



Modelo de predicción de demanda de energía para ucp's en Colombia bajo el método stacking y minería de texto

Javier Hernán Velasco Castillo

Universidad Del Valle
Facultad De Ingeniería, Escuela De Ingeniería de Sistemas y Computación
Santiago De Cali, Colombia
2019

Modelo de predicción de demanda de energía para ucp's en Colombia bajo el método stacking y minería de texto

Javier Hernán Velasco Castillo

Trabajo de investigación presentado como requisito parcial para optar al título

de:

Magister en Ingeniería

Director:

Ph.D. Andrés Mauricio Castillo Robles

Línea de Investigación:

Gestión de la información y el conocimiento

Grupo de Investigación:

Grupo de Univalle en Inteligencia Artificial - GUIA

Universidad Del Valle

Facultad De Ingeniería, Escuela De Ingeniería de Sistemas y Computación

Santiago De Cali, Colombia

2019

A la memoria de mis abuelos Aura y Silvio, quienes fueron los pilares fundamentales para mi formación personal y profesional. Todo lo que soy se lo debo a ellos y lastimosamente este éxito no podemos celebrarlo juntos; como dijo mi abuelo “es el ciclo de la vida y debes continuar, estuve aquí ayudándote hasta donde más pude”.

A mi madre María Eugenia, a mi esposa Angela, a mi hija Mariana, quienes con su apoyo, comprensión y fortaleza, contribuyeron para continuar a pesar de los golpes de la vida.

Con mi eterna gratitud,

Javier Hernán Velasco Castillo

Agradecimientos

Agradezco y dedico este proyecto a todas aquellas personas que contribuyeron y me motivaron para finalizar esta meta propuesta hace algunos años.

A mi familia, que con su apoyo y comprensión me dieron fuerzas para continuar.

Al tutor de investigación, Ph.D. Andrés Mauricio Castillo Robles, porque con su ayuda, confianza, orientación y apoyo logramos grandes éxitos como la publicación del artículo y la finalización del proyecto.

Al profesor M.Sc. Oswaldo Solarte Pabón por su colaboración en el grupo de investigación GUIA.

A los evaluadores del anteproyecto y del proyecto Ph.D. Oscar Fernando Bedoya Leyva y Ph.D. Raúl Ernesto Gutiérrez de Piñerez, por el tiempo de dedicación a la revisión del proyecto de investigación y sus aportes.

A Ph.D. Víctor Andrés Bucheli Guerrero, por su colaboración para participar como ponente en el 13vo Congreso Colombiano de Computación.

A Nelcy Bastidas Marín y Lady Linibec Díaz Molano, por su colaboración y atención.

A la Escuela de Ingeniería de Sistemas y Computación de la universidad del Valle, por darme nuevamente la oportunidad de continuar la maestría.

A Dios, por darme salud y fortaleza para finalizar el proyecto

Resumen

Los modelos de predicción de demanda de energía eléctrica son usados por los comercializadores de energía para planificar la demanda hora a hora de la siguiente semana. Usando estas predicciones, la compañía de expertos en mercados (XM) planifica la asignación de la generación de energía eléctrica del siguiente día a los distintos generadores conectados al Sistema Interconectado Nacional (SIN). Un buen modelo de pronóstico es una garantía para controlar los costos operacionales.

Como objetivo de esta investigación se planteó el desarrollo de un modelo que mejorará las predicciones disminuyendo el indicador de error con que son evaluadas las UCP's en Colombia, como resultado de esta investigación se definió un modelo que a partir de métodos estadísticos se generarán 12 predicciones que serán fuente de información para la red neuronal como meta-clasificador, las predicciones resultantes se analizarán y se determinará si presentan valores por encima de las predicciones de la UCP. Con la información extraída de la red social Twitter se analizará las posibles causas que determinaron que los valores de predicción se desfasaran de los reales.

Como resultados de la ejecución del modelo con los datos públicos del administrador de la UCP UCode, se realizó la comparación de los datos de predicción del modelo contra las predicciones hechas por la UCP y se demostró que el modelo logra disminuir el error. La falta de la información de la demanda no atendida impide actualizar los datos de la demanda real y esto genera que se presenten algunas inconsistencias en algunas horas del día. También se evidenció que con el acceso a la información más actualizada (SCADA, balances y medidores frontera), el sistema está en la capacidad de mejorar sus predicciones.

Palabras clave: Modelo de predicción, modelo estadístico, demanda de energía eléctrica, pronóstico, stacking, redes neuronales artificiales.

Abstract

Prediction models of Energy demand are used by energy marketers to plan the demand hour by hour of the following week. Using these predictions, the company of experts in markets (XM) plans the allocation of the next day's power generation to the different generators connected to the National Interconnected System (NIS). A good forecasting model is a guarantee to control operational costs.

The objective of this research was to develop a model that will improve predictions by decreasing the error indicator used to evaluate PCUs in Colombia. As a result of this research, a model was defined that will generate 12 predictions from statistical methods. they will be a source of information for the neural network as a meta-classifier, the resulting predictions will be analyzed and it will be determined if they present values above the predictions of the PCU. With the information extracted from the social network Twitter will analyze the possible causes that determined that the prediction values are offset from the real ones.

As a result of the execution of the model with the public data of the PCU UCodensa administrator, the comparison of the prediction data of the model against the predictions made by the PCU was made and it was demonstrated that the model manages to reduce the error. The lack of information from the unmet demand prevents updating the actual demand data and this generates that some inconsistencies appear at certain times of the day. It was also evidenced that with the access to the most up-to-date information (SCADA, balance sheets and border meters), the system is in the capacity to improve its predictions.

Keywords: Prediction model, statistics model, electry energy demand, forecast, Stacking, artificial neural network

Contenido

	Pág.
Resumen	V
Lista de figuras	IX
Lista de tablas	XI
Lista de abreviaturas	XII
Introducción	1
1. Problema de Investigación.....	3
1.1 Descripción del problema.....	3
1.2 Justificación	7
1.3 Objetivos.....	9
1.3.1 Objetivo general	9
1.3.2 Objetivos específicos.....	9
2. Marco teórico	10
2.1 Demanda de Energía	10
2.2 Métodos Estadísticos	13
2.3 Modelos de predicción.....	14
2.4 Modelo Arima.....	15
2.5 Modelo Dinámico Bayesiano.....	16
2.6 Modelo de regresión.....	18
2.7 Clasificadores individuales	19
2.8 Modelo de redes neuronales.....	19
2.9 Clasificación de las redes neuronales	21
2.10 Codificación de datos de entrada.....	22
2.11 Algoritmo Backpropagation	23
2.12 Desarrollo de un modelo de predicción	25
2.13 Modelos de combinación de clasificadores	26
2.13.1 Bagging.....	26
2.13.2 Boosting	26
2.13.3 Stacking	26
2.14 Validación	27
2.14.1 Validación cruzada	28
2.14.1.1 Método hold-out.....	28
2.14.1.2 Método k-fold	28
2.15 Errores en los pronósticos.....	29

2.16	Minería de texto	30
3.	Estado del Arte	32
3.1	Modelos públicos actuales	32
3.1.1	Modelo de predicción de Codensa.....	32
3.1.2	Modelo de predicción de Electricaribe	34
4.	Metodología.....	37
4.1	Selección de UCP	37
4.2	Revisión del modelo	39
4.3	Extracción de información	39
4.4	Selección de datos	41
4.5	Identificación de algoritmos y métodos.....	43
4.6	Modelo base	44
4.7	Stacking	45
4.7.1	Experimentación.....	46
4.7.2	Entrenamiento (Train).....	47
4.7.3	Prueba (Test).....	48
4.8	Minería de texto	49
5.	Resultados	52
5.1	Extracción de información de redes sociales	56
5.2	Análisis de datos	59
5.3	Ejecución del modelo en otra UCP	62
5.4	Prototipo.....	69
5.5	Discusión	72
6.	Conclusiones y trabajos futuros	74
6.1	Conclusiones	74
6.2	Trabajos futuros.....	76
	Bibliografía	77

Lista de figuras

	Pág.
Figura 1-1: Indicador de pronósticos resumen mensual agosto 2018.....	4
Figura 1-2: Temperatura para la ciudad de Cali el 1 de diciembre de 2017	6
Figura 1-3: Temperatura para la ciudad de Bogotá el 1 de diciembre de 2017	6
Figura 1-4: Curvas de Demanda y Temperatura en Cali en agosto de 2018.....	7
Figura 2-1: Cadena de servicios.....	10
Figura 2-2: Demanda de Energía en el SIN	11
Figura 2-3: Demanda de Energía de Colombia para el mes de agosto de 2018.....	12
Figura 2-4: Modelo general de medidores de una administradora de UCP.	13
Figura 2-5: Curva de demanda de energía de lunes a domingo en Ucodensa del 18 al 24 de septiembre de 2017	15
Figura 2-6: Curva de demanda de energía de lunes a domingo en UCali del 17 al 24 de septiembre de 2017	16
Figura 2-7: viernes 6 a.m. UCP Ucodensa entre mayo y septiembre de 2017.....	17
Figura 2-8: Nota del periódico EL TIEMPO sobre vacaciones en junio de 2017	17
Figura 2-9: Curva de regresión simple	18
Figura 2-10: Esquema de una red neuronal artificial	20
Figura 2-11: Red neuronal artificial con conexiones y nodos	23
Figura 3-1: Metodología aplicada por Codensa	33
Figura 3-2: Métodos usados por Codensa	33
Figura 3-3: Módulos usados por Electricaribe	35
Figura 3-4: Metodología aplicada por Electricaribe	36
Figura 4-1: Las 10 UCP's con mayor participación en el SIN en agosto de 2018	37
Figura 4-2: Participación de Codensa y Antioquia en la demanda de energía del SIN para el mes de agosto de 2018	38
Figura 4-3: Indicadores de pronósticos oficiales de demanda de Ucodensa.....	40
Figura 4-4: Análisis de componentes principales de Codensa.....	42

Figura 4-5:	Modelo Base	44
Figura 4-6:	Modelo de Validación cruzada.....	46
Figura 4-7:	Determinación de nodos por capa.....	47
Figura 4-8:	Red neuronal artificial con backpropagation	48
Figura 4-9:	Valores de diferencia por tipo de día	49
Figura 4-10:	Modelo de ETL y análisis de datos para minería de texto	50
Figura 4-11:	Modelo de predicción de demanda de energía - SILVCAS	51
Figura 5-1:	Demanda de energía	53
Figura 5-2:	Diferencia por tipo de día	54
Figura 5-3:	Demanda y diferencia día lunes	54
Figura 5-4:	Demanda y diferencia día miércoles	55
Figura 5-5:	Demanda y diferencia día domingo	55
Figura 5-6:	Demanda y diferencia día lunes festivo.....	56
Figura 5-7:	Datos del clima.....	57
Figura 5-8:	Demanda de energía domingo 11 de marzo de 2018.....	59
Figura 5-9:	Demanda y diferencia día lunes	59
Figura 5-10:	Publicación del IDEAM - Elecciones al congreso 11 de marzo 2018	60
Figura 5-11:	Publicación de SectorMovilidad - elecciones del 11 de marzo de 2018..	61
Figura 5-12:	Demanda día domingo de los meses de febrero y marzo	61
Figura 5-13:	Publicación de SectorMovilidad - elecciones del 27 de mayo de 2018 ...	62
Figura 5-14:	Demanda de energía	65
Figura 5-15:	Diferencia por tipo de día	65
Figura 5-16:	Demanda de energía día viernes	66
Figura 5-17:	Diferencia entre SILVCAS y la UCP para el día viernes.....	66
Figura 5-18:	Demanda y diferencia para el día DOALF de Uantioquia	67
Figura 5-19:	Demanda del tipo de día DOALF en Uantioquia	68
Figura 5-20:	Demanda de energía de la UCP Uantioquia para los días DOMMINGO y DOALF en mayo 2018	68
Figura 5-21:	Interfaz de inicio de la aplicación web.	69
Figura 5-22:	Interfaz de predicción de la primera semana de octubre de 2018.....	70
Figura 5-23:	Precio de bolsa y escasez	71
Figura 5-24:	Precio de bolsa por hora de la primera semana de octubre de 2018	71

