder mejorar los mecanismos de predicción de los consumos energéticos del hotel.

1.3.1 Antecedentes

Varios estudios energéticos se han realizado en el hotel, dirigidos la mayoría de ellos por el CEEMA. En el estudio realizado por Isdel Geroy Borlado[10] se define la variable Habitaciones - Días - Ocupadas - Equivalente (HDOeq) que describe mejor la variabilidad del consumo eléctrico, para el indicador mundial; es una combinación de factores de temperatura, carga de las habitaciones y de otros servicios. La carga de las habitaciones logra la diferenciación entre las habitaciones (Fc) y el factor de servicio adiciona consumos no considerados en las habitaciones ocupadas (Fcs), como puede ser el restaurante o la discoteca, afectado por la ocupación en horas de las áreas de servicio (SO).

$$HDOeq = Ft *(HDO * Fc + SO *Fcs) \quad (1.2)$$

[5] El indicador I-HDO = kWh/HDO, donde kWh son los kilo Watts por hora y HDO las habitaciones ocupadas en el día.

Escobar[8] asevera que la mayor influencia en el consumo eléctrico del hotel es la energía dedicada a la climatización y propone dos ecuaciones lineales para pronosticar los consumos y evaluar la eficiencia de la instalación, en las dos estaciones de tiempo principales del país, invierno y verano.

$$E = 26;61 * HDOeq + 69967(Invierno) \quad R2=0;81$$

$$E = 26;86 * HDOeq + 125492(Verano) \quad R2=0;97 \qquad (1.3)$$

En este estudio las ecuaciones fueron obtenidas en un período comprendido entre enero de 2003 y mayo del 2004, lo que nos deja prácticamente una sola repetición de datos de cada estación climatológica. Esta pobreza en los datos, nos sugiere no confiar ciegamente en los valores de correlación obtenidos, hasta contrastar las ecuaciones con un número mayor de datos.

Geroy[10] realiza una caracterización energética del hotel basada en un estudio de los datos recogidos entre los años 2007 y 2009 concluye que el 16% del total de los gatos del hotel son por concepto de portadores energéticos y de ellos la electricidad asume el peso principal aportando más de la mitad de los gastos representados en el 16%.

Para tratar de mejorar o controlar este proceso se trabajó, en varios sentidos obteniéndose los resultados que se relacionan a continuación:

1. Se concluye que el indicador HDO, no es un indicador confiable, ni preciso para calificar el consumo energético de los hoteles. Las conclusiones fueron obtenidas mediante correlaciones lineales entre las variables “Habitaciones - Días – Ocupadas” y “kWh”. Para tratar de encontrar las variables que más influían en el consumo eléctrico se ejecutaron otros análisis estadísticos, el mejor modelo matemático resultante describe el 73% de la variabilidad del gravamen energético; la ecuación es lineal, en función de las variables “Horas- Grado” y “Turistas Opcionales”.

2. Creación de un índice de consumo teórico el cual permite valorar el comportamiento del consumo energético.

3. Definición de una metodología de trabajo para hacer un uso eficiente de los nuevos avances.

La tesis resume también un grupo de deficiencias y recomendaciones, siendo éstas:

1. No se realizó un estudio exhaustivo de todos los factores que pueden estar incluyendo en el consumo energético, siendo éste todavía un campo abierto y no concluido.

2. La metodología propuesta es muy dependiente de la sistematicidad de la recolección de los datos y de instrumentos precisos para su medida.

3. El análisis de los datos puedes ser bastante complejo y requiere por parte del operador conocimientos de Excel y de varios diagramas, pudiendo pasar por alto algún desajuste importante en los datos o bien no poder percatarse de ligeros cambios, que pudieran ser de interés. Se sugiere la creación de un sistema más fácil de manejar.

1.4 Compresión de los datos

Los datos de esta investigación son de disímiles fuentes y formatos. El Departamento de Energía proporcionó medidas muy precisas tomadas diariamente por un período de varios años. Los datos brindados por el Departamento de Reserva son resúmenes mensuales obtenidos por el sistema informático de comercialización. Los datos del clima se extrajeron de la página web http://espanol.wunderground.com, no es oficial, pero los datos que brinda son muy fiables; se realizó una comparación con datos proporcionados por el Centro Meteorológico de Cienfuegos y las diferencias no fueron significativas.

1.5 Colección de los datos iniciales

Todas las colecciones de datos estaban en formato de Excel, esta misma herramienta se utilizó para el engorroso proceso de integración. Las colecciones finales quedaron divididas en dos archivos Excel, DS-diario.xls y DSmensual.xls, la primera con todos los datos medidos diariamente y la segunda con todos los resúmenes mensuales.

1.6 Descripción de los datos

Como se ha mencionado con anterioridad la fuente de los datos fue diversa. Es posible agrupar el origen de los datos en tres grupos fundamentales, los datos energéticos, los datos de la gestión comercial y los datos del tiempo climatológico.

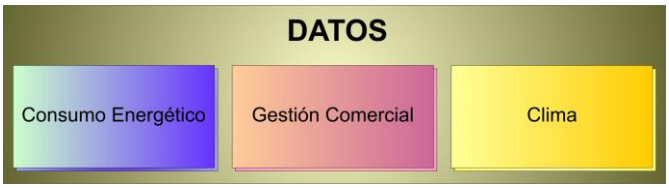


Figura 1.2: Agrupación según el origen de los datos.

Los datos energéticos fueron provistos por el Departamento de Energía. Estos datos se caracterizan por ser muy fieles y granulados, miden los consumos energéticos, pérdidas y demandas. Son el producto de las lecturas de los operarios del metro contador en tres momentos del día establecidos por la tarifa de la empresa eléctrica, estos son los horarios picos, de madrugada y día.

Los datos comerciales se extrajeron de resúmenes mensuales aportados por el sistema de gestión informático del Departamento de Reserva. Los parámetros coleccionados describen los turistas hospedados en el hotel y su relación con las habitaciones.

Los datos climatológicos son medidas diarias tomadas en un amplio período de tiempo, proceden del sitio en Internet mencionado con anterioridad.

1.6.1Tablas de datos.

Los datos se agrupan en dos archivos Excel, DS-diario.xls contiene una tabla con todos los datos medidos diariamente. Las variables medidas son meteorológicas, energéticas y la variable de ocupación habitacional. Los registros de esta tabla pertenecen al período de 1 de enero de 2007 al 31 de diciembre de 2009. Los datos cuentan con un total de 1096 registros y 28 factores.

DS-mensual.xls contiene las medidas mensuales que son las medidas de los parámetros de los turistas y las variables energéticas. La tabla recoge las variables de fecha.

La Tabla mide un período desde enero del 2004 a diciembre del 2009 y tiene 72 registros y 16 factores.

1.6.2 Descripción de las variables del estudio.

En las tablas que siguen se describen las variables[6] agrupadas por la fuente. La primera columna hace una breve descripción (Descripción), la segunda es la unidad de medida (UM), la tercera una abreviatura (Abrev.) y la cuarta el tipo de variable (Tipo).

Las variables energéticas, son las variables que se desean pronosticar. Para ellas se quiere encontrar un modelo de predicción y un indicador.

[6] En este documento se emplean como términos homólogos variable y factor.

Descripción	UM	Abrev.	Tipo
Consumo en la madrugada	kW	CM	Continuo
Consumo en el horario del día	kW	CD	Continuo
Consumo en el horario pico	kW	CP	Continuo
Pérdidas en transferencia	kW	PT	Continuo
Demanda máxima de consume	kW	DC	Continuo
Consumo total (CM+CD+CP)	kW	CT	Continuo

Tabla 1.1: Variables energéticas.

El Hotel Jagua tiene 149 habitaciones, cuando las variables que miden las habitaciones se evalúan mensualmente, se asume que se tienen 149 habitaciones por la cantidad de días del mes, por ejemplo, las plazas disponibles en el mes son la suma de la cantidad de habitaciones disponibles cada día del mes.

Los turistas de paquete son aquellos cuya reserva fue hecha a través de un paquete turístico, y los opcionales, es cualquier otro viajero extranjero que hace una reservación en el hotel.

Descripción	Abrev.	Tipo
Plazas disponibles por día	PD	Continuo
Turistas fisicos extranjeros	TF_ E	Continuo
Turistas fisicos nacionales	TF_ N	Continuo
Turistas extranjeros por día, de paquetes	TD_ E_ P	Continuo
Turistas extranjeros por día, opcionales	TD _E_ O	Continuo
Turistas por día, nacionales	TD _N	Continuo
Habitaciones por día disponibles	HDD	Continuo
Habitaciones por día ocupadas por extranjeros de paquete	HDO_ E_ P	Continuo
Habitaciones por día ocupadas por extranjeros opcionales	HDO_ E_ O	Continuo
Habitaciones por día ocupadas por nacionales	HDO_ N	Continuo

Tabla 1.2: Variables comerciales.

Además de las variables anteriores se adicionaron las variables temporales, que sitúan la medición en una fecha determinada.

Descripcion	UM	Abrev.	Tipo
Temperatura maxima	Grados celcios	TempX	Continuo
Temperatura media	Grados celcios	TempA	Continuo
Temperatura minima	Grados celcios	TempI	Continuo
Punto de rocio máximo	Grados celcios	RocX	Continuo
Punto de rocio medio	Grados celcios	RocA	Continuo
Punto de rocio mínimo	Grados celcios	RocI	Continuo
Humedad maxima	%	HumX	Continuo
Humedad media	%	HumA	Continuo
Humedad minima	%	HumI	Continuo
Presión al nivel del mar máxima	hPa	PresX	Continuo
Presión al nivel del mar media	hPa	PresA	Continuo
Presión al nivel del mar mínima	hPa	PresI	Continuo
Visibilidad maxima	km	VisiX	Continuo
Visibilidad media	km	VisiA	Continuo
Visibilidad minima	km	VisiI	Continuo
Velocidad del viento maxima	km/h	VieX	Continuo
Velocidad del viento media	km/h	VieA	Continuo

Ráfaga	km/h	Raf	Continuo
Nubosidad		Clo	Ordinal
Eventos		Ev	Discreto

Tabla 1.3: Variables del tiempo climático.

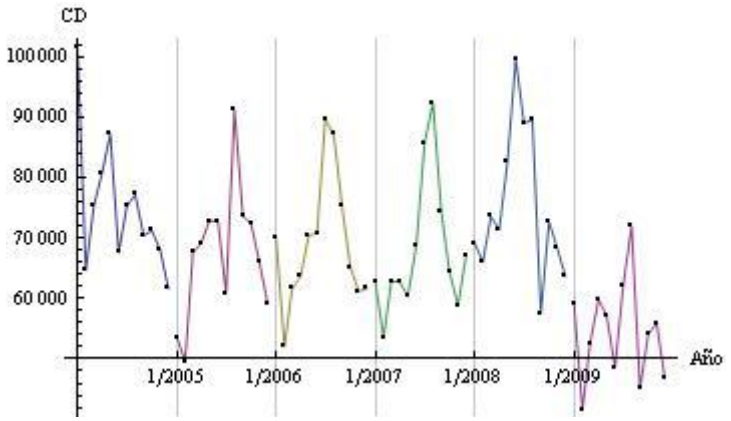
Descripción	Abrev.	Tipo
Año	Año	Continuo
Mes	Mes	Continuo

Tabla 1.4: Variables de tiempo.

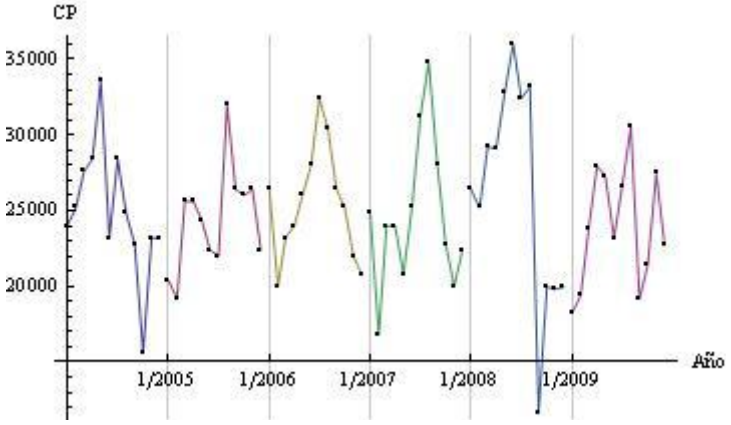
1.7 Exploración de los datos.