Lista de tablas

	Pág.
Tabla 2-1: Clasificación de las redes neuronales (Álvarez)	21
Tabla 4-1: Estructura de la tabla datosucp	41
Tabla 4-2: Tipos de días definidos por XM	41
Tabla 4-3: Cruce de métodos con los tipos de muestreo	43
Tabla 4-4: Valor equivalente por tipo de día.....	45
Tabla 4-5: Registro por rango de minutos de temperatura y clima	51
Tabla 5-1: Cantidad de muestras por tipo de día	52
Tabla 5-2: Predicción del modelo SILVCAS vs UCP Codensa.....	53
Tabla 5-3: Datos del clima (Datos abiertos del IDEAM)	58
Tabla 5-4: Datos del clima (Weather.com)	58
Tabla 5-5: Valores de predicciones de SILVCAS y Uantioquia	64
Tabla 5-6: Valor acumulado por diferencias en tipos de día	64
Tabla 5-7: Valor acumulado de diferencias para el día viernes de Uantioquia	67
Tabla 5-8: Valor acumulado de diferencias para el día doalf de Uantioquia	67
Tabla 5-9: Diferencia en MWh en la primera semana de octubre de 2018	70
Tabla 5-10: Calculo del valor en pesos de la diferencia	72
Tabla 5-11: Predicciones con 4 y 1 semana de datos históricos.	72

Lista de abreviaturas

Abreviatura	Término
-------------	---------

<i>CND</i>	Centro nacional de despacho
<i>CNO</i>	Consejo nacional de operación
<i>CREG</i>	Comisión de regulación de energía y gas
<i>DOALF</i>	Domingo antes de lunes festivo
<i>ENS</i>	Energía no suministrada
<i>GWd</i>	Gigavatio-día
<i>GWh</i>	Gigavatio-hora
<i>LAC</i>	Liquidador y administrador de cuentas
<i>LF</i>	Lunes festivo
<i>MADLF</i>	Martes después de lunes festivo
<i>MAPE</i>	Error porcentual absoluto medio
<i>MLP</i>	Perceptrón multicapa
<i>MWh</i>	Megavatio-hora
<i>OR</i>	Operadores de red
<i>PCA</i>	Análisis de componente principal
<i>PENS</i>	Porcentaje de energía no suministrada
<i>RNA</i>	Redes neuronales artificiales
<i>SAALF</i>	Sábado antes de lunes festivo
<i>SCADA</i>	Sistema de supervisión control y adquisición de datos
<i>SDE</i>	Desviación típica de los errores
<i>SDL</i>	Sistema de distribución local
<i>SIN</i>	Sistema interconectado nacional
<i>STR</i>	Sistema de transmisión regional
<i>STN</i>	Sistema de transmisión nacional
<i>TF-IDF</i>	Frecuencia de Termino-frecuencia inversa de documento
<i>UCP</i>	Unidad de control de pronóstico

Introducción

La planeación a corto plazo de la demanda de energía eléctrica en Colombia está bajo la administración de la Compañía de Expertos en Mercados (XM) que es una filial de la firma estatal colombiana de transmisión ISA especializada en gestión de sistemas de tiempo real. XM tiene a cargo la planeación a corto, mediano y largo plazo de los recursos de generación de Colombia, es decir, plantas hidroeléctricas, térmicas y eólicas y los recursos de transmisión de acuerdo con la demanda de energía eléctrica de cerca de 42 millones de habitantes. Esta planeación de corto plazo, comprende la recepción de las ofertas diarias que presentan los generadores en la Bolsa de Energía, donde se asignan hora a hora las plantas que suministrarán la energía al día siguiente. XM, realiza esta selección con criterios de seguridad y economía para garantizarles a los usuarios el servicio con estándares de calidad, confiabilidad y eficiencia¹.

Para garantizar estos estándares, el Consejo Nacional de Operación mediante el acuerdo 349 de 2006 aprueba los procedimientos de creación y modificación de Unidades de Control de Pronóstico las cuales son un conjunto de elementos que conforman una unidad típica de un sistema eléctrico donde se realiza un balance de energía (Wood, 2013). Dentro de sus responsabilidades la UCP debe reportar a más tardar los viernes a las 13:00 horas el pronóstico de la semana siguiente al área de planeación de corto plazo de XM, con el fin de determinar la demanda hora a hora de la semana siguiente y proceder a definir la cantidad de energía a ofertar con los generadores conectados al sistema eléctrico, programar las unidades de generación, fijar el precio spot de oferta de la electricidad en el mercado y evitar el sobre costo en la producción de energía (Rueda, 2001).

El pronóstico de demanda de energía realizado por las UCP no tiene penalización, sin embargo, en el capítulo 3 de la Resolución de la Comisión de Regulación de Energía y

¹ <http://www.xm.com.co/corporativo/Paginas/Nuestra-empresa/que-hacemos.aspx>

Gas (CREG) 094 de 2012 para realizar la estimación de la Energía No Suministrada (ENS) y del Porcentaje de Energía No Suministrada (PENS) se utiliza como base el pronóstico oficial de demanda de energía en la hora del evento ocurrido en el STR. En los casos que el PENS sea mayor al 2%, el Centro Nacional de Despacho (CND) deberá enviar a la Superintendencia de Servicios Públicos Domiciliarios (SSPD) un informe detallado del evento ocurrido, considerando que hubo un ENS y se deberá estimar cual es la ENS a compensar para que la superintendencia de servicios públicos domiciliarios determine si hay lugar a aplicar dicha compensación. A esta compensación será asignada la mayor ENS estimada dentro el periodo del evento y será el Centro Nacional de despacho (CND) el encargado de suministrar al Liquidador y Administrador de Cuentas (LAC) la información requerida para el cálculo de las compensaciones correspondientes al mes a facturar. Debido a esto, cada UCP ha implementado su propio modelo de predicción de demanda basado en su conocimiento del negocio y en la experticia de sus analistas.

La mayoría usa métodos estadísticos (Mínimos Cuadrados, Mediana, Promedio (Codensa, 2016) y Medias Móviles) combinados con conocimientos de entorno como, por ejemplo, considerar los períodos de vacaciones, las épocas de sequía o inviernos fuertes, etc. Sin embargo, considerar y modelar todos los factores externos es una tarea compleja ya que no existen sistemas confiables en Colombia que los registre. A pesar del esfuerzo puesto por las UCP en estos modelos hasta el momento no ha sido posible incluir, de forma sistemática, dichos factores en los modelos de predicción. En este documento describimos un modelo de predicción de demanda basado en métodos estadísticos y de regresión, esto permiten aumentar la precisión de todos los modelos de predicción probados con datos históricos.

1. Problema de Investigación

1.1 Descripción del problema

En Colombia, existen plantas de generación hidráulica, térmica, solar y eólica que se encuentran conectadas a los Sistema de Distribución Local (SDL) y Sistema de Transmisión Regional (STR). Cada planta es programada para generar los **megavatio-hora (MWh)** en la hora pactada para poder atender la demanda de energía del Sistema Interconectado Nacional (SIN). Esta demanda debe ser con base en la generación neta de las plantas e incluye: hidráulicas, térmicas, plantas menores, cogeneradores, demanda no atendida, limitación del suministro e importaciones (XM, 2017). (ver Ecuación (1.1)).

$$\text{La Ecuación de la Demanda Energía SIN: } \mathbf{Generación + Demanda No} \quad (1.1) \\ \mathbf{Atendida + Importaciones - Exportaciones.}$$

Esta energía generada por las diferentes plantas no puede ser almacenada a bajo costo como otros productos, sino que debe ser consumida de forma instantánea y el sistema eléctrico colombiano debe estar en la capacidad de atender toda la demanda (Franco, 2012). Por este motivo los modelos de predicción de cada administradora de Unidad de Control de Pronóstico (UCP) debe representar en la forma más fiel posible la demanda real para evitar que se produzcan desviaciones grandes en el pronóstico que puedan implicar grandes costos de operación (Rueda, 2001). Estos errores de predicciones son publicados en el portal de XM mensualmente en el archivo “Indicadores Calidad Pronóstico Oficial”, en donde utilizan como indicador de desempeño el error absoluto medio porcentual (MAPE) (ver **Figura 1-1**), y se determina la clasificación del error, el porcentaje de participación en el SIN, el periodo y el día de mayor desviación para cada una de las 36 UCP's.

La clasificación del error está dividida en cuatro categorías:

1. Errores de desviación que se encuentran menor a 3%.

2. Errores entre el 3% y el 5%.
3. Errores mayores al 5%.
4. Cargas industriales.

En la parte final del archivo de Indicadores de Calidad de Pronóstico Oficial se encuentran las observaciones y recomendaciones para los mercados de comercialización o UCP's que durante tres meses consecutivos presentaron desviaciones mayores a 5% y por lo tanto serán tema en el próximo Consejo Nacional de Operación.

Figura 1-1: Indicador de pronósticos resumen mensual agosto 2018.



Nombre de la fuente: <http://www.xm.com.co/Paginas/Home.aspx>

Aunque el pronóstico de demanda de energía realizado por las UCP's no tiene penalización en el momento, la comunidad se ha enfocado en disminuir el porcentaje de desviación del

pronóstico de la demanda, por lo que han propuesto diferentes modelos y han combinado distintos métodos, pero todos convergen en que hay factores externos complejos que afectan sus predicciones, y cada UCP ha tratado de incluir este conocimiento en sus modelos, algunas veces por medio de apreciaciones humanas. Hasta el momento no ha sido posible incluir, de forma sistemática, dichos factores en los modelos de predicción.

En el momento todas las UCP's tienen modelos y técnicas diferentes para predecir la demanda de energía a corto plazo. Como ejemplo tenemos algunos operadores de red (OR) que han publicado sus modelos, como Codensa y Electricaribe (Franco, 2012). Dentro de los modelos que más se utilizan se encuentran los modelos estadísticos (su base es la información histórica), modelos empíricos (dependen de la intuición y el juicio humano) y la inteligencia artificial (Franco, 2012).

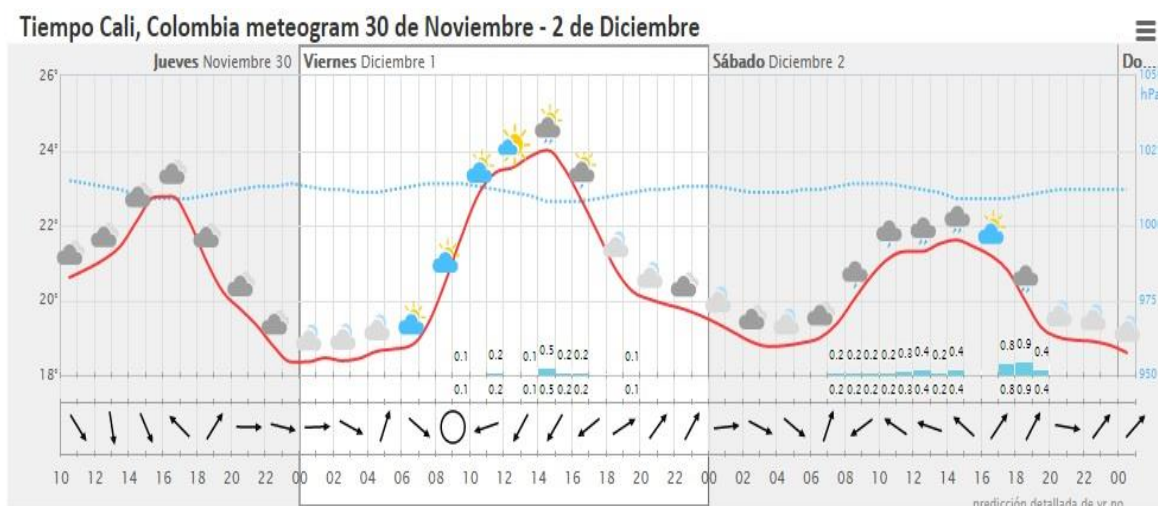
Los métodos de ensamble y los meta-algoritmos han mostrado mejorar la capacidad de predicción que tienen los modelos de predicción por separado. La idea detrás de estos métodos es que siempre es mejor tener la opinión de un panel de expertos que la de uno solo cuando se deben tomar decisiones importantes. En el caso del modelo de predicción usado por Codensa, se observa este tipo de algoritmos (Valencia Cárdenas, 2013).