Una primera exploración se enfocó en las variables objetivos. Se orientó el trabajo alrededor de las preguntas: ¿Cómo es el comportamiento de las variables energéticas? ¿Entre ellas existen diferencias significativas?. ¿Existe algún comportamiento periódico?

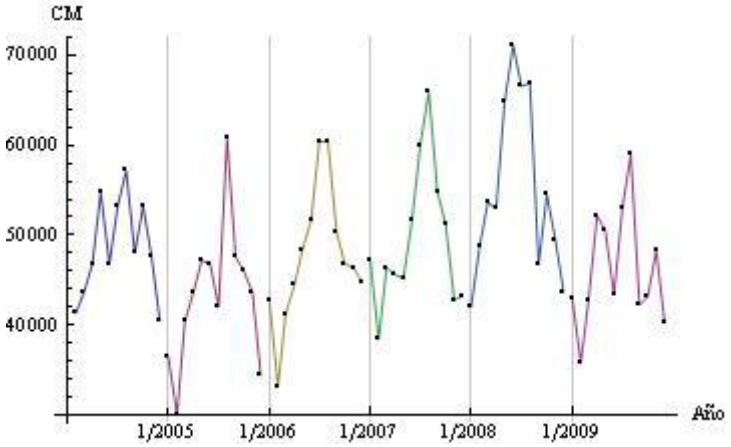
Las variables energéticas tienen un comportamiento bastante periódico en períodos de un año. La variabilidad de las variables es muy similar para cada año. Se concluye una similitud en proporciones muy parecidas entre todas las variables objetivos. En la figura 1.3 se muestran las gráficas de las mediciones mensuales de las variables energéticas.


CD
100000
90000
80000
70000
60000
1/2005
1/2006
1/2007
1/2008
1/2009
Año

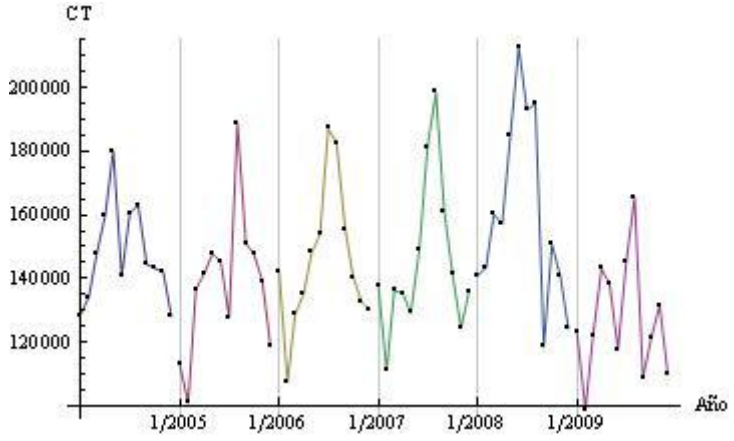
CD


CP
35000
30000
25000
20000
1/2005
1/2006
1/2007
1/2008
1/2009
Año

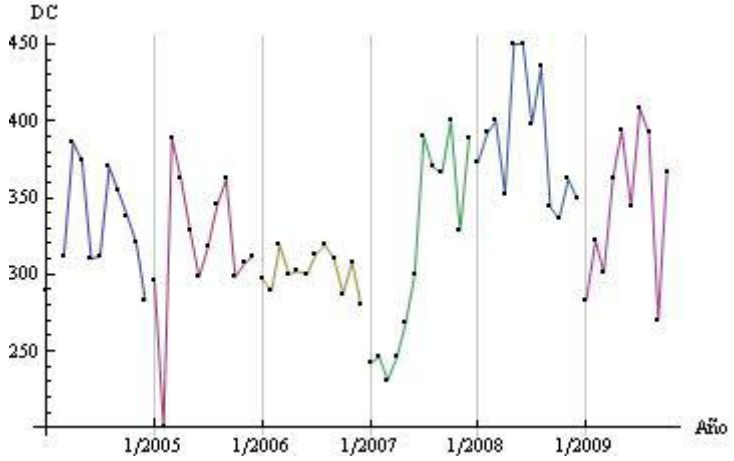
CP


CM
70000
60000
50000
40000
1/2005
1/2006
1/2007
1/2008
1/2009
Año

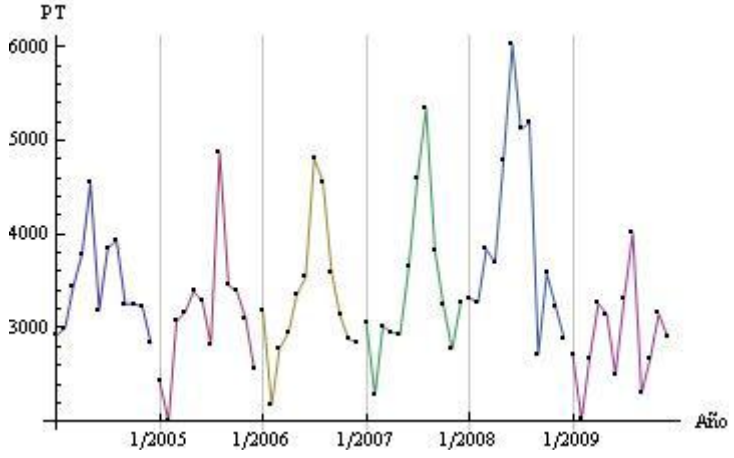
CM



CT



DC



PT

Figura 1.3: Gráficos de las medidas mensuales de las variables energéticas.

Analizando las gráficas anteriores se puede ver que las variables CM, CT y PT tiene un comportamiento periódico en el período de un año. El alcance de la tesis se estuvo sólo en el análisis de la variable consumo total como variable objetivo.

1.8 Verificación de la calidad de los datos

El completamiento de los datos es afectado por su procedencia. Algunos datos como los obtenidos del Departamento de Energía son muy completos y abarcan un período de tiempo amplio. Otros como los de la Gestión Comercial del Hotel, son de períodos cortos de tiempo. Los rangos temporales de medición diferentes, hacen que los datos en su conjunto total no sean completos en un mismo período de tiempo, de preferencia amplio. Los valores perdidos no son muy frecuentes si los datos son separados por su origen, pero si se integran buscando un solo conjunto homogéneo, en un período largo, la ausencia de mediciones en determinadas fechas, puede hacer que los valores perdidos sean un problema realmente importante.

Para hacerse una rápida idea sobre el comportamiento de los valores perdidos se modelaron algunas gráficas. La Figura 2.4 muestra las variables del conjunto de datos diarios. La variable HDO es la de peores medidas.

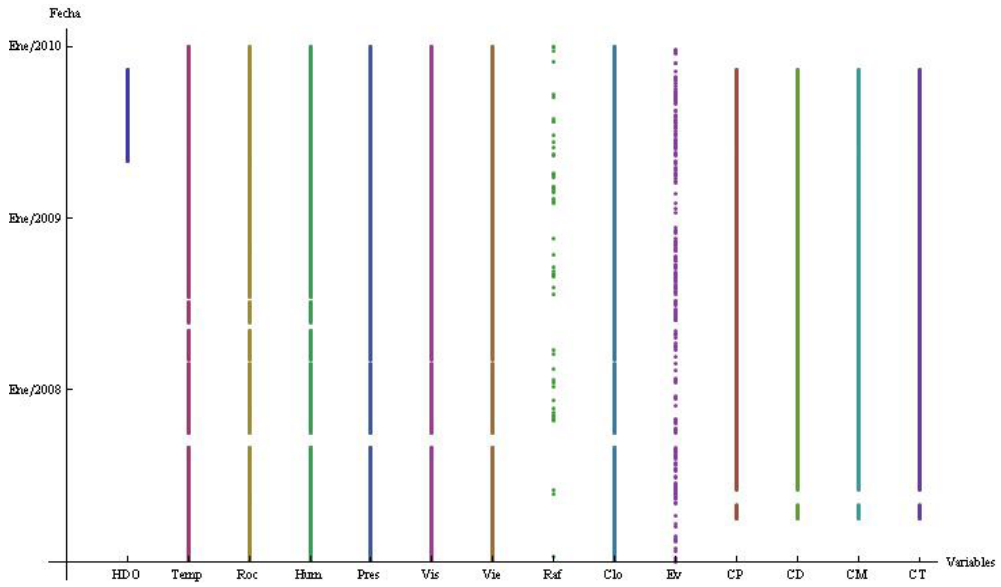


Figura 1.4: Gráfica de valores perdidos para los datos medidos diariamente.

Para el conjunto de datos medidos mensualmente, no presentan prácticamente ningún valor perdido. La figura 1.5 visualiza este comportamiento.

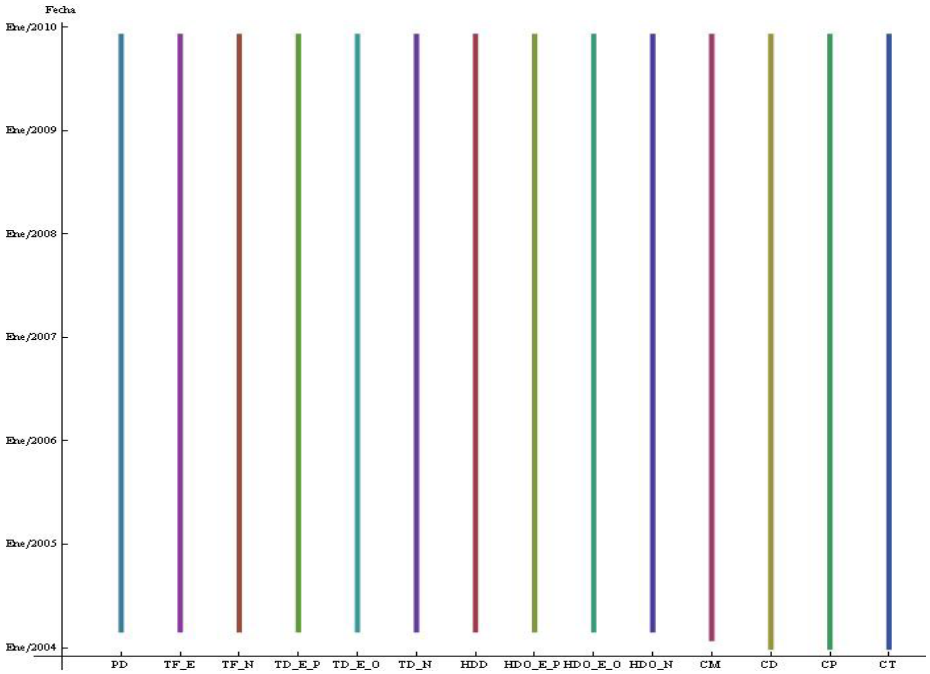


Figura 1.5: Gráfica de valores perdidos para los datos medidos mensualmente.

Conclusiones parciales del capítulo.

Se pudieron hacer descripciones mencionando sus características esenciales acerca de los métodos utilizados, *CfsSubsetEval,* ACP, MLP, Regresión Lineal.

Por el análisis del negocio se concluye que con la obtención de un indicador más fiable y ajustado a las características del hotel, se pueden mejorar los mecanismos de predicción de los consumos energéticos del hotel.

Se realizó un análisis de las variables de los conjuntos de datos obtenidos en el Hotel Jagua, luego una exploración y una verificación de la calidad de estos datos.

CAPITULO 2

Introducción:

En este capítulo se justifican las herramientas utilizadas durante el desarrollo de la investigación. Se definen los atributos seleccionados a partir del conjunto de datos inicial, usando el método de selección de atributos *CfsSubsetEval* y el de extracción *PrincipalComponents*. Además se muestran los resultados alcanzados en los entrenamientos del MLP y los resultados utilizando *LinearRegression*.