Estos modelos estadísticos son una gran herramienta de predicción, pero es de destacar que la demanda de energía es un proceso estocástico no estacionario integrado por varios componentes individuales que pueden ser influenciados por varios factores externos relacionados con el consumo de energía en una hora determinada como el clima, temperatura, economía, demografía, eventos deportivos importantes, éxodo de personas, etc. Se ha determinado que existen patrones que se repiten en el futuro (Sarmiento, 2008). Por ello se requiere de un método de extracción de esos patrones para validar los datos históricos y que colaboran a generar predicciones más acertadas (Ortiz, 2014).

Por ejemplo, estudios realizados por Mesa Bustelo, han demostrado que la variación en la temperatura en Madrid, España, afecta la demanda de energía ya sea positiva o negativamente. La variación se calcula en 12%, concluyendo que a menor temperatura menos consumo de energía y a mayor temperatura mayor consumo, debido al uso aires acondicionados (Fernández de Mesa Bustelo, 2016). De ahí que se tiene como hipótesis

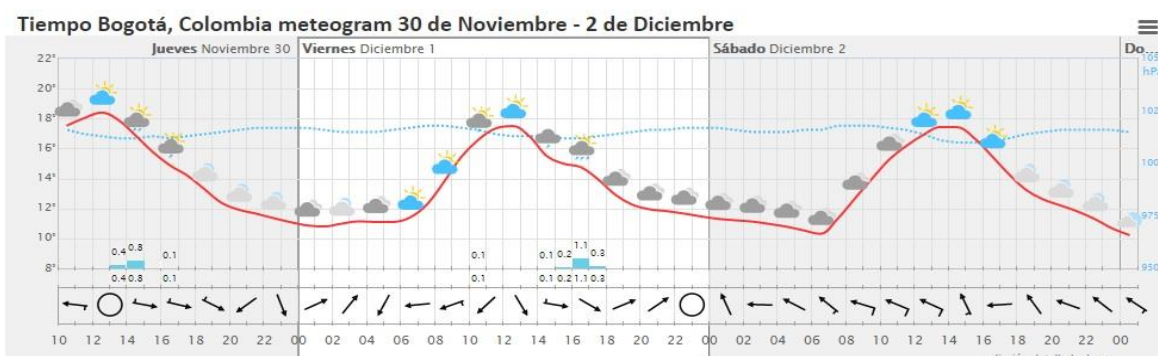
que la temperatura es uno de los factores externos que más afectan la demanda de energía en Colombia, como se observa en la **Figura 1-2** y **Figura 1-3** las temperaturas presentan un comportamiento de una onda sinusoidal donde el aumento de temperatura inicia a partir de las 10 horas y disminuyendo a partir de las 16 horas, de acuerdo a la ubicación de la región estas temperaturas tienen mayor o menor intensidad.

Figura 1-2: Temperatura para la ciudad de Cali el 1 de diciembre de 2017



Nombre de la fuente: <https://meteobox.co/cali/>

Figura 1-3: Temperatura para la ciudad de Bogotá el 1 de diciembre de 2017

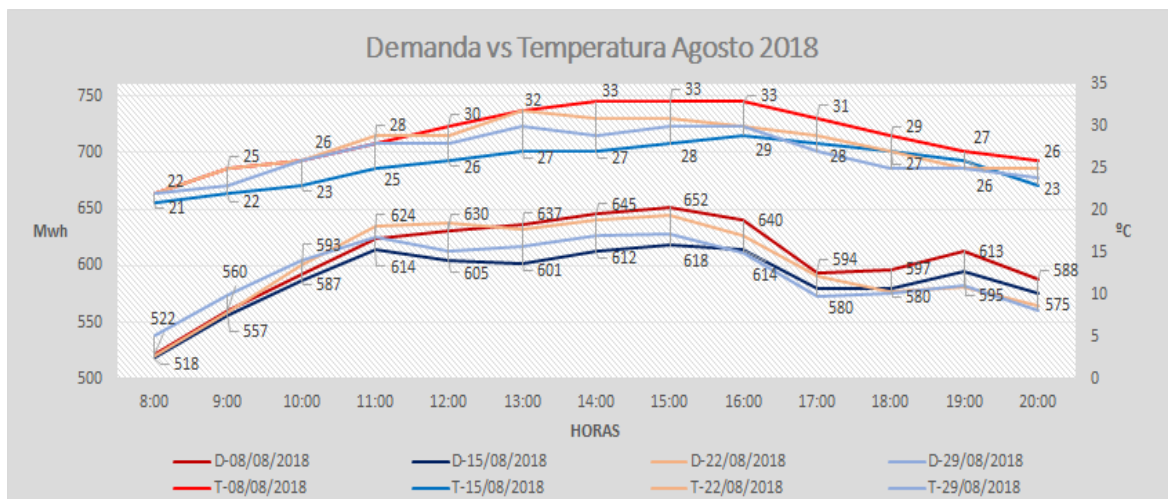


Nombre de la fuente: <https://meteobox.co/bogota/>

Realizando la comparación de los días miércoles de agosto de 2018 entre la demanda de energía de la administradora Ucali (Escala de valores lado izquierdo) y la Temperatura en Cali (Escala de valores lado derecho), se observa que existe una correlación entre

consumo y temperatura, por lo que se plantea la siguiente hipótesis: la temperatura es directamente proporcional al consumo. En la **Figura 1-4** la curva de color rojo representa los valores más altos de temperatura y consumo de energía y corresponde al día 08 de agosto de 2018 y la curva en color azul representa los valores más bajo de temperatura y consumo de energía que corresponde al día 15 de agosto de 2018.

Figura 1-4: Curvas de Demanda y Temperatura en Cali en agosto de 2018



La ucp UCodensa en el momento presenta la mayor participación en el Sistema Interconectado Nacional y se mantiene entre los valores de menor desviación del MAPE (por debajo del 2%) y con resultados estables, es de considerar que Codensa ha socializado el modelo de predicción de demanda de energía que utiliza. Por lo tanto, surge la siguiente pregunta de investigación:

¿Diseñar un modelo genérico de predicción de demanda energía que mejore las predicciones de Codensa?

1.2 Justificación

Existen varios factores que influyen en el pronóstico de la demanda de energía a corto plazo, que cuando no son tenido en cuenta pueden producir desvíos considerables en los pronósticos generando sobrecostos en la producción de energía. Este trabajo busca mejorar la confiabilidad de los pronósticos de la demanda de energía de cada mercado de

comercialización que debe reportar a XM a través de un modelo de predicción que se adapte a cada UCP.

La investigación busca desarrollar e implementar un modelo capaz de predecir la demanda de energía eléctrica horaria a corto plazo de las UCP's de Colombia, con la cual se podría ayudar a disminuir la desviación y mantener a la mayoría de UCP's de Colombia con errores medios porcentuales por debajo de los actuales disminuyendo los costos de operación.

Las UCP's con mayor participación en el Sistema Interconectado Nacional (SIN) de Colombia han enfocado parte de sus recursos en la búsqueda de nuevos modelos que mejoren la predicción de la demanda real por hora, la cual es usada por XM para planificar la compra de energía a los generadores. Es de conocimiento entre los administradores de las UCP's que existen factores externos que afectan sus predicciones, y cada operador ha tratado de incluir este conocimiento en sus modelos. Por ejemplo, es evidente que la temperatura durante el día afecta el consumo de energía en ciudades calientes como Cali o Cartagena debido al uso de ventiladores y aires acondicionados. Por tanto, la expectativa de tener una semana calurosa o fría hace que las estimaciones, que estos operadores realizan sobre el consumo, suban o bajen para las horas de la tarde. Debido a que estas apreciaciones son realizadas por cada operador, no existe un modelo general que pueda ser implementado por cualquier UCP de Colombia y que de acuerdo a sus propios factores externos se pueda alimentar el modelo.

El problema principal para incluir este tipo de información en los modelos actuales, es el deficiente sistema de información climático colombiano. No existe ningún sistema que provea esta información. El IDEAM realiza reportes poco detallados sobre el clima, y la información de las estaciones de monitoreo, no se encuentran disponibles al público. Sin embargo, hemos notado que el IDEAM es muy activo reportando y recibiendo reportes del clima en su cuenta de Twitter, además en la web existen proveedores de servicios de previsión meteorológica que actualizan constantemente los datos como temperatura, precipitación y clima.

1.3 Objetivos

1.3.1 Objetivo general

Desarrollar un modelo general de predicción de demanda de energía a corto plazo para las distintas UCP's de Colombia que provea errores medios porcentuales por debajo de los actuales.

1.3.2 Objetivos específicos

- Determinar los modelos base de predicción que harán parte del modelo de ensamble.
- Identificar, extraer, convertir y guardar las fuentes de datos a usar por el modelo de predicción.
- Identificar, descargar y procesar las fuentes de información textual relacionada con el consumo de energía.
- Validar la predicción con datos reales.
- Implementar los resultados en un prototipo que se encuentre disponible en la web.
- Incorporar los descriptores de la información textual en los modelos de predicción implementados en el objetivo 1.

2.Marco teórico

En el momento se han realizado investigaciones enfocadas en encontrar un modelo, método o técnica que se ajuste a los valores de las demandas de energía en una UCP, es así que XM proporciona un pronóstico base de aproximación de la demanda de energía por cada UCP de la semana siguiente, que puede ser utilizado por la UCP y modificado de acuerdo a los valores de demanda de energía no atendida que se tiene programada.

2.1 Demanda de Energía

El sector de energía eléctrica en Colombia, está constituido por cuatro tipos de agentes, clasificados en generadores, transmisores, distribuidores o comercializadores de acuerdo a sus funciones que se enmarcan en la comisión reguladora de energía y gas CREG (Bolaño Cuesta, 2015). En la **Figura 2-1** se observa la cadena de servicios de los agentes y que en algunos casos las empresas de energía pueden tener 1 y hasta los 4 tipos de agentes de la cadena de servicios.