Primera iteración.

2.1. Preparación de datos.

2.1.1 Selección de datos.

Dentro de la selección de datos puede haber selección de atributos y/o selección de registros. El primer problema presentado con los datos obtenidos para este trabajo fue el período al que pertenecen. El primer grupo de datos son medidas diarias de las variables climatológicas tomadas en el período comprendido entre los años 2007 y 2009. El segundo grupo de datos son medidas mensuales de las variables comerciales y medidas diarias del consumo eléctrico del hotel entre los años 2004 y 2009.

Si los datos están completos o no, es debido a su procedencia. Algunos datos como los obtenidos del Departamento de Energía son muy completos y abarcan un período de tiempo amplio. Otros como los de la gestión comercial del Hotel, son de períodos cortos de tiempo. Los rangos temporales de mediciones diferentes, hacen que los datos en su conjunto total no sean completos en un mismo período de tiempo, de preferencia amplio. Los valores perdidos no son muy frecuentes si los datos son separados por su origen, pero si se integran buscando un conjunto homogéneo, en un período largo, la ausencia de mediciones en determinadas fechas, puede hacer que los valores perdidos sean un problema realmente importante.

Por esta razón se decidió considerar sólo el período 2007-2009, que tiene muy pocas valores perdidos.

Para aplicar los métodos de clasificación se decidió realizar una selección de atributos que permitan comparar los diferentes resultados para seleccionar, el o los mejores de ellos.

El conjunto de datos inicial usado para la selección de atributos fue CD_1.

CD_1: El conjunto de datos inicial con todos los atributos.

Mes ,Año,PD ,TF_E,TF_N,TD_E_P ,TD_E_O, TD_N , HDD,HDO_E_P,HDO_E_O ,HDO_N ,TempX ,TempA ,TempI,RocX,RocA,RocI ,HumX , HumA,HumI, PresX,PresA ,PresI ,VisiX ,VisiA ,VisiI ,VieX ,VieA ,Raf,Clo ,Lluvia,Tormenta Niebla,Lluvia-Tormenta,Niebla-Lluvia ,Nada ,CT

A partir del conjunto de datos CD_1, incluyendo sólo las variables climatológicas y la variable objetivo.CD_2.

CD_2: El conjunto de datos inicial con todos la variables climatológicas.

Mes ,Año,TempX ,TempA ,TempI,RocX,RocA,RocI ,HumX , HumA,HumI, PresX,PresA ,PresI ,VisiX ,VisiA ,VisiI ,VieX ,VieA ,Raf,Clo ,Lluvia,Tormenta Niebla,Lluvia-Tormenta,Niebla-Lluvia ,Nada ,CT

En investigaciones anteriores las variables climatológicas eran uno de los factores que no faltaba en ningún modelo predictor del consumo eléctrico. Esta razón fue la raíz de explorar estas variables individualmente en este experimento.

La selección de atributos y extracción de atributos se realizó utilizando los métodos siguientes respectivamente.

Correlation-based Feature Subset Selection (*CfsSubsetEval Eval*)

PrincipalComponents.

2.1.1.1 *CfsSubsetEval*

CfsSubsetEval utiliza un método de búsqueda. En este trabajo se seleccionó el método de búsqueda *Best First*. Este algoritmo busca en el espacio, los sub conjuntos de atributos por ascensión en colina argumentado con una facilidad backtracking. *Best First* se puede iniciar con un conjunto de atributos vacíos y buscar hacia adelante (*Froward*), o iniciar un conjunto de atributos completos y buscar hacia atrás (*Backward*), o iniciar en cualquier punto y buscar en ambas direcciones. (Considerando todas las posibilidades de las adiciones y eliminaciones de atributos en un punto dado).

El modo de selección de atributos utilizado es validación cruzada con el número de pliegues 10 y la semilla de números aleatorios 1.

La clase de los datos es numérica. La configuración de los parámetros en CfsSubsetEval se fijó como se indica a continuación:

locallyPredictive: True

missingSeparate: False

Para los significados de cada parámetro ver Anexo 1.

La configuración del método de búsqueda:

Direction:Forward

lookupCacheSize:1

searchTermination:5

startSet: [vacío]

Con la opción *direction* se especifica la dirección de la búsqueda. La opción *lookupCacheSize* ajusta el tamaño máximo de caché para los sub-conjuntos de datos evaluados. Por defecto toma el valor 1. La búsqueda termina por cumplir la máxima cantidad de *backtracking* especificada en la opción *searchTermination*. El punto inicio de la búsqueda se ajusta con la opción *startSet*. Puede ser una lista de números correspondientes a cada atributo; empieza en 1. Se pueden incluir rangos. Ejemplo: 1, 2,5-9,17. En este caso es vacío.

Corridas de *CfsSubsetEval*

El conjunto de datos CD_1 se corrió con *CfsSubsetEval* y se obtuvieron los resultados siguientes.

Evaluation mode: 10-fold cross-validation

=== Attribute selection 10 fold cross-validation seed: 1 ===

number of folds (%) attribute

1(10 %) 1 Mes

0(0 %) 2 Año

0(0 %) 3 PD

0(0 %) 4 TF_E

1(10 %) 5 TF_N

0(0 %) 6 TD_E_P

0(0 %) 7 TD_E_O

0(0 %) 8 TD_N

0(0 %) 9 HDD

0(0 %) 10 HDO_E_P

5(50 %) 11 HDO_E_O

10(100 %) 12 HDO_N

10(100 %) 13 TempX

10(100 %) 14 TempA

10(100 %) 15 TempI

10(100 %) 16 RocX

10(100 %) 17 RocA

10(100 %) 18 RocI

0(0 %) 19 HumX

6(60 %) 20 HumA

5(50 %) 21 HumI

1(10 %) 22 PresX

0(0 %) 23 PresA

0(0 %) 24 PresI

10(100 %) 25 VisiX

10(100 %) 26 VisiA

10(100 %) 27 VisiI

10(100 %) 28 VieX

10(100 %) 29 VieA

9(90 %) 30 Raf

0(0 %) 31 Clo

7(70 %) 32 Lluvia

9(90 %) 33 Tormenta

4(40 %) 34 Niebla

10(100 %) 35 Lluvia-Tormenta

0(0 %) 36 Niebla-Lluvia

9(90 %) 37 Nada

Los resultados obtenidos con el conjunto de datos CD_1 corridos con el método *CfsSubsetEval* arroja que los atributos: HDO_N, TempX, TempA, TempI, RocX, RocA, RocI, VisiX, VisiA, VisiI, VieX ,VieA y Lluvia-Tormenta.

De acuerdo con los resultados obtenidos se formaron dos conjuntos de datos. CD_3 con todos los atributos que reportó cfs como valioso y CD_4 con todos aquellos atributos que el *CfsSubsetEval* reportó más del 75% de las veces que fue corrido (10 veces).

CD_3: Todos los atributos que reportó *CfsSubsetEval* con el conjunto de datos CD_1como valioso

Mes ,TF_N ,HDO_E_O ,HDO_N ,TempX,TempA ,TempI,RocX ,RocA ,RocI,HumA ,HumI ,PresX ,VisiX ,VisiA ,VisiI ,VieX,VieA ,Raf ,Lluvia ,Tormenta ,Niebla ,Lluvia-Tormenta ,Nada , CT

CD_4: Todos aquellos atributos que el *CfsSubsetEval* reportó más del 75% de las veces que fue corrido con el conjunto de datos CD_1.

HDO_N ,TempX,TempA ,TempI ,RocX ,RocA ,RocI ,VisiX ,VisiA ,VisiI ,VieX ,VieA ,Raf ,Tormenta ,Lluvia-Tormenta ,Nada , CT

Con el conjunto de datos CD_2.

Evaluation mode: 10-fold cross-validation

=== Attribute selection 10 fold cross-validation seed: 1
===

number of folds (%)
attribute

0(0 %) 1 Mes

6(60 %) 2 Año

0(0 %) 3 TempX

0(0 %) 4 TempA

8(80 %) 5 TempI

0(0 %) 6 RocX

0(0 %) 7 RocA

0(0 %) 8 RocI

1(10 %) 9 HumX

0(0 %) 10 HumA

0(0 %) 11 HumI

0(0 %) 12 PresX

0(0 %) 13 PresA

0(0 %) 14 PresI

10(100 %) 15 VisiX

8(80 %) 16 VisiA

0(0 %) 17 VisiI

1(10 %) 18 VieX

0(0 %) 19 VieA

7(70 %) 20 Raf

0(0 %) 21 Clo

3(30 %) 22 Lluvia

10(100 %) 23 Tormenta

3(30 %) 24 Niebla

10(100 %) 25 Lluvia-Tormenta

0(0 %) 26 Niebla-Lluvia

0(0 %) 27 Nada

Con estos resultados obtenidos se creó un conjunto de datos CD_5 con todos los atributos que *CfsSubsetEval* reportó como valioso.

CD_5: Todos los atributos que reportó *CfsSubsetEval* con el conjunto de datos CD_2 como valioso

Año ,TempI ,HumX ,VisiX ,VisiA ,VieX ,Raf ,Lluvia ,Tormenta ,Niebla ,Lluvia-Tormenta, CT

2.1.1.2 *PrincipalComponents.*

.PrincipalComponents realiza un análisis de los Componentes Principales y una transformación de datos. Se usa de conjunto con la búsqueda *Ranker*. Este método logra la reducción de la dimensionalidad, escogiendo suficientes vectores de *eigen*[7]. Este proceso se realiza de acuerdo a un porcentaje de varianza de los datos originales. Por defecto es 95%. Los ruidos de los atributos pueden ser eliminados transformándolos en un espacio de componentes principales. Esto se hace eliminando algunos de los peores vectores de *eigen*. Y luego transformando los atributos a su espacio original.

¿Por qué el método *PrincipalComponents*?

- El método usado dentro de *PincipalComponents* es Análisis de Componentes Principales, siendo éste muy utilizado para muchas aplicaciones en áreas de inteligencia artificial, gráficos por computadoras, etc.
- Revela estructuras simples en los conjuntos de datos complejos.

[7] Los **vectores propios**, **autovectores** o **eigenvectores** de un operador lineal son los vectores no nulos que, cuando son transformados por el operador, dan lugar a un múltiplo escalar de sí mismos, con lo que no cambian su dirección.

- Es un método completamente no paramétrico. Cualquier conjunto de datos puede ser analizado por este método. No requiere ningún parámetro, ni importa la manera en que están organizados los datos, la respuesta del método es única e independiente del usuario.

Configuración usada:

CenterData: False

maximumAttributeNames: 5

transformBackToOriginal: False

varianceCovered: 0.95

El modo de selección de atributos es utilizar el conjunto de entrenamiento completo.

El método de *PrincipalComponents* se utiliza de conjunto con el método de búsqueda *Ranker*.

La búsqueda *Ranker* ordena los atributos de acuerdo a sus evaluaciones individuales. Se debe usar también de conjunto con un evaluador de atributos. WEKA facilita *RiliefF*, *GainRatio*, *Entropy*, etc. La búsqueda está implementada usando *ReliefF*.

Configuración de la búsqueda utilizada:

generateRanking: True

numToSelect:-1

startSet: [vacío]

threshold:-1.7976931348623157E308 (valor por defecto)

GenerateRanking es una opción constante que siempre toma el valor *True*. *Ranker* sólo es capaz de ordenar los atributos. La opción *numToSelect* posibilita especificar la cantidad de atributos a retener. El valor de este campo por defecto es -1, significa

que debe retener todos los atributos. Se tiene que usar esta opción o la opción *threshold* para reducir el conjunto de atributos.

El campo *startSet* especifica un conjunto de atributos a ignorar. Cuando la búsqueda *Ranker* ordena los atributos no evalúa los atributos que están especificados en este campo. Estos atributos especifican como una lista separada por coma. También se pueden incluir rangos. El indexado de los atributos debe empezar con 1. *Threshold* es el umbral. Este especifica cuáles atributos pueden ser eliminados. El valor por defecto no desecha ningún atributo. Debe utilizar esta opción o la opción *numToSelect* para reducir el conjunto de atributos.