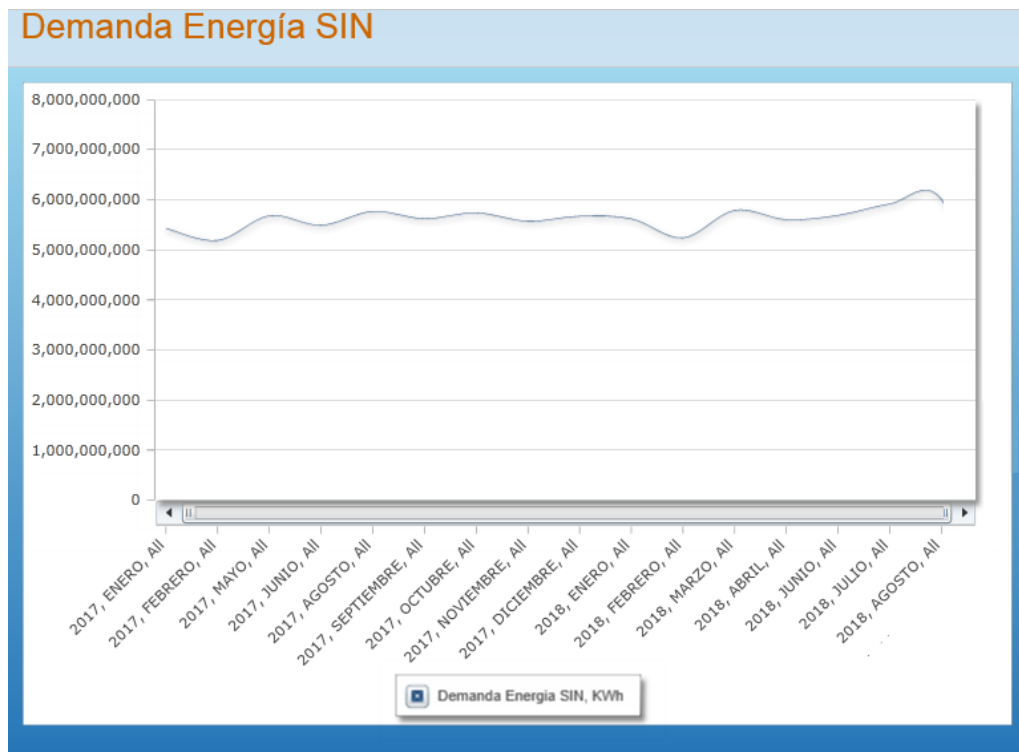
Figura 2-1: Cadena de servicios



Nombre de la fuente: <http://elmeridiano.co/la-energia-electrica-nos-llega-de-muy-lejos/37041>

La demanda de energía de Colombia o del Sistema Interconectado Nacional (SIN) es calculada por la suma de los consumos reportados por cada UCP que tiene participación en el SIN, estos consumos o demandas de energía pueden ser variables es así que para el año 2017 se presentó un crecimiento a partir del segundo trimestre del año por la recuperación de los hábitos de consumo en sectores como el residencial y pequeños negocios. La demanda de energía durante 2017 tuvo un crecimiento del 1.3% con respecto al año 2016, con un valor de 66,893 GWh (XM, 2017). Como se observa en la **Figura 2-2** la demanda de energía presenta los valores más bajos de demanda de energía en el mes de febrero, por ser uno de los meses que tiene menos días, además realizando una comparación entre el año 2017 con lo transcurrido hasta el año 2018(enero a agosto) se observa que la demanda presenta una tendencia a crecer.

Figura 2-2: Demanda de Energía en el SIN



Nombre de la fuente: <http://elmeridiano.co/la-energia-electrica-nos-llega-de-muy-lejos/37041>

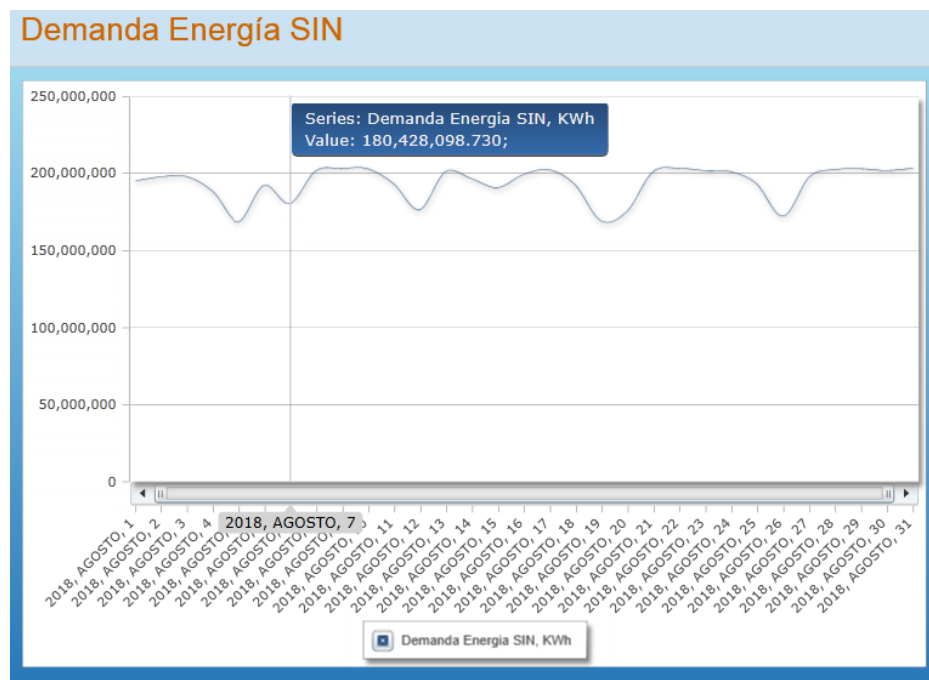
La demanda de energía se clasifica en tres tipos de días: 1) los días comerciales(lunes a viernes), 2) los sábados y 3) domingos y festivos; para el año 2017 los días comerciales

presentaron un crecimiento de 1.1% con un consumo promedio de 190 GWd; los días sábados presentaron un crecimiento de un 1.7% con un consumo promedio de 180.6 GWd, y el consumo en los días domingos y festivos el crecimiento fue de 1.6% con un consumo promedio de 162 GWd (XM, 2017).

De acuerdo al tipo de mercado la demanda de energía presentó un crecimiento del 1.9% para el mercado regulado (consumo residencial y pequeños negocios) y permaneció igual para el mercado no regulado (industria y comercio) respecto al 2016 (XM, 2017).

Los días domingo son fácilmente reconocibles debido a que es el día con menor demanda de energía en el SIN, pero como se puede observar en la **Figura 2-3**, el día martes 7 de agosto de 2018 la demanda fue baja por ser un día festivo, pero no llegó hasta los valores de un domingo. Es de destacar que un día martes festivo puede afectar considerablemente el Lunes, tipificando el día lunes como un día atípico debido a su comportamiento.

Figura 2-3: Demanda de Energía de Colombia para el mes de agosto de 2018.

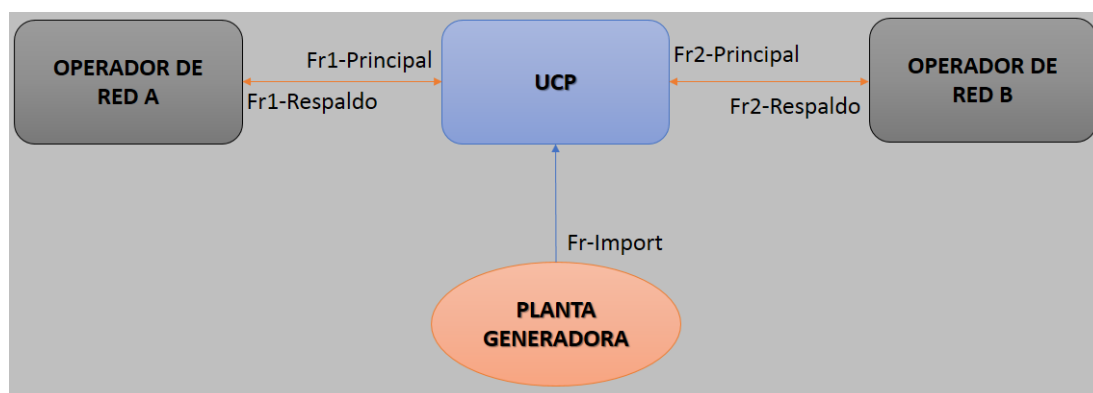


Nombre de la fuente:

[http://informacioninteligente10.xm.com.co/demanda/Paginas/Demanda%20Energia%20SIN%20\(kWh\).aspx](http://informacioninteligente10.xm.com.co/demanda/Paginas/Demanda%20Energia%20SIN%20(kWh).aspx)

Para realizar el cálculo de la demanda de energía por hora, se hace por medio de los medidores frontera, estos medidores son instalados uno principal y uno de respaldo en cada una de las fronteras con cuenta la administradora UCP. De acuerdo a los datos obtenidos en los medidores se pueden obtener que son exportaciones o importaciones de energía, y para el caso de las plantas generadoras solo se presenta importaciones ya que son los recursos que inyectan energía a las redes. En la **Figura 2-4** se observa un esquema general de una ecuación de demanda de energía.

Figura 2-4: Modelo general de medidores de una administradora de UCP.



Nombre de la fuente: Elaborado por Javier H. Velasco

2.2 Métodos Estadísticos

Los métodos estadísticos son utilizados en investigación y nos permiten ver con claridad y discernir una tendencia lo suficientemente confiable como para ser útil (Wood, 2013). Los métodos pueden enmarcarse en estadística descriptiva y en inferencia estadística. La estadística descriptiva es describir y resaltar numéricamente lo que es esencial en los resultados de un estudio usando los métodos apropiados y la inferencia estadística permite obtener conclusiones de los datos que se están investigando, de una manera objetiva (Madero, 2006).

Los métodos estadísticos se clasifican en (Fernández de Mesa Bustelo, 2016):

- **Recopilación:** Los datos pueden ser obtenidos interna o externamente, los datos externos son publicaciones.

- **Organización:** Se inicia por ordenar y corregir los datos recopilados, se continúa con la clasificación y se procede a tabular.
- **Presentación:** Los datos se muestran mediante texto, tablas o gráficas estadísticas.
- **Análisis:** Se utilizará el análisis estadístico simple que proporciona el fundamento básico para el análisis.
- **Interpretación:** Conclusión después de los resultados de la Interpretación.

Los métodos estadísticos son usados en la verificación de resultados de una tarea de minería de datos. Por ejemplo, luego de hacer un modelo de clasificación o predicción, el modelo debe ser verificado por medio de un test de hipótesis, usando las técnicas clásicas de estadística (García Ubilla).

2.3 Modelos de predicción

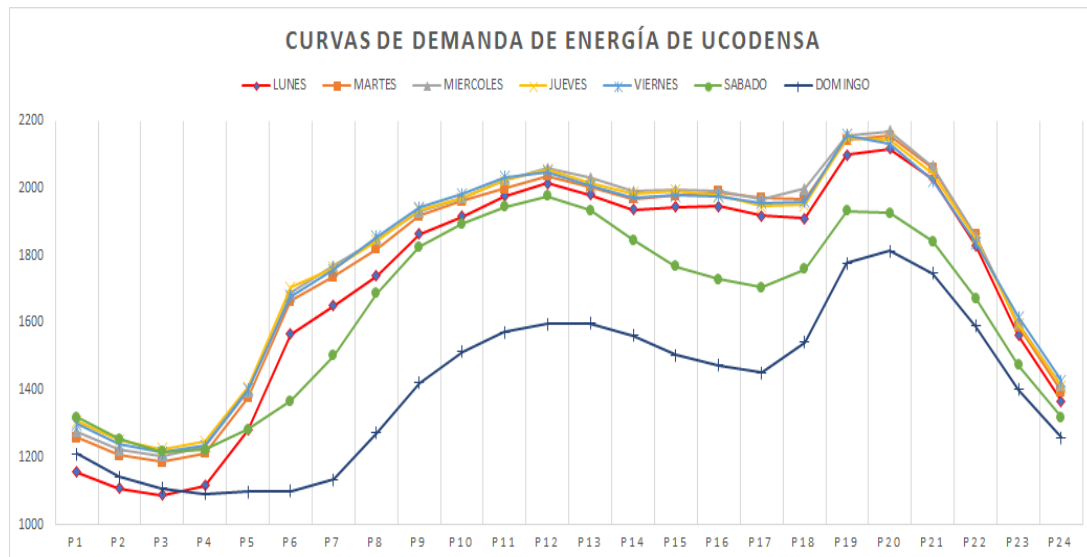
Los modelos de predicción son herramientas matemáticas (Cárdenas, s.f.) o algoritmos para estimar la probabilidad que un determinado evento ocurra (diagnóstico) o de predecir que ocurra (pronóstico) (Reyes S., 2014). Los modelos de predicción pueden ser modelos estadísticos que explican o predicen los valores de los datos. Estos datos son comparados para ver qué tan bien se ajustan los datos a los modelos, examinando los residuos o desviaciones. También es posible usar modelos para simular datos, con el fin de explorar propiedades de procedimientos o conceptos (Estrella). Los modelos son empleados en distintas áreas de conocimiento y su utilización resulta transversal en todos los campos de estudio, siendo una herramienta vital para establecer el escenario inicial de distintos sectores, permitiendo la definición de políticas de acuerdo a la información disponible (Prieto, 2015).

En los modelos de predicción se pueden tener desenlaces binarios, ordinales, continuos y de supervivencia (Reyes S., 2014). En el caso específico de diagnóstico, generalmente se está refiriendo a desenlaces binarios, por lo que el modelo de regresión logístico es el más utilizado como técnica estadística, ya que es flexible y puede combinar predictores categóricos, continuos, transformaciones no lineales y términos de interacción (Reyes S., 2014).

2.4 Modelo Arima

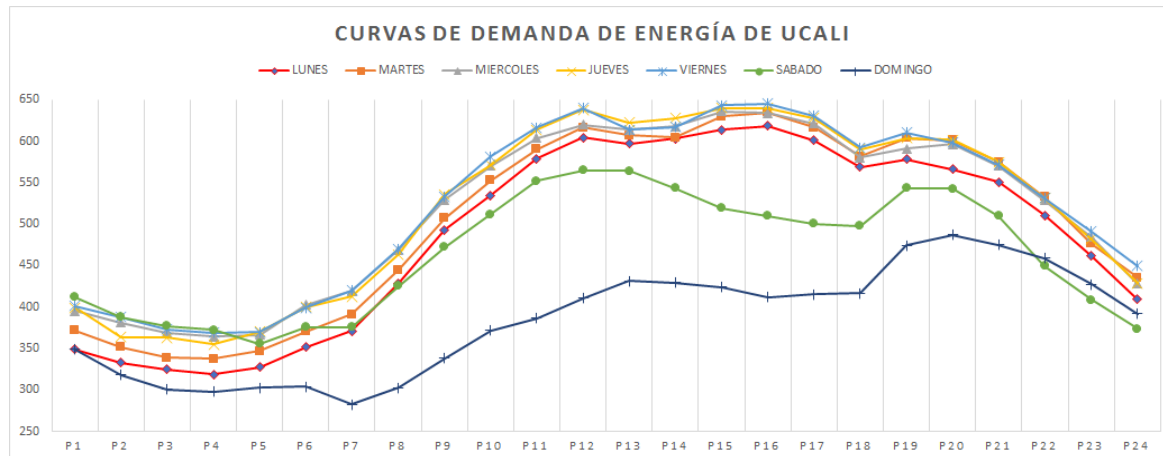
El análisis exploratorio mostró que el consumo de energía se incrementa a través de los años, los mayores consumos son los días martes, miércoles, jueves y viernes, seguidos del sábado y lunes, siendo este último el variable. Para el día domingo el consumo de energía es menor, la demanda en horas de la madrugada es baja, luego se presenta un incremento entre las 6 y 7 a.m. en día de semana debido a la iniciación de actividades. Seguidamente se observa un crecimiento progresivo hasta las 12 del día donde presenta un pico (almuerzo y noticieros); allí comienza un descenso que finaliza a las 6 p.m. y el consumo se incrementa hasta las 8 p.m. donde la demanda es máxima (alumbrado público, retorno a casa y la cena). A partir de las 9 p.m. la demanda desciende hasta finalizar el día (descansar) (**Figura 2-5 y Figura 2-6**). la exactitud de los pronósticos de demanda depende de confiabilidad de la predicción de la tasa de crecimiento del PIB, lo cual se convierte en una desventaja teniendo en cuenta que determinar el PIB en el futuro es un problema que encierra mucha incertidumbre. La implementación de modelos ARIMA no es operacionalmente eficiente para realizar pronósticos de demanda horaria en el corto plazo, ya que para realizar predicciones se requieren tantos modelos como horas-días por estimar (Barrientos).

Figura 2-5: Curva de demanda de energía de lunes a domingo en Ucodensa del 18 al 24 de septiembre de 2017



Nombre de la fuente: Elaborado por Javier H. Velasco

Figura 2-6: Curva de demanda de energía de lunes a domingo en UCali del 17 al 24 de septiembre de 2017

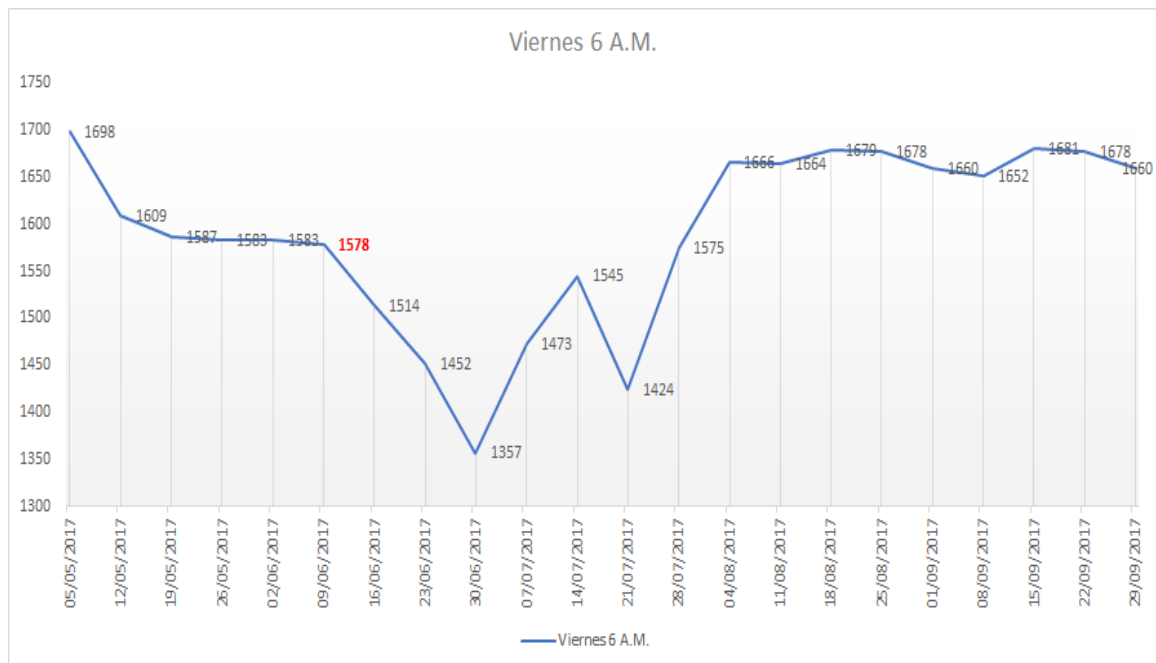


Nombre de la fuente: Elaborado por Javier H. Velasco

Otros estudios encontraron que el modelo econométrico con errores SARIMA es el óptimo para lograr el menor error de pronóstico, acercándose más a la realidad del comportamiento de la energía diaria. Esto se debe a que, al incorporar variables significativas como los retardos, las variables ficticias o de nivel, explicaran mejor los cambios estructurales en la serie (Tabares Muñoz, 2014).

2.5 Modelo Dinámico Bayesiano

El modelo clásico ARIMA, exige numerosos datos históricos para su estimación. Además, sobre dicha demanda inciden factores como la temporada, clima, temperatura, entre otros aspectos, que no pueden incorporarse. A medida que el tiempo evoluciona, la información relevante se recibe, revisa y actualiza continuamente. Esto indica que los nuevos pronósticos estarán condicionados en dicha información actualizada. Este estudio realiza el pronóstico con menos datos, pero al no tener un histórico se llega al error de aumentar el MAPE, es decir la semana antes de semana santa tiene factores que influyen como si la gente viaja o no, de acuerdo a la economía, el clima y eventos que atraen turistas, además otra fecha significativas que afectan la demanda de energía son las vacaciones de los colegios ya que se observa un descenso en unas determinadas horas (**Figura 2-7 y Figura 2-8**) (Valencia Cárdenas, 2013).

Figura 2-7: viernes 6 a.m. UCP Ucodensa entre mayo y septiembre de 2017.

Nombre de la fuente: Elaborado por Javier H. Velasco

Figura 2-8: Nota del periódico EL TIEMPO sobre vacaciones en junio de 2017

No se aplazarían vacaciones de medio año por paro de maestros

Por: Redacción Vida/Educación | 13 de junio 2017, 12:19 p.m.

El viernes pasado, algunos colegios oficiales salieron a vacaciones y este que viene, según el calendario escolar, debería salir el resto. Sin embargo, desde hace más de un mes, ocho millones de estudiantes están en un descanso obligado y anticipado. A estas alturas del paro, en las que aún no se ve un avance significativo, los padres se preguntan si habrá o no vacaciones de mitad de año.

Según la ministra de Educación, Yaneth Giha, “muchos colegios están en vacaciones y otros saldrán el próximo viernes, así que **no hemos hablado de aplazamiento de vacaciones**. Lo que sí les hemos pedido a las Secretarías de Educación es que hagan los ajustes necesarios al calendario escolar para que los maestros repongan las clases de manera efectiva y que los niños y jóvenes no se vean afectados”.

Nombre de la fuente: <http://www.eltiempo.com/vida/educacion/paro-de-maestros-no-aplazara-vacaciones-de-medio-ano-98480>

2.6 Modelo de regresión

Los modelos de regresión analizan el efecto que tiene una serie de variables explicativas x_i sobre una variable respuesta y (Fernández de Mesa Bustelo, 2016). Los modelos de regresión pueden ser simples o múltiples, en las regresiones simples se tiene una única variable predictora (ver Ecuación (2.1)), pero cuando se requieren dos o más variables predictoras se utiliza la regresión múltiple (Carrasquilla-Batista, 2016).

La ecuación de la regresión lineal simple es:

$$\hat{Y}_t = a + bt \quad (2.1)$$

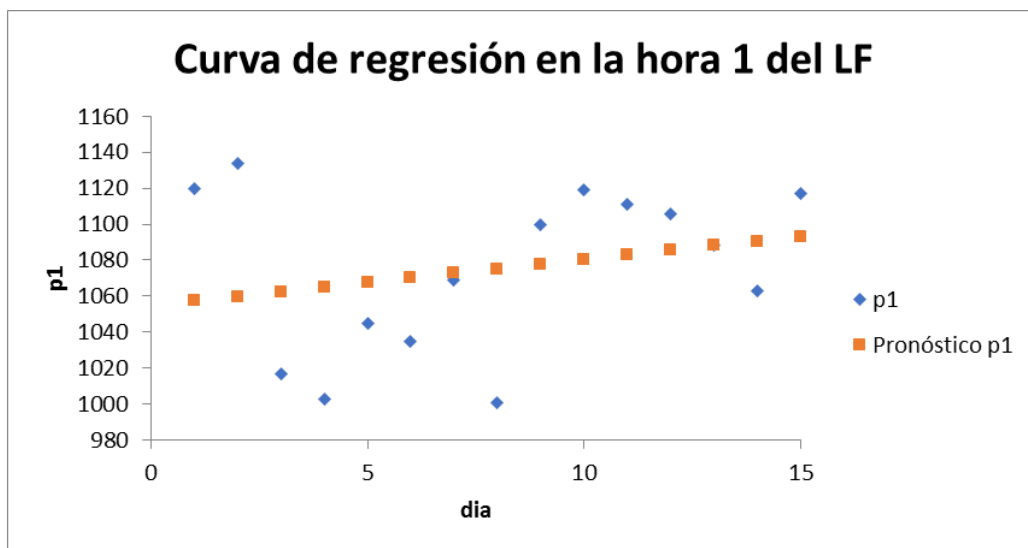
$\hat{Y}_t$: pronóstico del periodo t

a : intersección de la línea con el eje

b : pendiente (positiva o negativa)

t : periodo de tiempo

Figura 2-9: Curva de regresión simple



Nombre de la fuente: Elaborado por Javier H. Velasco

Como se observa en la **Figura 2-9**, las distancias entre P1 y la regresión Pronóstico P1 se llama residuos, es decir que la diferencia entre el valor observado y el valor aproximado es el residuo, entre más cerca se encuentren los puntos a la línea de regresión mejor es el ajuste entre la línea y el dato (Carrasquilla-Batista, 2016).