Corridas de *PrincipalComponents.*

Con el conjunto de datos CD_1:

eigenvalue	proportion	cumulative	
15.30983	0.41378	0.41378	-0.249RocI-0.249RocA-0.247RocX-0.234TempI-0.225TempA...
4.44832	0.12022	0.534	0.376HDO_N+0.366TD_N+0.362TF_N-0.313HumX-0.25Año...
3.50055	0.09461	0.62861	-0.367TD_E_O-0.35PD-0.35HDD-0.342HDO_E_O-0.311VisiI...
2.28336	0.06171	0.69033	0.399Clo-0.301Año-0.292Lluvia-Tormenta+0.261Lluvia+0.243HumI...
1.79772	0.04859	0.73891	-0.446Raf+0.369VisiI-0.346VieX-0.288Niebla+0.237Nada...
1.52279	0.04116	0.78007	-0.276PresI-0.26PresA+0.26 TempX-0.255PresX-0.249HumA...
1.26767	0.03426	0.81433	-0.467Niebla-Lluvia+0.41 HDO_E_O-0.344VisiX+0.256TD_E_O…
0.99836	0.02698	0.84131	0.485Niebla-Lluvia+0.314Lluvia-Tormenta-0.312Niebla+0.309Raf
0.96834	0.02617	0.86748	0.507Niebla-0.351Lluvia-0.344HDD-0.344PD-0.27Niebla-Lluvia...
0.8039	0.02173	0.88921	-0.435Niebla+0.396HDO_E_O+0.259Lluvia-Tormenta…
0.76942	0.0208	0.91001	0.476Tormenta+0.405Clo-0.27Niebla-Lluvia+0.262Raf-0.245Año...
0.65868	0.0178	0.92781	0.468Raf+0.405Mes+0.363Nada-0.316Lluvia-0.293Clo...
0.55031	0.01487	0.94268	-0.395Mes+0.319Raf-0.313Niebla-Lluvia-0.261HumX-0.256VieX...
0.45789	0.01238	0.95506	0.487Clo-0.275TD_E_P-0.263HDO_E_P+0.261VisiI+0.257Mes...

Como se muestra en la tabla anterior el *PrincipalComponents* extrae 14 combinaciones lineales de las variables originales. Estos 14 nuevos atributos explican un 95 % de la variablilidad de todo el conjunto.

CD_6: 14 vectores lineales PC del conjunto de datos CD_1.

V1,V2,V3,…,V14, CT

Con el conjunto de datos CD_2:

eigenvalue	proportion	cumulative	
12.80026	0.474083	0.47408	-0.274RocA-0.273RocI-0.271RocX-0.257TempI-0.248TempA...
3.16106	0.11708	0.59116	0.366VisiX+0.355VisiA+0.324Mes+0.301HumX+0.266VisiI...
2.08183	0.0771	0.66826	-0.451VisiI-0.4VisiA+0.372Raf+0.257Clo+0.251Lluvia...
1.75179	0.06488	0.73315	-0.451Año-0.396Raf-0.317Niebla-Lluvia-0.307Niebla-0.283VieX...
1.56206	0.05785	0.791	-0.382PresI-0.368PresA-0.334PresX+0.31 VieX-0.296Año...
0.9579	0.03548	0.82648	0.64 Niebla-Lluvia-0.634Niebla+0.186Lluvia+0.164Raf+0.156Lluvia-Tormenta...
0.87439	0.03238	0.85886	-0.668Tormenta-0.374Clo+0.348Año+0.262Nada-0.23VisiX...
0.71869	0.02662	0.88548	-0.646Niebla-Lluvia-0.438Niebla+0.265Año+0.255VieX+0.239Lluvia-Tormenta...
0.67014	0.02482	0.9103	-0.554Raf+0.457Lluvia-0.44Mes-0.343Nada-0.161VieA...
0.55574	0.02058	0.93088	0.649Clo+0.429Año+0.335VisiI-0.264VisiX-0.236HumX...
0.45225	0.01675	0.94763	-0.476HumX-0.475VieX+0.291Raf-0.234Lluvia-0.226TempX...
0.43375	0.01606	0.9637	-0.473Mes+0.408Raf-0.295VieA-0.284HumI-0.278Lluvia-Tormenta...

El *PrincipalComponents* extrae 12 combinaciones lineales de las variables originales que explican el conjunto de datos CD_2 en un 96%.

CD_7: 12 vectores lineales PC del conjunto de datos CD_2.

V1, V2, V3, …,V12, CT

Para crear estos nuevos conjuntos de datos, se multiplican las matrices de los vectores de *eigen* resultantes de cada conjunto de datos con las matrices del conjunto de datos originales correspondientes.

2.1.2 Limpiar los datos.

La limpieza de los datos aumenta la calidad de los datos al nivel requerido por las técnicas de análisis de datos. Esta puede incluir la selección de un conjunto de datos limpios, la sustitución de los valores perdidos con valores adecuados.

En este caso todos los valores perdidos del conjunto de datos se sustituyeron con la media de la temporada correspondiente. Este valor se calculó utilizando la función calcularpromedio(x). Ver capítulo 3 para mayor información. .

2.1.3 Construcción de datos.

La variable eventos (Ev) es una variable que supuestamente tiene muchos valores perdidos, pero esto no es así. Muchos de los valores perdidos de esta variable no son por ausencia de información, es la no ocurrencia de ningún evento en ese momento. La variable se volvió a codificar, lo que provocó la aparición de un nuevo valor que corresponde a la no ocurrencia de eventos. El valor NADA, permite la diferenciación entre el desconocimiento del valor y la no ocurrencia de evento alguno. Los valores que toma la variable NADA se calcularon utilizando la función nada (num). Consulte el capítulo 3 para más información.

La variable evento es una variable discreta. Que puede tomar los valores siguientes: NADA, lluvia, lluvia-tormenta, tormenta, niebla, niebla-lluvia. Para tener una mejor comprensión de la variable evento se crearon seis variables numéricas una para cada valor del dominio y los valores se calcularon usando la siguiente fórmula (F(x, j) donde x es alguno de los valores posible del dominio de la variable Evento).

$$F(x, j) = \frac{\text{cantidad de veces que tomó el valor x en el mes j}}{\text{Cantidad total de casos en el mes j.}} \quad \text{.....(2.1)}$$

Los valores que toma cada variable creada a partir de la variable evento, son probabilidades de ocurrir el evento correspondiente a cada variable en cada mes del período considerado.

Este cálculo se hizo utilizando la función testing(x). Ver capítulo 3 para información.

2.1.4 Integrar datos.

Al comienzo los datos climatológicos y los datos energéticos se contenían en dos tablas diferentes. Para formar el conjunto de datos inicial se integraron las dos tablas formando una sola que contiene nuevos registros. Cada uno de estos registros se conforma con los dos registros correspondientes de cada tabla anterior.

2.1.5 Formatear los datos en formato ARFF.

La herramienta utilizada en este trabajo para minería de datos (WEKA) sólo permite un formato de archivos. El formato arff. Como el conjunto de datos inicial estaba en formato xls (hojas de Excel), primeramente tenía que convertirse a formato dlm, un archivo que contiene datos separados por comas y luego convertirse al formato arff. Los atributos pueden ser principalmente de dos tipos: numéricos de tipo real o entero (*real* o *integer*), y simbólicos (especificando los valores posibles que pueden tomar entre llaves).

La estructura que debe tener este archivo para poder ser leído por WEKA es la siguiente:

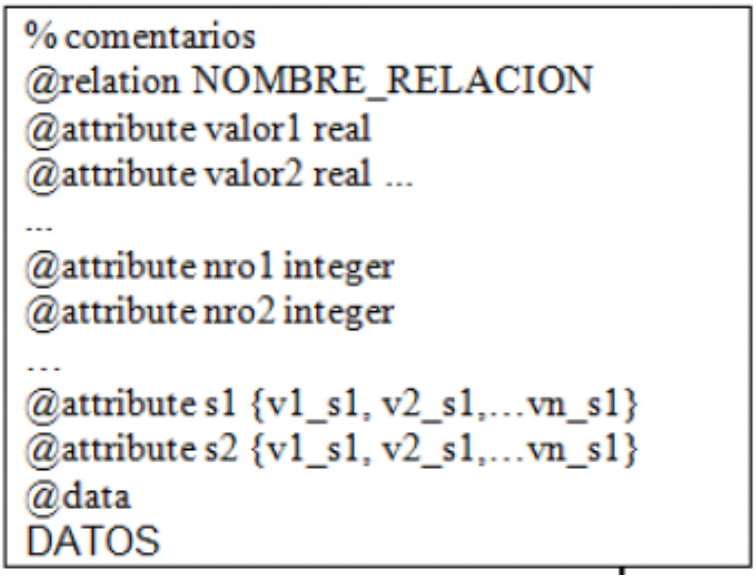


Figura 2.1: Estructura de archivo .arff

Esta conversión de formato se hizo utilizando el algoritmo writetext_final. El código y la descripción del algoritmo está en el capítulo 3.

```
@relation CD_4__Rcfs_condatoscomerciales_mayor75
@attribute HDO_N numeric
@attribute TempX numeric
@attribute TempA numeric
@attribute TempI numeric
@attribute RocX numeric
@attribute RocA numeric
@attribute RocI numeric
@attribute VisiX numeric
@attribute VisiA numeric
@attribute VisiI numeric
@attribute VieX numeric
@attribute VieA numeric
@attribute Raf numeric
@attribute Tormenta numeric
@attribute Lluvia-Tormenta numeric
@attribute Nada numeric
@attribute CT numeric
@data
427,28.5161,24.1935,19.9032,19.4839,17.9355,16.0645,9.83871
433,28.75,24.0357,19.1786,19.4643,17.1429,15.4286,9.75,9.092
484,29.6129,24.5161,19.6129,18.5484,16.5161,14.4839,9.83871
```

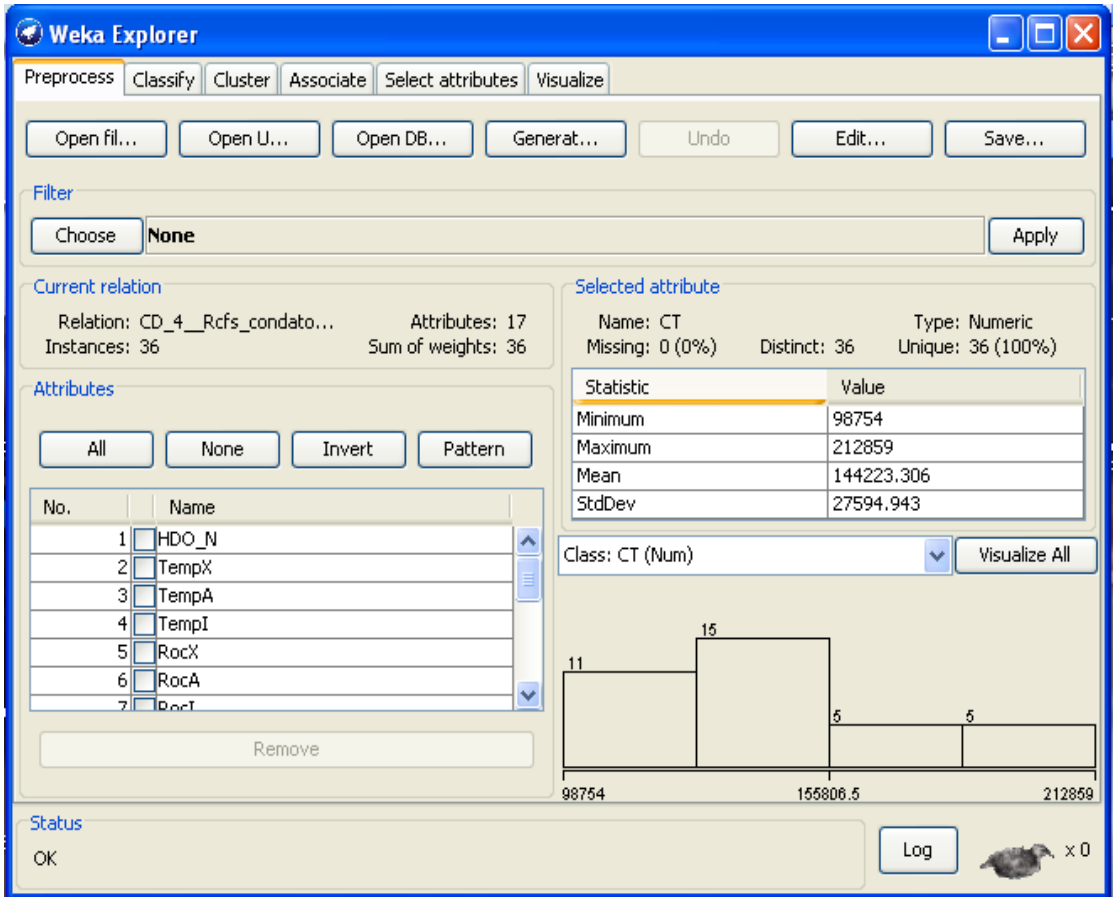


Figura 2.2: Ambiente WEKA después de introducidos los datos: CD_4.

2.2 Construir y correr el modelo.

Para la construcción del modelo se utilizaron dos modelos de regresión que están implementados como funciones en WEKA:

LinearRegression

MultilayerPerceptron

Los conjuntos de datos resultantes del proceso de selección de datos, CD_3, CD_4, CD_5, CD_6 y CD_7, se corrieron con los métodos de clasificación anteriormente mencionados.

2.2.1 *LinearRegression.*

Este método utiliza regresión lineal para la predicción. Usa el criterio akaike[8] para la selección de los modelos, es capaz de trabajar con instancias que tienen pesos asignados.

Se utilizó la *LinearRegression* porque:

- Cuando la salida o la clase es numérica, y todos los atributos son numéricos, la regresión lineal es una técnica natural para considerar.
- Es un modelo muy simple.

Configuración utilizada:

attributeSelectionMethod:*M5 method*

debug:False

eliminateColinearAtributes:True

ridge:1.0E-8

Para el significado de cada opción (ver Anexo 1).

2.2.2 *MultilayerPerceptron*

Es un modelo de regresión que usa retro propagación para pronosticar instancias. Esta red se puede crear a mano o creando un algoritmo o ambos. La red puede ser monitoreada y modificada durante el tiempo de entrenamiento. Todos los nodos de esta red son de tipo *sigmoid* (excepto para cuando la clase es numérica, los nodos de salida se convierten en unidades lineales sin umbrales). El *MultiLayerPerceptron* tiene retro propagación que utiliza el procedimiento gradiente decente *(descent)* como método de aprendizaje y su función de transferencia es la función sigmoide:

[8] El criterio de información de Akaike es una medida relativa de mejor ajuste (*goodness of fit*) de un modelo estadístico. Fue desarrollado por Hirotsugu Akaike.

$$f(x)=\frac{1}{1+e^{-x}}.$$ …….(2.2)

La función de error cuadrático de entrenamiento para una sola instancia es:

$$E=\frac{1}{2}(y-f(x))^2$$ ……(2.3)

Donde f(x) es la predicción de la red obtenida por la unidad de salida y (y), es el valor de la clase que pertenece a la instancia.[11]

Se utilizó el *MultilayerPerceptron* porque:

- El aprendizaje adaptativo: Es la habilidad de aprender cómo hacer tareas basado en los datos dados en su entrenamiento.
- Está probado que una red neuronal con dos capas con suficientes nodos ocultos es un aproximador universal.[12],[13]

Configuración utilizada:

En el caso del campo *hiddenLayers* se debía buscar mejor combinación de las capas ocultas y de las neuronas en cada capa oculta. Para eso se corrió *MultilayerPerceptron* con diferentes conjuntos de datos variando la combinación de las capas ocultas y la cantidad de neuronas en cada capa oculta en el campo hiddenLayers. La mejor combinación encontrada fue, 3 capas ocultas y en cada capa oculta 20,6 y 3 neuronas respectivamente.

GUI: *true*

autoBuild: true

debug: false

decay:false

hiddenLayers: 20,6,3

learninRate:0.3

momentum:0.2

nominalToBinaryFilter: false

normalizeAttributes: true

normalizeNumericClass: trae

resetTrue

seed:0

traningTime:500

validationSetSize:0

validationThreshold: 20

Para mayor información (ver Anexo 1).

2.2.3 Resultados de las corridas con los métodos de regresión:

MLP:

Tipo de error	CD1	CD3	CD4	CD5	CD6	CD7
Correlation coefficient	0.6933	0.7403	0.7101	0.7593	-0.035	0.0367
Mean absolute error	20917.5939	16874.1215	17722.7253	16910.0631	22757.9275	22147.1914
Root mean squared error	24514.3214	20867.4658	23256.2763	19737.8972	29851.8475	30355.3822

Tabla 2.1: Resultados de las corridas con MLP.

LinearRegression:

Tipo de error	CD1	CD3	CD4	CD5	CD6	CD7
Correlation coeficient	0.323	0.6256	0.7133	0.6457	0.6788	0.49
Mean absolute error	28446.5714	19963.0743	17635.2651	15789.6445	15640.196	19325.7867
Root mean squared error	58341.2654	24760.5596	19894.3075	21317.6872	20068.5881	24179.7882

Tabla 2.2: Resultados de las corridas con LinearRegression.

Como se muestra en la tabla 2.1 el mejor modelo obtenido con *MultilayerPerceptron* es el modelo correspondiente al conjunto de datos CD_5 con un coeficiente de correlación de 0.7593, el error absoluto medio de 16910.0631 y la raíz del error cuadrático medio de 19737.8972.

Según la tabla 2.2 de *LinearRegression* el modelo correspondiente al conjunto de dato CD_4 tiene el mejor coeficiente de correlación y baja *Root mean squared error*, respecto a otros conjuntos de datos. En el caso de *Mean absolute error*, el valor más bajo reporta el modelo correspondiente del conjunto de dato CD_6. Pero la diferencia es aproximadamente 13% que se puede considerar como muy bajo. Por lo tanto se consideró el modelo correspondiente al CD_4 como el mejor modelo en el caso de la *LinearRegression.*

Comparando los mejores modelos de los métodos *MultilayerPerceptron* y de *Linear Regression* se decidió aceptar el modelo correspondiente al conjunto de datos CD_4, de *LinearRegression* por la razón de que el modelo de *LinearRegressi*on es mucho más simple que el modelo de M*ultilayerPerceptron*. Aunque en el caso del modelo de *MultilayerPerceptron* del conjunto de datos CD_5, los valores de los errores son bajos. Pero las diferencias de los errores con el modelo de *Linear Regression* correspondiente al conjunto de datos CD_4, son menos de 5% que se puede considerar como no significativo.

Conclusiones parciales del capítulo

Se realizó la preparación de los datos iniciales. Los datos quedaron limpios a través de las tareas de sustitución de valores perdidos. Se seleccionaron los rasgos más influyentes en la clase. Se crearon los conjuntos de datos que serán utilizados para la corrida de los modelo.

Luego de aplicar los métodos de regresión se concluye que el modelo correspondiente al conjunto de datos CD_4 es el mejor. Este modelo se obtuvo con el método de *LinearRegression.*

CAPITULO 3

Introducción.

En el presente capítulo se describe el mejor resultado alcanzado luego de aplicar los métodos de regresión: *MultilayerPerceptron* y *LinearRegression* a los conjuntos de datos resultantes de las selecciones de atributos y al conjunto de datos original. .Además se describen los parámetros necesarios y la salida de las funciones que se crearon en MATLAB, y que se utilizaron para los cálculos.

3.1 Resultados obtenidos:

El mejor modelo obtenido fue por el método de *LinearRegression* correspondiente al conjunto de datos CD_4.

El mejor modelo obtenido:

CT = 66.7461 * HDO_N +6821.2005 * TempX + (-11220.4888) * RocX +6174.8114 * RocI +(-42069.6058) * VisiA + 14727.1566 * VisiI +

(-3037.3133)* VieX + 4600.5699 * VieA +123736.617 * Lluvia-Tormenta +316637.5

Correlation coefficient	0.7133
Mean absolute error	17635.2651
Root mean squared error	19894.3075
Relative absolute error	80.2832 %

Root relative squared error	70.624 %
Total Number of Instances	36

Tabla 3.1: Errores reportados del modelo de LinearRegression correspondiente al conjunto de datos CD_4.

Entrevistando a los expertos del Departamento de Energía del Hotel Jagua y analizando las investigaciones realizadas anteriormente en este campo en el Hotel, se decidió considerar para su aceptación el modelo para el pronóstico del consumo eléctrico del Hotel Jagua, sólo si éste tenía una precisión de 95% o más.

Según la tabla 3.1 se puede ver que el *Mean absolute error* y el *Root mean squared error* del modelo son significativamente altos. Aunque es el mejor modelo obtenido, el sólo logra una precisión de 20% o menos.

Esto se puede deber a que los datos no son suficientemente buenos .Por ejemplo algunos datos como los comerciales eran resúmenes mensuales de datos primarios, a los que no se pudo acceder. Y además el conjunto homogéneo de datos que se formó sólo abarca un período corto de 2007-2009.

El modelo obtenido por el método de *LinearRegression*, se implementó como una función en MATLAB. Al pasar un vector columna con todas las variables en el orden requerido, la función devuelve el valor del consumo eléctrico. Para mayor información consultar capítulo 3, la función model_regression.

3.2 Funciones en MATLAB:

Las funciones utilizadas para los cálculos fueron implementadas en MATLAB.

3.2.1 Función- valorperdido

```
function [q] = valorperdido(P)

m=0;

for n=1:length(P)
```

```
    a=isnan(P(n));
    if a==1
       m=m+1;
    end
end
q=(m/length(P))*100
return;
```

La función valorperdido se halla en el valor perdido de una matriz columna P, de acuerdo a la fórmula siguiente:

valorperdido=(cantidad de valores vacíos en una matriz columna P/cantidad de casos)*100%

3.2.2 Algoritmo writetext_final.

```
load textdata;
a= rot90(textdata,3);
b={'@attribute'};
n=repmat(b(1),31,1);
C=horzcat(n,a);
fid = fopen('attributes.txt','wt');
for i=1:size(C,1)%cant rows in the matrix C
    fprintf(fid, '%s %s\n', C{ i,:});
end
dlmwrite('attributes.txt',['%Nimali Liyanage , Tutora:Prof.Maia Viera.' 13 10 fileread('attributes.txt')],'delimiter','');
load data;
```

```
dlmwrite('data.txt', data,'-append', 'precision', '%g','newline', 'pc');

dlmwrite('data.txt',['@data' 13 10 fileread('data.txt')],'delimiter','');

!for %f in ("attributes.txt", "data.txt") do type "%f" >> "version1.txt"
```

La función writetext_final crea un archivo de texto con el mismo formato que tiene un archivo arrf. Primero se carga una matriz fila que se llama *text data* que contiene todos los encabezados del conjunto de datos considerados. Luego se escriben estos encabezados en un archivo de texto que se ha creado, llamado *attributes.txt.* Luego se cargan los datos correspondientes a los encabezados y se escriben en un archivo de texto llamado *data.txt.* Finalmente se combinan estos dos archivos de textos y se crea un nuevo archivo de texto llamado *version1.txt.*

3.2.3 Función-testing

```
function [r] = testing(X)

m=0;

for n=1:length(X)

   if strcmpi(X(n),'lluvia')

      m=m+1;

   end

end

r=m/length(X)

return;
```

La función testing, se utiliza para hallar el valor de cada atributo derivado del atributo evento, de acuerdo a la siguiente fórmula:

Valor del atributo y=(cantidad de veces que aparece (y), en la matriz/la cantidad total de casos en la matriz).

La matriz que se necesita como parámetro es una matriz columna con los datos pertenecientes al atributo evento.

3.2.4 Función-nada

```
function [r] = nada(num)

m=0;

for n=1:length(num)

   a=isnan(num(n));

   if a==0

      m=m+num(n);

   end

end

r=1-m

return;
```

La función nada se utiliza para buscar el valor que toma el atributo NADA en cada mes. El parámetro que debe pasar es un vector columna llamado num, que contiene todos los valores del mes a considerar de los atributos derivados en el atributo evento.