2.7 Clasificadores individuales

Los clasificadores tienen algunos de los métodos de clasificación más comunes:

- **Métodos estadísticos clásicos:** Clasificador bayesiano simple o Naive Bayes comprende la estructura de una red bayesiana que considera todas las variables predictoras como condicionalmente independientes de acuerdo a la variable a clasificar (Soler Flores, 2014)
- **Aprendizaje simbólico:** El Árboles de decisión representa una disyunción de conjunciones sobre los valores de los atributos. Donde cada rama, de la raíz a un nodo hoja, tiene correspondencia a una conjunción de atributos y el árbol en sí, a una disyunción de estas conjunciones (Arencibia R, 2012).
- **Modelos de dependencia:** Redes bayesianas, las cuales pueden calcular la distribución de probabilidad para cualquier subconjunto de variables de la red dado los valores o distribuciones de las variables restantes (Arencibia R, 2012).
- **Redes neuronales artificiales:** Pueden definirse como el conjunto de unidades computacionales llamadas neuronas interconectadas por medio de arcos pesados a manera de grafo dirigido (Arencibia R, 2012).
- **Máquinas de soporte vectorial** (Support Vector Machine, SVM) fundamenta las decisiones de clasificación, no basadas en todo el conjunto de datos, sino en un número finito y reducido de casos, que constituyen los “vectores soporte” (Arencibia R, 2012).

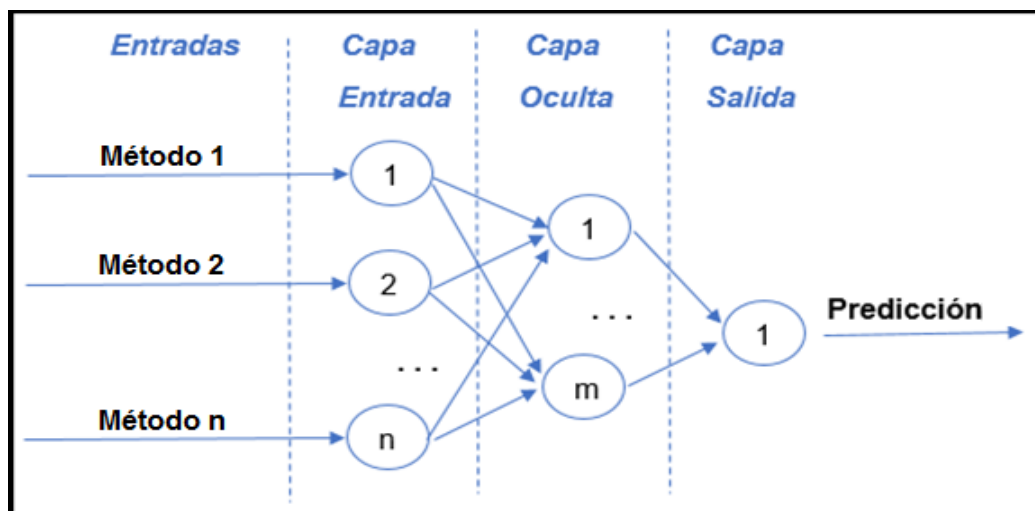
2.8 Modelo de redes neuronales

El análisis Los modelos de redes neuronales artificiales (RNA) son robustos, adaptativos y tienen como características principales la auto-organización y adaptabilidad, las RNA tiene como base los algoritmos de aprendizaje adaptativo que tiene un procesamiento no lineal y que aumenta la capacidad de la RNA para aproximar funciones y clasificar los patrones

(Vásquez Macedo, 2013). Algunas de las RNA más conocidas y utilizadas para predicción son las redes del perceptrón multicapa o Multilayer Perceptron (MLP), que usando neuronas ocultas con funciones no lineales es capaz de aproximar cualquier tipo de función continua y brindar excelentes resultados en las tareas de clasificación (Arencibia R, 2012).

De acuerdo a estudios de (Tudela, Sanjinés, y Gimmy Nardó) el mínimo error se logra utilizando redes neuronales artificiales, donde concluyen que el beneficio económico por la implementación de los modelos de redes neuronales artificiales en la predicción de la demanda de potencia eléctrica se presenta por la reducción de los costos incrementales generados por el error de pronóstico. Las Redes Neuronales no requieren que la tarea a ejecutar se programe, sino que ellos generalizan y aprenden de la experiencia. Estas RNA se utilizan para la clasificación y predicción de datos y variables, el tipo de red neuronal artificial utilizado es el Perceptrón Multicapa (**Figura 2-10**) (Tabares Muñoz, 2014).

Figura 2-10: Esquema de una red neuronal artificial



Nombre de la fuente: Elaborado por Javier H. Velasco

La utilización de inteligencia artificial para semanas normales es confiable, pero es deficiente para semana santa y semana de vacaciones. Es posible hacer la separación de las semanas en 5 grupos para la predicción. 1) Demanda en semanas normales. Este es el grueso de las semanas de predicción en el año. 2) Demanda en semanas con lunes festivo: lunes y martes cambian su comportamiento por efecto del festivo. Los demás días

siguen un comportamiento normal. 3) Demanda en semanas dentro del periodo de vacaciones. 4) Demanda en semana correspondiente a Semana Santa. 5) Demanda en semanas con días festivos diferente de lunes (Sarmiento, 2008).

2.9 Clasificación de las redes neuronales

Las redes neuronales pueden ser clasificadas como supervisadas, no supervisadas e híbridos como se observa en la siguiente tabla:

Tabla 2-1: Clasificación de las redes neuronales (Álvarez)

SUPERVISADO	UNIDIRECCIONALES	Perceptrón	Rosenblatt	1958
		Adaline	Widrow y Hoff	1960
		Perceptrón Multicapa Backpropagation	Rumelhart, Hinton y Williams	1986
		Correlación en cascada	Fahlman y Lebiere	1990
		Máquina de Bolzman	Ackley, Hinton y Sejnowski	1985
		Learning Vector Quantization	Kohonen	1988
		General Regresión Neural Network	Specht	1991
	RECURRENTES	Recurrent backpropagation	Pineda	1989
		Brain State in a Box	Anderson, Silverstein, Ritz y Jones	1977
		Jordan	Jordan	1986
NO SUPERVISADO	UNIDIRECCIONALES	Neocognitrón	Fukushima, Miyake e Ito; Fukushima	1983; 1988
		Principal Component Analysis	Oja; Hertz	1982; 1991
		Mapa autoorganizado	Kohonen	1982; 1995
	RECURRENTES	Red de Hopfield	Holfield	1982
		Resonancia Adaptativa (ART)	Carpenter, Grossberg y Rosen	1987; 1990; 1991
		Bidirectional Associative Memory	Kosko	1992
HIBRIDO		Radial Basis Funtion	Broomhead y Love; Moody y Darken	1988; 1989
		Contrapropagación	Hetcht-Nielsen	1990

- **Aprendizaje supervisado:** El proceso de aprendizaje se realiza mediante un entrenamiento controlado por un agente externo (supervisor, maestro) que determina la respuesta que debería generar la red a partir de una entrada determinada. El supervisor controla la salida de la red y en caso de que ésta no coincida con la deseada, se procederá a modificar los pesos de las conexiones, con el fin de conseguir que la salida obtenida se aproxime a la deseada (Matich, 2001).
- **Aprendizaje no supervisado:** conocidas como autosupervisado, no requieren influencia externa para ajustar los pesos de las conexiones entre sus neuronas. La red no recibe ninguna información por parte del entorno que le indique si la salida generada en respuesta a una determinada entrada es o no correcta. Estas redes deben encontrar las características, regularidades, correlaciones o categorías que se puedan establecer entre los datos que se presenten en su entrada (Matich, 2001).
- **Redes monocapa:** Establecen la conexión entre las neuronas que pertenecen a la única capa que constituye la red. Las redes monocapas se utilizan generalmente en tareas relacionadas con lo que se conoce como autoasociación (regenerar información de entrada que se presenta a la red de forma incompleta o distorsionada) (Matich, 2001).
- **Redes multicapa:** Las redes multicapas son aquellas que disponen de un conjunto de neuronas agrupadas en varios niveles o capas. Normalmente, todas las neuronas de una capa reciben señales de entrada desde otra capa anterior (la cual está más cerca a la entrada de la red), y envían señales de salida a una capa posterior (que está más cerca a la salida de la red). A estas conexiones se las denomina conexiones hacia delante o feedforward. Sin embargo, en un gran número de estas redes también existe la posibilidad de conectar la salida de las neuronas de capas posteriores a la entrada de capas anteriores; a estas conexiones se las denomina conexiones hacia atrás o feedback (Matich, 2001).

2.10 Codificación de datos de entrada

Para realizar la codificación de los datos de entrada a una red neuronal artificial se deben codificar previamente, es decir se deben encontrar valores apropiados para representar

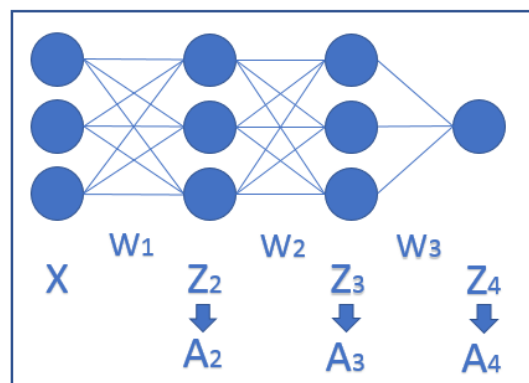
las características simbólicas en términos de números. Las variables a codificar se pueden clasificar en los siguientes tipos:

- **Variables o atributos numéricos (frecuentemente llamadas continuas):** Un atributo numérico es aquel que puede tomar cualquier valor dentro de un cierto intervalo $[a, b]$; donde a puede ser $-\infty$ (menos infinito) y b , ∞ (infinito).
- **Variables o atributos simbólicos (frecuentemente llamados discretos):** Un atributo simbólico es cuando los pesos son dados por un cierto número de términos, semejantes a: alto o bajo, sí o no, etc. Por lo tanto, los valores discretos se pueden construir en un atributo continuo pseudodiscreto. Donde se asume que todas las entradas deben transformarse dentro de los intervalos $[0, 1]$ o $\{0, 1\}$ y que por extensión sería $[-1, 1]$ o $\{-1, 1\}$ (Matich, 2001).

2.11 Algoritmo Backpropagation

El análisis El nombre de backpropagation resulta de la forma en que el error es propagado hacia atrás desde la capa de salida a través de la red neuronal artificial. Esto permite que los pesos sobre las conexiones de las neuronas ubicadas en las capas ocultas cambien durante el entrenamiento (Matich, 2001). El algoritmo Backpropagation para redes multicapa trabaja bajo el aprendizaje supervisado, por lo que necesita un conjunto de instrucciones de entrenamiento que le describa cada una de las salidas y su valor de salida esperado (Lozano, 2017). Ver **Figura 2-11**

Figura 2-11: Red neuronal artificial con conexiones y nodos



Nombre de la fuente: Elaborado por Javier H. Velasco

El algoritmo Backpropagation utiliza el método de propagación que es el proceso del cálculo hacia delante (Feedforward o propagation). Las primeras entradas de la red, son calculadas como las salidas de la primera capa oculta. Luego, usando estos valores como entradas a la segunda capa oculta, son calculadas las salidas de esta capa, y así sucesivamente para finalmente obtener la salida de la red.