3.2.5 Función calcularpromedio.

```
function [Z] = calcularpromedio(x)
B = sum(x);
A=size(x,1);
Z=B./A;
End
```

La función calcularpromedio se halla la media para cada columna, de una matriz x, de acuerdo a la fórmula siguiente:

Media = Suma de los valores de columna j de la matriz x

Cantidad de elemntos en la columna j en la matriz x

La matriz x puede ser una matriz columna o una matriz con i filas y j columnas. La fución devuelve una matriz fila con las medias correspondientes a cada columna j de matrz x.

3.2.5 Función model_regression

```
function [q] = model_regression(P)

   HDO_N=P(1,:);

   TempX=P(2,:);

   RocX=P(3,:);

   RocI=P(4,:);

   VisiA=P(5,:);

   VisiI=P(6,:);

   VieX=P(7,:);

   VieA=P(8,:);

   LTor=P(9,:);

     [CT]=(66.7461*HDO_N)+(6821.2005*TempX)-
(11220.4888*RocX)+(6174.8114*RocI)-(42069.6058*VisiA)+(14727.1566*VisiI)-
(3037.3133*VieX)+(4600.5699*VieA)+(123736.617*LTor)+(316637.5);

   disp(['El consumo total del mes es: ',num2str(CT) ' kWh'])

end
```

La función model_regression calcula el consumo eléctrico .Debe pasar como parámetro un vector columna con los valores correspondientes de las variables HDO_N,TempX, RocX, RocI, VisiA,VisiI,VieX, VieA, Lluvia-Tormenta manteniendo este mismo orden.

Conclusiones Parciales del capítulo.

El mejor modelo obtenido por el método *LinearRegression* logra una precisión aproximadamente de 20%.

Las funciones implementadas en MATLAB facilitaron los cálculos necesarios del trabajo.

Conclusiones.

Se obtuvo un modelo matemático para la predicción del consumo eléctrico.

El modelo obtenido a partir de los datos comerciales y climatológicos del Hotel Jagua no supera la precisión del 95% para la predicción del consumo eléctrico mensual en el hotel.

Según el estudio realizado se concluye que los datos climatológicos por sí solos no logran explicar el consumo total del Hotel Jagua.

Recomendaciones.

Es necesario acceder a los datos primarios de las variables comerciales, pues en este trabajo se utilizaron los resúmenes mensuales de indicadores bastante generales.

Conseguir los datos recientes de las variables y volver a aplicar las técnicas de minería de datos.

Analizar si el hotel mide otras variables, y en caso afirmativo incorporarlas al estudio.

Intentar encontrar modelos por separado para todas las variables objetivos e integrarlos, entiéndase un modelo para el horario pico, para el horario de la madrugada y el horario del día.

Referencia Bibliográfica.

[1] Insttituto Tecnológico Hotelero, “La importancia de la eficiencia energética para los hoteles,” 2009.

[2] Luis A. Marín Llanes and Juan C. Carro Cartaya, “LA MINERÍA DE DATOS COMO HERRAMIENTA EN EL PROCESO DE INTELIGENCIA COMPETITIVA..”

[3] Mark A. Hall, “Correlation-based Feature Selection for Machine Learning,” University of Waikato., 1999.

[4] U. M. Fayyad and K. B. Irani, *Multi-interval discretisation of continuous-valued attributes for classification learning.*, 1993.

[5] Karl Pearson, “On Lines and Planes of Closest Fit to Systems of Points in Space.,” *Phylosophical Magazine.*, 1901, pp. 559-572.

[6] Janny Hermosa Morell Díaz, “Propuesta de técnica de Inteligencia Artificial para la detección de anomalías en la red de datos de la Universidad de Cienfuegos.,” Univercidad de Cienfuegos-Cuba, 2010.

[7] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Pe- ter Reutemann, and Ian H. Witten, Nueva zelanda: Universidad de Waikato, .

[8] José Carlos Escobar Palacio, “ANALISIS ESTACIONAL DEL COMPORTAMIENTO ENERGÉTICO
DEL HOTEL JAGUA.,” 2004.

[9] Aníbal E. Borroto Nordelo and José P. Monteagudo Yanes, “Gestión y economía energética.,” 2006.

[10] Isdel Geroy Borlado., “Propuesta de un Sistema de Monitoreo y Control Energético en el Hotel Gran Caribe Jagua de Cienfuegos,” 2009.

[11] Ian H. Witten and Eibe Frank, *Data mining : practical machine learning tools and techniques*, The Morgan Kaufmann, .

[12] G.Cybenko, “aproximation by superpositions of a sigmoidal function,” 1989.

[13] K.Hornik, M.Stinchcombe, and H.White, “Multilayer feedforward networks are universal approximators,” 1989.

Bibliografías.

[1] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Pe- ter Reutemann, and Ian H. Witten, Nueva zelanda: Universidad de Waikato, .

[2] U. M. Fayyad and K. B. Irani, *Multi-interval discretisation of continuous-valued attributes for classification learning.*, 1993.

[3] "[1201.0633] General bound of overfitting for MLP regression models."

[4] Yiming Yang and Jan O. Pedersen, "A Comparative study on Feature Selection in Text Categorization.."

[5] Brian R. Hunt, Ronald L.Lipsman, and Jonathan M.Rosenberg, "A Guide to Matlab for beginners and Experienced Users."

[6] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga., "A review of feature selection techniques in bioinformatics.," 2005.

[7] Jonathon Shlens, "A Tutorial on Principal Component Analysis," Apr. 2009.

[8] Lindsay I Smith, "A tutorial on Principal Components Analysis," Feb. 2002.

[9] Isabelle Guyon and André Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, 2003.

[10] María García Jiménez and Aránzazu Álvarez Sierra, "Análisis de Datos en WEKA – Pruebas de Selectividad."

[11] José Carlos Escobar Palacio, "ANALISIS ESTACIONAL DEL COMPORTAMIENTO ENERGÉTICO DEL HOTEL JAGUA.," 2004.

[12] "Analysis Ch2 - Data Cleaning."

[13] Zaily Alfonso Cuencio, "Aplicación de Técnicas de Clasificación de Inteligencia Artificial en el Pronóstico de Temperaturas en el Centro Meteorológico de Cienfuegos.," Univercidad de Cienfuegos-Cuba, 2011.

[14] G.Cybenko, "aproximation by superpositions of a sigmoidal function," 1989.

[15] "cleaning.pdf (Objeto application/pdf)."

[16] “cleaning-unece.pdf (Objeto application/pdf).”

[17] K. Selvakuberan,, M. Indradevi, and Dr.R.Rajaram, “Combined Feature Selection and classification – A novel approach for the categorization of web pages,” Apr. 2008.

[18] Mark A. Hall, “Correlation-based Feature Selection for Machine Learning,” University of Waikato., 1999.

[19] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth, “CRISP-DM 1-
Step-by-step data mining guide,” Aug. 2000.

[20] Ian H. Witten and Eibe Frank, *Data mining : practical machine learning tools and techniques*, The Morgan Kaufmann, .

[21] Petr Somol, Jana Novovi˘cová, and Pavel Pudil, “Efficient Feature Subset Selection
and Subset Size Optimization.”

[22] “Estimation of Prostate Cancer Probability by Logistic Regression: Free and Total Prostate-specific Antigen, Digital Rectal Examination, and Heredity Are Significant Variables.”

[23] Dennis W. Ruck, Steven K. Rogers, and Matthew Kabrisky, “Feature Selection Using a Multilayer Perceptron,” *Journal of Neural Network Computing,*, vol. 2, Nov. 1989, pp. 40-48.

[24] Ira Cohen, Qi Tian, Xiang Sean Zhou, and Thomas S. Huang, “Feature Selection Using Principal Feature
Analysis.”

[25] J.L. CUBERO, F. BERZAL, and F. HERRERA, “FUNDAMENTOS DE MINERÍA DE DATOS.”

[26] Aníbal E. Borroto Nordelo and José P. Monteagudo Yanes, “Gestión y economía energética.,” 2006.

[27] Zikander M. Mirza, “Introduction to Matlab.”

[28] “Journal of Medical Systems, Volume 29, Number 3 - SpringerLink.”

[29] Insttituto Tecnológico Hotelero, “La importancia de la eficiencia energética para los hoteles,” 2009.

[30] Luis A. Marín Llanes and Juan C. Carro Cartaya, “LA MINERÍA DE DATOS COMO HERRAMIENTA EN EL PROCESO DE INTELIGENCIA COMPETITIVA..”

[31] Tom Mitchell, *Machine Learning*.

[32] Salvador Ramírez, “Matlab,” Mar. 2002.

[33] David Kuncicky, “Matlab programming.”

[34] Jialong He, “MATLAB Quick Reference.”

[35] “Microsoft Linear Regression Algorithm Technical Reference.”

[36] K.Hornik, M.Stinchcombe, and H.White, “Multilayer feedforward networks are universal approximators,” 1989.

[37] Karl Pearson, “On Lines and Planes of Closest Fit to Systems of Points in Space.,” *Phylosophical Magazine.*, 1901, pp. 559-572.

[38] “PLoS ONE: Artificial Neural Networks in the Outcome Prediction of Adjustable Gastric Banding in Obese Women.”

[39] Mark A. Hall and Lloyd A. Smith, “Practical Feature Subset Selection for Machine Learning.”

[40] Janny Hermosa Morell Díaz, “Propuesta de técnica de Inteligencia Artificial para la detección de anomalías en la red de datos de la Universidad de Cienfuegos.,” Univercidad de Cienfuegos-Cuba, 2010.

[41] Isdel Geroy Borlado., “Propuesta de un Sistema de Monitoreo y Control Energético en el Hotel Gran Caribe Jagua de Cienfuegos,” 2009.

[42] “Support vector regression to predict porosity and permeability: Effect of sample size 10.1016/j.cageo.2011.06.011 : Computers & Geosciences | ScienceDirect.com.”

[43] “Waikato Environment for Knowledge Analysis (WEKA).”

[44] William F. Ryan, “Weka: A Useful Tool for Air Quality Forecasting,” 2007.

[45] Ian H. Witten, Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, and Saly Jo Cunningham, “Weka:Practical Machine learning tools and Tecniques with Java Implementations,” Aug. 1999.

[46] Rossen Dimov, “Weka:Practical Machine learning tools and Tecniques with Java Implementations,” Apr. 2007.

[47] Manu Gupta, “Why MultiLayer perceptron.”

Anexo1

Métodos utilizados para la selección y de extracción de atributos.

CfsSubset Eval

Opciones

Opción	**Descripción**
locallyPredictive	Identifica atributos predictivos locales. Iterativamente añade atributos que tengan mayor correlación con la clase, mientras no tenga un atributo en el subconjunto que tenga mayor correlación con el atributo que ya está.
missingSeperate	Trata valores perdidos como un valor separado (*separate value*). En otro caso cuenta los valores perdidos que están distribuidos entre otros valores en la proporción de su frecuencia.

Capacidades

.**Capacidad**	**Soportado**
Class	Valores perdidos en la clase, clase numérica, clase nominal, clase de tipo fecha, clase binaria.
Attributes	Atributos nominales vacíos, atributos nominales, atributos numéricos, atributos unarios, atributos de tipo fecha, atributos binarios, valores perdidos.
Min # of instances	1

PrincipalComponents

Opciones

Opción	**Descripción**
maximumAttributeNames	El máximo número de atributos a incluir en los nombres de los atributos transformados.
normalize	Normaliza datos de entrada.
transformBackToOriginal	Transforma a través del espacio PC y regresa al espacio original. Si sólo los mejores n componentes principales están retenidos (configurando varianceCovered<1), esta opción dará un conjunto de datos en el espacio original pero con menos ruido de atributos.
varianceCovered	Retiene suficientes atributos PC de acuerdo a la proporción de la varianza.

Capacidades

Capacidad	**Soportado**
Class	Clase fecha, clase numérica, clase binaria, clase nominal, valores pérdidas de clase, no clase.

Attributes	Atributos numéricos, atributos nominales , Atributos binarios, valores perdidos, atributos fecha, atributos nominales vacíos, atributos unarios
Min # of instances	1

Modelos de regresión utilizadas en WEKA.

MultilayerPerceptron

Opciones

.**Opción**	**Descripción**
GUI	Muestra la interface GUI. Esta opción brinda la posibilidad de pausar y alterar la red neuronal mientras esté entrenando. Click izquierdo- se usa para adicionar un nodo (este nodo se seleccionará automáticamente, cuando se da click no debe haber ningún nodo seleccionado). Para seleccionar un nodo se debe dar click izquierdo encima del nodo. Para conectar un nodo, primero se debe seleccionar el nodo(s) inicio(s), luego dar click en el nodo final o un espacio vacío (este creará un nodo que está conectado con los nodos seleccionados.) después de hacer las conexiones entre ellos, los nodos seguirán seleccionados. (Estos son conexiones dirigidas, una conexión entre dos nodos se establecerá sólo una vez y no establecerá

	conexiones que no sean válidas.) Para eliminar una conexión se selecciona un nodo(s) conectado(s) en la conexión que se quiera eliminar y luego se da click derecho en el otro nodo (no es importante si el nodo es de inicio o de final, la conexión será removida) . Para eliminar un nodo se da click derecho encima del nodo, mientras no haya otro nodo seleccionado (incluyendo el nodo que se quiere eliminar). También se eliminarán las conexiones a este nodo. Para deshacer una selección se da click izquierdo mientras esté presionada la tecla control o se da click derecho en un espacio vacío. Las instancias se pueden proveer con las etiquetas de la izquierda. Los nodos rojos son de capas ocultas. Los nodos naranja son nodos de salida. Las etiquetas que están en la derecha muestran la clase que representa el nodo de salida, con una clase numérica el nodo de salida se convierte en una unidad lineal sin umbral. Se puede hacer alteraciones a una red neuronal mientras la red no esté corriendo. Esto también se aplica al campo learning rate y a los otros campos del panel de control. Puede dar click en aceptar y la red termina en

	ese momento. La red está pausada automáticamente en el comienzo. Muestra algunas indicaciones mientras la red está corriendo, como: en que iteración la red actualmente está y cuál fue el error (aproximado) para esa iteración(o para la validación si es usado). Observar que el valor del error está basado en una red que está cambiando mientras se va calculando el valor. (También depende si la clase es normalizada en caso del error reportado para la clase numérica.) Cuando la red termina de entrenar será pausada automáticamente esperando para ser aceptada o para continuar el entrenamiento. Si la GUI no está marcada en la configuración, la red no requiere ninguna interacción.
autoBuild	Adiciona y conecta las capas ocultas en la red.
debug	Si está configurado como true , el clasificador puede dar algunas informaciones adicionales a la consola (panel de control).
decay	Este causará decrecimiento de la tasa de aprendizaje (*learning rate*). Este divide la velocidad de aprendizaje inicial entre la cantidad de iteraciones para determinar cuál será la velocidad de aprendizaje actual. Este ayuda a la red para que no haya divergencias respecto a las salidas deseadas y además mejora la actuación

	general. El decrecimiento de la tasa de aprendizaje no se muestra en GUI, sólo la tasa de aprendizaje inicial. Si se cambia la tasa de aprendizaje usando GUI, es tratada como la tasa de aprendizaje inicial.
hiddenLayers	Define las capas ocultas de la red neuronal. Es una lista de números enteros positivos, esta lista debe ser separada por coma. Para no tener capas ocultas en la red debe poner un solo 0 en este campo. Este sólo se debe usar si se está utilizando autobuild. También hay valores *wildcard* 'a' = (atributos + clases) / 2, 'i' = atributos, 'o' = clases, 't' = atributos + clases.
learningRate	El rango entre pesos actualizados.
momentum	El impulso aplicado en el momento de actualización de los pesos.
nominalToBinaryFilter	Este preprocesará las instancias con el filtro, ayudando a mejorar el comportamiento si hay atributos nominales en los datos.
normalizeAttributes	Normaliza los atributos y puede ayudar a mejorar la actuación de la red. No sólo normaliza los atributos numéricos sino también atributos nominales (luego estos atributos corren a través del filtro nominal to binary si está en uso). Entonces los valores nominales serán entre -1 y 1.
normalizeNumericClass	Normaliza la clase numérica y así se puede mejorar la actuación de la red. Normaliza la clase que esté entre -1 y 1.este proceso sólo es interno y la salida será la escala original.

randomSeed	Seed es usado para iniciar el generador de números aleatorios. Los números aleatorios son usados para establecer los pesos iniciales en las conexiones entre nodos y también para mezclar datos de entrenamiento.
reset	Permite a la red establecer una velocidad baja de aprendizaje. Si la red diverge de la solución este campo se restablecerá automáticamente con una baja tasa de aprendizaje y empezará a entrenar nuevamente. Esta opción sólo está disponible cuando GUI no está establecido. Si la red diverge y no tiene permiso para restablecerse, fallará el proceso de entrenamiento y devolverá un mensaje de error.
trainingTime	Número de iteraciones para entrenar. Si la configuración de la validación no es cero entonces se puede terminar el proceso de la red tempranamente.
validationSetSize	El tamaño del conjunto de validación en porcentaje. (el entrenamiento continuará hasta que la observación del error de validación esté constantemente empeorando o alcance la cantidad de iteraciones establecidas). Si éste está en 0 no se usa un conjunto de validación y la red entrenará hasta la cantidad de iteraciones especificadas.
validationThreshold	Usado para terminar las pruebas de validación. El valor que está establecido en este campo dice

	cuántas veces en una fila se puede empeorar el error de validación antes de terminar el entrenamiento.

Capacidades

.**Capacidad**	**Soportado**
Class	Clase nominal, clase binaria, valores perdidos de clase, clase de tipo fecha, clase numérica.
Attributes	Atributos nominales vacíos, atributos numéricos, atributos binarios, valores perdidos, atributos fecha, atributos unarios, atributos nominales.
Min # of instances	1

LinearRegression.

Opciones

Opción	**Descripción**
attributeSelectionMethod	Selecciona el método usado para la selección de atributos y para la utilización en regresión lineal. Los métodos disponibles son: no selección de atributos, selección de atributos usando el método de M5(avanza a través de atributos eliminando el que tiene el mínimo coeficiente estandarizado y ningún mejoramiento observado en el error estimado dado por el criterio de

	información akaike) y una selección ávida(*greedy selection*) usando la métrica de información Akaike(*Akaike information metric*)
debug	Entrega las informaciones de *debug* a la consola.
eliminateColinearAttributes	Elmina atributos colineales.
ridge	El valor del parámetro *ridge*.

Capacidades

Capacidad	**Soportado**
Class	Valores perdidos de clase, clase numérica, clase de tipo fecha.
Attributes	Atributos numéricos, atributos nominales, atributos binarios, atributos fecha, valores perdidos, atributos nominales vacíos, atributos unarios.
Min # of instances	1

Anexo 2

Corridas en *PrincipalComponents*- CD_1

Los vectores de *eigen* resultantes.

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	
-0.1353	-0.187	-0.2008	0.1704	0.1452	0.0461	-0.0184	0.1212	0.0056	0.0913	0.0019	0.4046	-0.3952	0.2568	Mes
0.0479	-0.25	0.0873	-0.3013	-0.2122	-0.1126	0.2493	0.2771	0.103	0.0866	-0.245	-0.0568	0.2193	0.1878	Año
0.091	0.0788	-0.35	0.1982	0.1539	0.2201	0.0818	-0.1713	-0.3438	-0.2146	-0.1093	-0.0823	-0.0547	-0.069	PD
0.2069	0.1601	-0.0786	0.1639	-0.057	-0.0853	-0.0606	0.2327	0.0166	-0.0341	-0.0434	0.1229	0.1439	0.2194	TF_E
-0.0684	0.3623	0.2087	0.1235	-0.0152	0.1208	0.1347	0.0195	-0.1007	0.0242	0.1281	-0.0292	-0.0824	-0.1043	TF_N
0.2093	0.1467	-0.0542	0.1689	0.0397	0.1276	0.0634	0.1752	0.0155	-0.1165	0.0061	0.1612	0.1591	-0.2753	TD_E_P
0.026	0.0632	-0.3672	0.0642	0.0236	0.2102	-0.2558	-0.2944	0.2545	0.2195	-0.1264	0.0203	0.1866	0.0725	TD_E_O
-0.1122	0.3655	0.0421	0.0729	0.0478	-0.221	0.2146	0.0527	-0.089	-0.1203	0.0527	0.0925	0.0371	0.01	TD_N
0.091	0.0788	-0.35	0.1982	0.1539	0.2201	0.0818	0.1713	-0.3438	-0.2146	-0.1093	-0.0823	-0.0547	-0.069	HDD
0.2132	0.1349	-0.0579	0.1724	-0.033	0.1359	0.0499	0.1506	0.0328	-0.1133	-0.0669	-0.1618	-0.1494	-0.2626	HDO_E_P
-0.0021	0.0537	-0.3419	0.1305	0.1527	-0.111	-0.4095	0.0065	-0.087	-0.3961	-0.1691	-0.0049	-0.2553	-0.143	HDO_E_O
-0.0852	0.3759	0.1405	-0.0857	-0.0428	0.1651	0.1408	-0.0716	-0.1662	-0.0629	0.1516	0.053	-0.1679	0.0022	HDO_N
-0.2034	0.1441	0.0425	-0.2272	0.0055	0.2597	0.0495	0.0135	-0.0018	0.082	-0.12	-0.021	-0.0544	-0.0273	TempX
-0.2254	0.1209	-0.0168	-0.1669	0.0076	0.1885	0.0689	0.0173	-0.0122	-0.0899	-0.0605	-0.0077	-0.066	-0.0362	TempA
-0.2337	0.1142	-0.0043	-0.0944	-0.0273	0.1498	-0.077	-0.0076	-0.0297	-0.0939	-0.0303	-0.0046	-0.0639	-0.018	TempI
-0.2466	0.022	-0.004	-0.0788	-0.0189	0.0882	0.0799	0.0108	-0.0015	-0.1315	-0.0103	-0.0054	-0.0711	-0.0326	RocX
-0.2492	-0.0043	-0.0039	-0.0646	-0.0178	0.0473	-0.0818	-0.0354	-0.0161	-0.0934	-0.0113	-0.0127	-0.0852	-0.0379	RocA
-0.2493	-0.0103	-0.0099	-0.0566	-0.026	0.0351	-0.0762	-0.0252	-0.0009	-0.0864	0.003	-0.0117	-0.1032	-0.0468	RocI
-0.1422	0.3129	-0.0427	-0.0874	-0.0268	0.1402	-0.006	-0.0154	-0.0084	-0.2006	-0.0722	-0.0232	-0.2614	-0.1907	HumX
-0.1992	-0.0509	-0.0005	-0.1958	-0.0482	-0.0509	-0.0005	-0.0005	-0.0005	-0.06	-0.0509	-0.043	-0.0509	-0.0509	HumA