Cálculo de Propagación:

$$a) \quad XW_1 = Z_2 \quad (2.2)$$

$$b) \quad A_2 = F(Z_2) \quad (2.3)$$

$$c) \quad A_2 W_2 = Z_3 \quad (2.4)$$

$$d) \quad A_3 = F(Z_3) \quad (2.5)$$

$$e) \quad A_3 W_3 = Z_4 \quad (2.6)$$

$$f) \quad Y^* = F(Z_4) \quad (2.7)$$

La función de costo o función objetivo es representada por $J(Y^*)$. Donde el costo de la red neuronal artificial es la suma de todos los errores (ver Ecuación (2.8)).

Función de costo:

$$Jy^* = 2y \cdot y^*2 \quad (2.8)$$

La salida de la red es comparada con la salida deseada, la diferencia de estos dos valores, llamado el error, es entonces usado para ajustar los pesos, primero en la capa de salida, luego en la capa anterior, y así sucesivamente, a fin de disminuir el error. Este recorrido del cálculo del error de la salida hacia la entrada es llamado la retropropagación (backpropagation) del error.

Cálculo de Backpropagation:

$$\partial Jw3 = -Y - Y^*F'Z4A3T \quad (2.9)$$

$$\partial Jw2 = -Y - Y^*F'Z4W3TF'Z3A2T \quad (2.10)$$

$$\partial Jw1 = -Y - Y^*F'Z4W3TF'Z3W2TF'Z2XT \quad (2.11)$$

Las neuronas de la capa de entrada no realizan ningún cálculo, ellas simplemente distribuyen las entradas x_i a los pesos w_{ij} de la capa oculta. Las neuronas en la capa oculta contienen la función de activación, en tanto las neuronas de salida son lineales. Al final se cuenta con una red completamente conectada, con un número amplio de conexiones en la red. Con n nodos de entrada, h nodos ocultos y m nodos de salida, para obtener la fórmula $h(m + n)$ conexiones en la red neuronal artificial (Hernández, 2018).

Después del proceso de entrenamiento los pesos de las conexiones en la red neuronal artificial quedan fijos. Como paso siguiente se debe comprobar si la red neuronal puede resolver nuevos problemas, del tipo general, para los que ha sido entrenada. Por lo tanto, con el propósito de validar la red neuronal se requiere de otro conjunto de datos, denominado conjunto de validación o testeo (Matich, 2001).

2.12 Desarrollo de un modelo de predicción

El desarrollo de modelos de predicción, generalmente está dividido en tres pasos generales:

- **Paso 1 - Construcción del modelo de predicción:** En esta etapa el modelo es desarrollado a partir de los predictores candidatos. Se utiliza la información de los datos históricos de demanda de energía al cual se han medido los diferentes predictores (Reyes S., 2014).
- **Paso 2 - Validación:** Evalúa la aproximación de la predicción al valor real teniendo como base el valor propuesto por la UCP. La validación puede ser realizada con validación cruzada para evaluar el modelo con los diferentes datos de prueba (Reyes S., 2014).
- **Paso 3 - Evaluación del modelo de predicción:** Se determina si la aplicabilidad del modelo en la práctica disminuye el error porcentual absoluto con el que son medidas las UCP y de esta manera determinar una reducción en los costos operacionales (Reyes S., 2014).

2.13 Modelos de combinación de clasificadores

Los modelos de combinación de clasificadores o multclasificadores son desarrollados por algoritmos que combinan las salidas y tiene como base la selección de diversos clasificadores, ya que se puede obtener una medida de correlación que existe entre los resultados de los diferentes clasificadores evitando la mala clasificación por los clasificadores de base. La forma en que se combinan las salidas de los clasificadores de base determina el resultado del multclasificador, agrupándolos en dos estrategias de combinación, que son selección y fusión. Entre los multclasificadores más utilizados tenemos Bagging, Boosting y Stacking (Arencibia R, 2012).

2.13.1 Bagging

Bagging es un modelo que construye N clasificadores base y cada uno de ellos utilizando el mismo algoritmo, la diferencia radica en que utiliza diferentes conjuntos para realizar el entrenamiento. Cada uno de estos conjuntos de entrenamiento se obtiene del muestreo con reemplazamiento (bootstrap) de un determinado porcentaje de instancias del conjunto de entrenamiento original. Este multclasificador, realiza predicciones a partir de las predicciones de los clasificadores base, teniendo como resultado que la clase con más votaciones será la seleccionada (Arencibia R, 2012).

2.13.2 Boosting

Boosting es un modelo que utiliza un único modelo de clasificación, el cual es entrenado utilizando muestras con reemplazo de la base de entrenamiento y al igual que Bagging utiliza voto mayoritario para combinar las salidas de los diferentes modelos. La diferencia entre los modelos Bagging y Boosting radica en que Boosting es iterativo. Mientras que en el Bagging cada modelo se construye por separado (Arencibia R, 2012).

2.13.3 Stacking

Stacking es un modelo conocido como Stacked Generalization. A diferencia de los modelos anteriores, Stacking utiliza diferentes modelos de clasificación como clasificadores de base y utiliza la combinación de las salidas de los clasificadores sin seleccionar el algoritmo con

mayor votación. Este meta clasificador tiene como objetivo aprender de las salidas de los clasificadores de base, cuáles son sus salidas más confiables y así determinar cuál es la mejor manera de poder combinarlas (Arencibia R, 2012). Es por ello que este modelo es seleccionado para realizar la investigación ya que cumple con las características que se plantean en el desarrollo de la investigación como son los valores reales históricos de la demanda de energía y la selección de los 12 modelos de predicción como clasificadores de base.

En principio, un modelo es mejor predictor que otro si las predicciones se encuentran más cercanas a los valores reales de la variable predicha. Para identificar esta situación se realizan predicciones ex-post, es decir, cuando los datos de entrada son divididos en 2 grupos, uno para realizar el entrenamiento y otro grupo para realizar las pruebas, este último grupo tiene como finalidad realizar las validaciones y poder tener una estimación de los valores (Vicentin Masaro, 2016).

2.14 Validación

La validación busca verificar si el modelo de predicción obtiene valores de predicciones más cercanas a los valores reales de la demanda de energía que los valores de los pronósticos reportados por las UCP's de Colombia. Existen dos clases de validaciones, la validación Externa y la Interna. En la validación externa, los datos son diferentes de los datos de donde fue desarrollado el modelo de predicción. Se pueden encontrar varios tipos de validación entre algunas de ellas tenemos: Validación temporal, Validación geográfica y Validación completa independiente.

La validación interna presenta diferentes técnicas entre algunas de ellas tenemos:

- **Validación interna aparente:** La validación es realizada con el 100% de la muestra con que fue desarrollado el modelo, produciendo estimaciones optimistas, entendiéndose optimistas como la diferencia entre el desempeño obtenido en la muestra y el desempeño verdadero del modelo (Reyes S., 2014).
- **Validación Split-sample:** La validación divide la muestra en forma aleatoria en dos partes. Una parte será destinada para desarrollar el modelo y la otra se realizará la validación interna. Este método de validación tiene varias desventajas como

creación de modelos inestables, desbalance en la división de la muestra, encontrar pobre desempeño del modelo (Reyes S., 2014).

- **Validación cruzada:** En la validación cruzada la muestra es dividida (1/7). El desarrollo del modelo se realiza con el 80% de la muestra. La validación es realizada en una de la décima parte de la muestra, pero se repite varias veces con las otras partes para obtener resultados estables (Reyes S., 2014).

2.14.1 Validación cruzada

La validación cruzada es una técnica que ayuda validar que los valores generados por el modelo se encuentran más cercanos a los valores reales. La validación cruzada o cross-validation utiliza los métodos hold-out o k-fold, para el desarrollo de esta investigación se utilizó el método k-Folds, con el cual se han obtenido resultados más estables. A continuación, se describirán los tipos de métodos utilizados en la validación cruzada.

2.14.1.1 Método hold-out

La Consiste en separar el conjunto de datos en dos subconjuntos, uno de los subconjuntos será utilizado para entrenar el modelo y el otro subconjunto se utilizará para realizar el test de validación. Con el subconjunto de Entrenamiento se creará únicamente el modelo. A partir del modelo creado se generarán datos de salida que se utilizarán para comparar con el subconjunto de datos de test para realizar la validación (Pérez-Planells, 2015).

2.14.1.2 Método k-fold

Está basado en el método hold-out, con la diferencia que el total de los datos se dividen en k subconjuntos, de manera que el método hold-out se repetirá k iteraciones, para cada k se utilizará un subconjunto distinto para validar el modelo entrenado con los otros k-1 subconjuntos (Pérez-Planells, 2015). Para el error de la validación cruzada de K iteraciones se realiza un cálculo de error obtenidos, ver Ecuación (2.12):

$$MAPE = \frac{\sum_{i=1}^n \left(\frac{|y - \hat{y}|}{y} \right) * 100}{n} \quad (2.12)$$

Finalmente se realiza el cálculo del indicador del error mediante MAPE a los resultados de cada iteración y también para los valores de la propuesta de la UCP y de esta forma determinar si el MAPE menor corresponde al modelo propuesto o al de la UCP, una de la desventaja de utilizar el método K-fold es que es lento desde el punto de vista computacional (Lang).

La estimación del error tiende a ser muy variable dependiendo de cuáles datos quedan en la tabla de entrenamiento y cuáles en la tabla de testing. se tiende a sobrestimar la estimación del error, es decir, es mucho mayor el error en la tabla de testing que en toda la tabla de datos (Rodríguez).

Si comparamos el método hold-out y k-fold, el método k-fold tiene mayor ventaja debido a que todos los datos serán utilizados para entrenar y validar, por lo que se obtendrán resultados más representativos a priori. Mientras que el método hold-out, realiza el proceso n veces de manera aleatoria, lo que no garantiza que los casos de entrenamiento y validación no se repetirán (Pérez-Planells, 2015).

2.15 Errores en los pronósticos

La forma de determinar si un pronóstico tendrá buenos resultados es mediante la medición de sus errores, para ellos existen diferentes indicadores que de acuerdo a sus características se calcula el valor del error, entre los indicadores de errores más utilizados tenemos:

- Error promedio(AE)
- Desviación absoluta media (MAD)
- Error cuadrado medio (MSE)
- Desviación típica de los errores (SDE)
- Error porcentual (PE)
- Error porcentual medio (MPE)
- Error Porcentual Absoluto Medio (MAPE)

Error Porcentual Absoluto Medio: (MAPE o *Mean Absolute Percentage Error*) es un indicador del desempeño del Pronóstico de Demanda que mide el tamaño del error (absoluto) en términos porcentuales. El hecho que se estime una magnitud del error porcentual lo hace un indicador frecuentemente utilizado por los encargados de elaborar pronósticos debido a su fácil interpretación (Santillán Arroyo), En la vida cotidiana el MAPE ha obtenido buenos resultados cuando el error porcentual se encuentra en valores por debajo del 20% (Ramón Sierra).

Cálculo del **MAPE** o **Error Porcentual Absoluto Medio** (ver Ecuación (2.13)):

$$MAPE = \frac{\sum_{t=1}^n \frac{|A_t - F_t|}{|A_t|}}{n} \quad (2.13)$$

El MAPE normaliza los cálculos de error mediante el cálculo de un error porcentual, permitiendo comparar los errores de pronóstico para distintos datos de series de tiempo. Por ejemplo, si una serie de tiempo tiene valores bajos de demanda y otra otros mucho más elevados, el MAPE será una forma precisa de comparar los errores para estas dos series de tiempo (Mariana Santillán Arroyo)

2.16 Minería de texto

En la extracción del conocimiento se han definido algunas etapas, donde una de las etapas es la minería de datos cuyo objetivo tiene la identificación de patrones válidos y modelos, a partir de técnicas de predicción como la minería de texto que se enfoca en datos textuales (Mariñelarena-Dondena, 2017). Las redes sociales como Twitter proveen información no estructurada y es una gran fuente de información subjetiva en tiempo real, debido a que millones de usuarios comparten opiniones en Twitter y toda esta información se descarga y se procesa con alto volumen de datos históricos (Maté, 2016).