	0.1863	0.0261			0.2486			0.0513		0.0479		0.0672	0.2101	
	-	-			-		-	-	-	-		-	-	
-0.1973	0.1426	0.0692	0.2425	0.0176	0.2243	0.0571	0.0104	0.0384	0.0477	0.1361	0.0393	0.1209	0.1579	HumI
			-		-			-				-		
0.2083	0.0071	0.1795	0.0136	0.1786	0.2551	0.1336	0.0559	0.0338	0.025	0.1389	0.0025	0.0283	0.0027	PresX
			-		-			-			-	-		
0.1977	0.0347	0.2004	0.0524	0.1783	0.2603	0.1702	0.093	0.0369	0.026	0.1541	0.0153	0.0147	0.0647	PresA
			-		-			-			-	-		
0.1865	0.0535	0.2125	0.0329	0.1911	0.2755	0.1881	0.1247	0.0473	0.0167	0.1678	0.0102	0.0084	0.1078	PresI
	-	-		-		-	-					-	-	
0.1076	0.2348	0.2034	-0.046	0.0579	-0.01	0.3442	0.1353	0.2639	0.1963	0.0289	-0.18	0.0126	0.2069	VisiX
	-	-				-	-					-		
0.1053	0.2221	0.2851	-0.163	0.2162	0.0364	0.2207	0.0148	0.1434	0.1185	0.1882	-0.1	0.0227	-0.046	VisiA
	-	-				-			-		-			
0.1007	0.1199	0.3112	-0.141	0.369	0.0474	0.1337	0.0993	0.0284	0.0778	0.1093	0.0826	0.0531	0.2607	VisiI
			-	-				-				-		
0.1698	0.0482	0.0528	0.0931	0.3456	0.2438	0.0173	-0.115	0.0381	0.2194	0.175	0.1387	0.2559	0.1525	VieX
		-		-		-	-					-		
0.222	0.0615	0.0274	0.0768	0.1438	0.1029	0.0166	0.1409	0.0044	0.094	0.1907	0.1531	0.2523	0.0905	VieA
	-			-	-				-				-	
-0.0134	0.1446	0.1081	0.1304	0.4461	0.0908	-0.114	0.3094	0.1739	0.0713	0.2616	0.4683	0.3185	0.0908	Raf
		-		-				-	-		-			
-0.1234	0.0502	0.0449	0.3989	0.0186	0.0027	0.0273	0.1023	0.0137	0.1749	0.405	0.2926	0.0707	0.4872	Clo
	-	-		-			-	-			-			
-0.1409	0.1605	0.0425	0.2613	0.2212	0.0358	-0.089	0.0957	0.3515	0.2034	0.0779	0.3164	0.2005	0.0748	Lluvia
		-	-	-	-	-						-	-	
-0.156	0.147	0.0346	0.1768	0.0354	0.2427	0.1629	0.0419	0.2103	0.0632	0.4755	0.0129	0.2062	0.1491	Tormenta
	-		-	-	-		-		-	-	-	-		
0.0899	0.0271	0.0885	0.0681	0.2884	0.2039	0.0686	0.3125	0.5072	0.4349	0.1415	0.2497	0.1884	0.2165	Niebla
			-	-	-						-	-	-	
-0.1676	0.0943	0.0448	0.2921	0.0339	0.0916	0.1377	0.3135	0.0384	0.2593	0.0454	0.1931	0.0751	0.1317	Lluvia-Tormenta
	-		-	-	-	-		-	-	-	-	-		
0.0833	0.0139	0.032	0.1448	0.1733	0.0804	0.4673	0.4846	0.2697	0.2128	0.2702	0.0785	0.3128	0.2213	Niebla-Lluvia
		-					-		-	-				
0.2099	0.0058	0.0101	0.0838	0.2375	0.1531	0.0272	0.0878	0.0305	0.1933	0.1777	0.3628	0.0362	0.0201	Nada

Corridas en *PrincipalComponents*- CD_2

Los vectores de *eigen* resultantes.

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	
-0.1504	0.3244	-0.037	0.1556	-0.0155	0.1102	-0.0063	0.0023	-0.4405	-0.0086	-0.149	-0.4731	Mes
0.0517	0.1954	-0.111	-0.4508	-0.2962	-0.0528	0.3483	0.2653	0.0471	-0.4289	-0.047	-0.1772	Año
-0.2229	-0.2484	-0.2025	-0.0718	0.0674	-0.0468	0.0459	0.015	-0.1123	0.0063	-0.226	0.2283	TempX
-0.2476	-0.1948	-0.1605	0.0358	0.0311	0.0323	0.0483	-0.0148	-0.09	0.0415	-0.15	0.127	TempA
-0.2574	-0.1584	-0.1195	-0.0247	0.0363	-0.03	0.0389	-0.0343	-0.0727	0.051	-0.068	0.0789	TempI
-0.2711	-0.0621	-0.0749	-0.0124	0.0388	0.0732	0.0614	-0.0459	-0.0706	0.0241	-0.102	0.0869	RocX
-0.2737	-0.0381	-0.0533	-0.0138	-0.0679	-0.04	0.0587	-0.0181	-0.0779	-0.0101	-0.106	0.0543	RocA
-0.2733	-0.0166	-0.052	-0.0105	-0.0764	-0.0544	0.0562	-0.0258	-0.0721	-0.0094	-0.108	0.0216	RocI
-0.1553	-0.301	-0.174	-0.041	-0.24	-0.0955	0.0659	-0.0901	-0.1143	-0.2359	-0.476	0.1738	HumX
-0.2184	-0.2464	-0.2028	-0.0905	-0.2003	-0.0363	-0.0013	-0.0231	0.0608	-0.1257	0.0415	-0.0762	HumA
-0.218	-0.2362	-0.2004	-0.1122	-0.1115	-0.0462	0.0163	-0.0697	0.0003	-0.0637	0.1601	-0.2835	HumI
-0.2337	-0.0833	-0.1214	-0.1311	-0.3345	-0.0459	-0.0693	0.0959	-0.0151	-0.0663	-0.15	0.0972	PresX
-0.2227	-0.1339	-0.1098	-0.117	-0.3684	-0.0638	-0.0683	0.1148	-0.0061	-0.0094	-0.185	0.0863	PresA
-0.2107	-0.1551	-0.1311	-0.1443	-0.3824	-0.0925	-0.081	0.1129	-0.0158	-0.0313	-0.182	0.0654	PresI
-0.1178	-0.3662	-0.169	-0.2032	-0.0983	-0.1365	-0.2297	0.0428	0.148	-0.2641	-0.0803	0.1423	VisiX
-0.1132	-0.3553	-0.4001	-0.0132	-0.0348	-0.0384	-0.1957	0.1298	-0.0181	-0.0063	-0.083	0.1845	VisiA
-0.1073	-0.2662	-0.4507	-0.1771	-0.0626	-0.0734	-0.1147	-0.045	-0.0411	-0.3348	-0.089	0.1338	VisiI
-0.1832	-0.1284	-0.0228	-0.2829	-0.3103	-0.0106	0.0393	0.2548	-0.1227	-0.1042	-0.475	-0.2424	VieX
-0.2387	-0.0398	-0.0483	-0.0126	-0.2698	-0.0351	-0.1426	0.0977	-0.1613	-0.0576	-0.165	-0.2946	VieA
-0.0148	0.114	0.3724	0.3957	0.0443	0.1642	-0.1119	0.1659	-0.5544	0.1228	0.2912	0.4076	Raf
-0.142	0.0463	0.2575	0.2546	0.2008	0.0963	-0.3743	0.0308	0.0672	0.6489	-0.176	0.0003	Clo

-0.1593	-0.2214	-0.2507	-0.0327	-0.2371	-0.1859	0.0564	0.2316	0.4573	-0.0523	-0.234	0.1855	Lluvia
-0.1663	-0.1353	-0.1179	-0.1198	-0.1293	-0.0213	-0.6675	0.0438	-0.1116	-0.2025	0.0098	-0.0958	Tormenta
0.1021	-0.0299	0.1731	-0.3072	-0.0907	0.6342	-0.1957	-0.4377	0.0906	0.2267	-0.102	-0.0715	Niebla
-0.1814	-0.1916	-0.1645	-0.2031	-0.2736	-0.1557	-0.086	0.2389	0.1222	0.0469	0.1968	-0.2782	Lluvia-Tormenta
0.0908	0.0206	-0.0404	-0.3166	-0.04	0.6398	-0.0092	-0.6462	0.0691	0.0063	-0.15	-0.053	Niebla-Lluvia
-0.2295	-0.0221	-0.0592	-0.2554	-0.0743	-0.099	0.2615	-0.1753	-0.3432	0.0165	0.055	0.095	Nada

Corridas en *MultilayerPerceptron*- CD_1

Correlation coefficient	0.6933
Mean absolute error	20917.5939
Root mean squared error	24514.3214
Relative absolute error	95.2257 %
Root relative squared error	87.0249 %
Total Number of Instances	36

Corridas en *MultilayerPerceptron*- CD_3

Correlation coefficient	0.7403
Mean absolute error	16874.1215
Root mean squared error	20867.4658
Relative absolute error	76.8181 %
Root relative squared error	74.0787 %
Total Number of Instances	36

Corridas en *MultilayerPerceptron*- CD_4

Correlation coefficient	0.7101
Mean absolute error	17722.7253
Root mean squared error	23256.2763
Relative absolute error	80.6813 %
Root relative squared error	82.5589 %
Total Number of Instances	36

Corridas en *MultilayerPerceptron*- CD_5

Correlation coefficient	0.7593
Mean absolute error	16910.0631
Root mean squared error	19737.8972
Relative absolute error	76.9818 %
Root relative squared error	70.0688 %
Total Number of Instances	36

Corridas en *MultilayerPerceptron*- CD_6

Correlation coefficient	-0.035
Mean absolute error	22757.9275
Root mean squared error	29851.8475
Relative absolute error	103.6037 %
Root relative squared error	105.9729 %
Total Number of Instances	36

Corridas en *MultilayerPerceptron*- CD_7

Correlation coefficient	0.0367
Mean absolute error	22147.1914
Root mean squared error	30355.3822

Relative absolute error	100.8234 %
Root relative squared error	107.7604 %
Total Number of Instances	36

Corridas en *LinearRegression*- CD_1

CT =

28.1169 * TF_E +
17.4007 * TD_N +
-18.6057 * HDO_E_P +
8822.3264 * TempA +
-6568.042 * PresX +
9895.6024 * PresI +
-1881.8126 * VieX +
-3429700

Correlation coefficient	0.323
Mean absolute error	28446.5714
Root mean squared error	58341.2654
Relative absolute error	129.5008 %
Root relative squared error	207.1092 %
Total Number of Instances	36

Corridas en *LinearRegression*- CD_3

CT =

227.1149 * HDO_E_O +
59.4489 * HDO_N +
7198.956 * TempX +
-2062.6078 * HumA +
1493.1216 * HumI +

3484.2524 * PresX +
-17275.6781 * VisiA +
-2177.6427 * VieX +
-74635.2503 * Lluvia +
-3374011

Correlation coefficient	0.6256
Mean absolute error	19963.0743
Root mean squared error	24760.5596
Relative absolute error	90.8803 %
Root relative squared error	87.899 %
Total Number of Instances	36

Corridas en *LinearRegression*- CD_4

CT =

66.7461 * HDO_N +
6821.2005 * TempX +
-11220.4888 * RocX +
6174.8114 * RocI +
-42069.6058 * VisiA +
14727.1566 * VisiI +
-3037.3133 * VieX +
4600.5699 * VieA +
123736.617 * Lluvia-Tormenta +
316637.5

Correlation coefficient	0.7133
Mean absolute error	17635.2651
Root mean squared error	19894.3075
Relative absolute error	80.2832 %
Root relative squared error	70.624 %
Total Number of Instances	36

Corridas en *LinearRegression*- CD_5

CT =

4289.1262 * TempI +
-4632.2134 * HumX +
-24245.425 * VisiA +
-2508.9502 * VieX +
-91410.1692 * Lluvia +
143720.498 * Tormenta +
755671.7

Correlation coefficient	0.6457
Mean absolute error	15789.6445
Root mean squared error	21317.6872
Relative absolute error	71.8811 %
Root relative squared error	75.677 %
Total Number of Instances	36

Corridas en *LinearRegression*- CD_6

CT =

-61.3183 * V3 +
-56.6697 * V7 +
59.3042 * V9 +
219624.5

Correlation coefficient	0.6788
Mean absolute error	15640.196
Root mean squared error	20068.5881
Relative absolute error	71.2008 %
Root relative squared error	71.2427 %
Total Number of Instances	36

Corridas en *LinearRegression*- CD_7

CT =

-6891.3651 * V8 +
12946.9865 * V10 +
-2146.1924 * V11 +
-5587288

Correlation coefficient	0.49
Mean absolute error	19325.7867
Root mean squared error	24179.7882
Relative absolute error	87.9791 %
Root relative squared error	85.8373 %
Total Number of Instances	36