Dentro de las tareas predictivas se tiene la construcción de clasificadores automáticos, que pueden estimar variables dependientes a partir de variables independientes que son extraídas de unidades de datos textuales como lo es Twitter. La extracción de las

características se puede realizar por medio de enfoques de vocabulario cerrado o abierto, para el caso de cerrado se emplean diccionarios manuales con palabras definidas por expertos y para el caso abierto el diccionario se determina a partir del conjunto de datos objeto de estudio (Mariñelarena-Dondena, 2017).

Existen algoritmos que se pueden utilizar para extracción de palabras textualmente como lo es el algoritmo TF-IDF (Term Frequency-Inverse Document Frequency), que extrae palabras que ayudan al cálculo del análisis y procesamiento multidimensional textual. Para realizar el cálculo se debe tener en cuenta las siguientes fórmulas, ya que para calcular la frecuencia de términos deben darse pesos de acuerdo al número de palabras y para la frecuencia de documentos inversos es el logaritmo del número de texto en documento sobre el número de palabras (Ren, 2017).

La fórmula de TF-IDF es:

$$w = TF * IDF \quad (2.14)$$

Donde TF es:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (2.15)$$

n_{ij} : Cantidad de frecuencias de ocurridas en el texto

$\sum_k n_{k,j}$: El denominador es la suma de todas las palabras en el texto

$$idf_i = \log \frac{|D|}{|\{j: t_i \in d_j\}|} \quad (2.16)$$

$|D|$: Representa el número de texto en un documento

$\{j: t_i \in d_j\}$: El denominador indica el número de palabras contenidas en el texto

Al realizar la extracción de palabras se debe tener en cuenta que existen palabras derivadas con las que no se puede realizar el procesamiento y se requiere obtener la palabra base o raíz a este método se llama Stemming, en algunos casos estas palabras no tienen sentido, pero al tener una palabra predecesora la palabra tiene sentido, por lo que se requiere la utilización de técnicas de Lemmatización (Agarwal, 2016).

3.Estado del Arte

3.1 Modelos públicos actuales

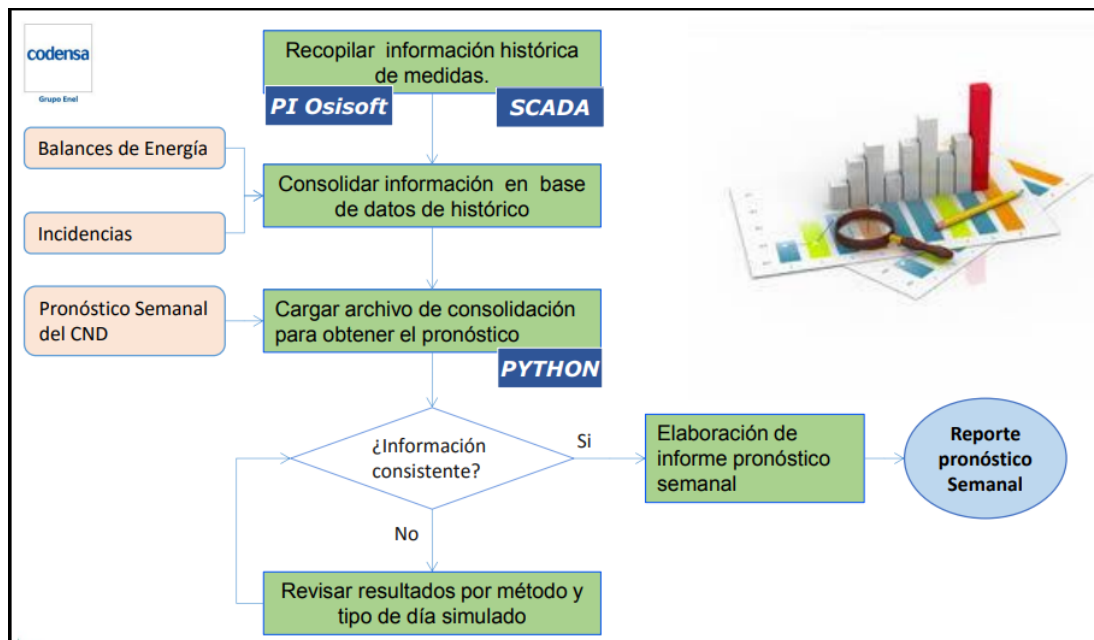
En Colombia existen 35 UCP's comerciales, pero solo Codensa y Electricaribe han hecho público la metodología y el método que utilizan para realizar el pronóstico de la demanda de energía, esta información se encuentra en el portal del Consejo Nacional de Operación (CNO) en la sección de noticias.

3.1.1 Modelo de predicción de Codensa

Codensa es una empresa dedicada a la Distribución y comercialización de energía eléctrica, líder en el mercado nacional con el 22,36% de la participación total. Tiene presencia en Bogotá, en 100 municipios de Cundinamarca, 8 municipios de Boyacá y 1 de Tolima².

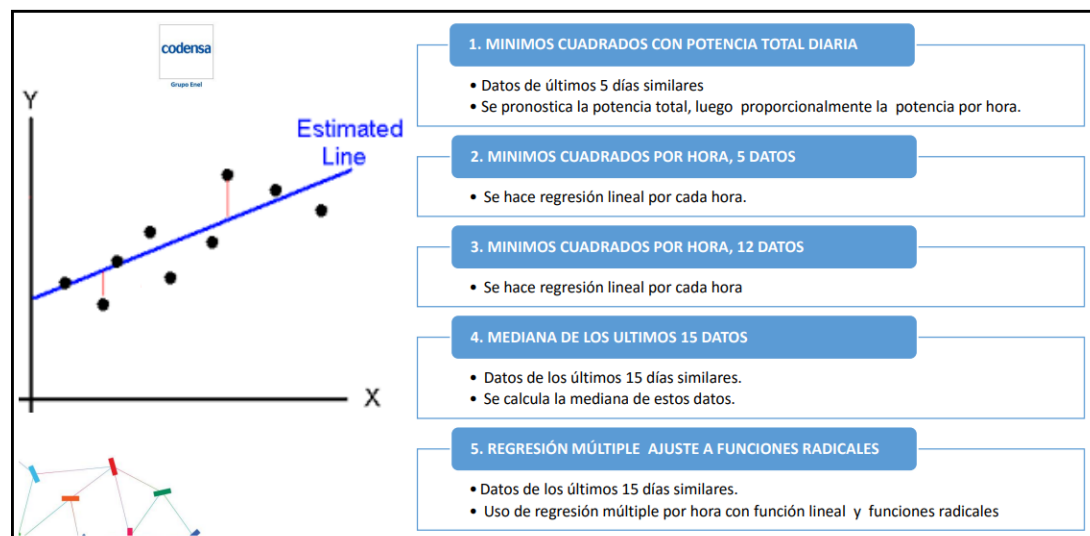
El modelo de predicción de Codensa dentro de su metodología (ver **Figura 3-1**) contempla la consolidación de la información de los datos históricos, obtenida de los medidores frontera, los balances de energía y las incidencias o demanda no atendida. Luego se procede a validar que la información del pronóstico semanal del Centro nacional de despacho (CND) sea consistente, pero en caso contrario esta información será revisada por el aplicativo que tiene Codensa para validar pronósticos de demanda, en caso de que los valores se encuentren correctos se procede a elaborar y reportar el pronóstico semanal (Codensa, 2016).

² <http://corporativo.codensa.com.co/es/conocenos/nuestronegocio/Paginas/home.aspx>

Figura 3-1: Metodología aplicada por Codensa

Nombre de la fuente:

https://www.cno.org.co/sites/default/files/documentos/noticias/Pronostico_Codensa_22072016_v1.pdf

Figura 3-2: Métodos usados por Codensa

Nombre de la fuente:

https://www.cno.org.co/sites/default/files/documentos/noticias/Pronostico_Codensa_22072016_v1.pdf

El proceso de validación de reporte del CND por parte de Codensa, tiene como base los métodos estadísticos: Mínimos cuadrados, Mediana y Regresión Múltiple ajuste a funciones radicales como se observa en la **Figura 3-2**. Cada uno de estos métodos son ejecutados de acuerdo al tipo de día que tiene la semana S+1 para pronosticar.

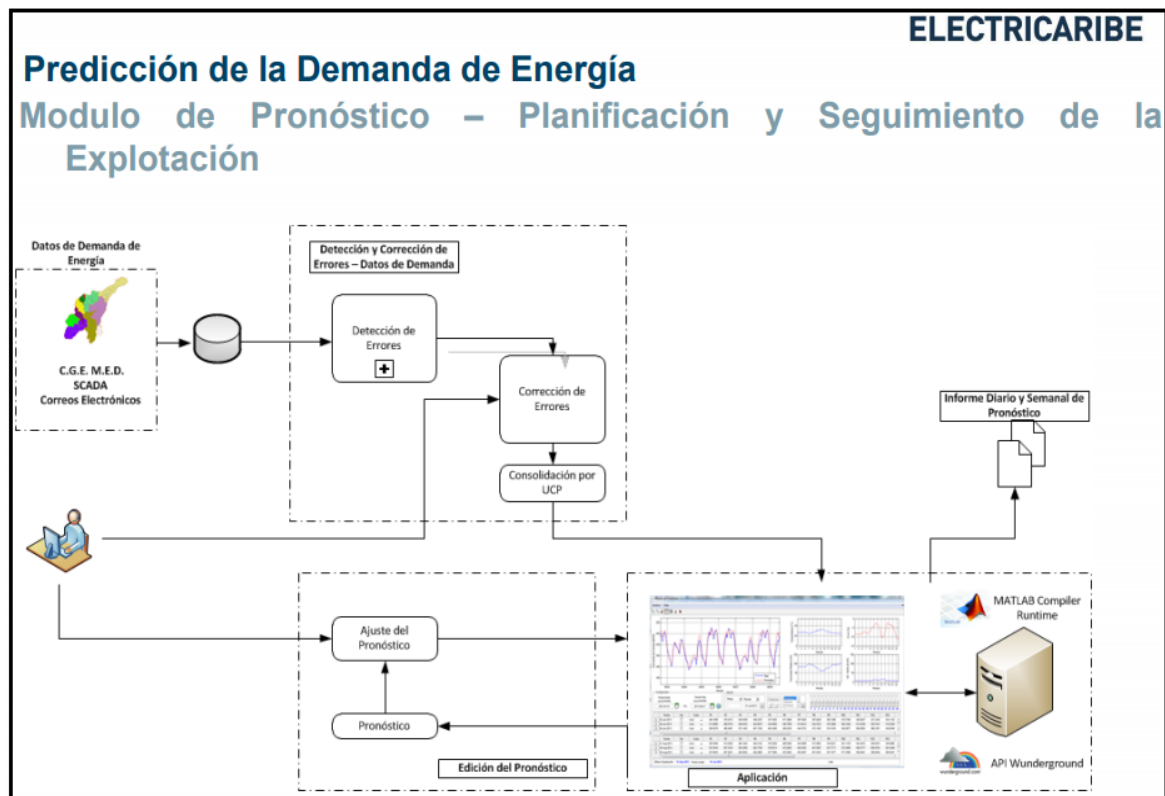
3.1.2 Modelo de predicción de Electricaribe

La compañía Electricaribe presta el servicio de distribución y comercialización de la energía eléctrica en 7 departamentos: Atlántico, Bolívar, Sucre, Córdoba, Cesar, Magdalena y La Guajira. Como operador de red tiene a su cargo la administración de las siguientes UCP's (Electricaribe, 2016):

- **UCP Barranquilla:** Demanda de energía del departamento del Atlántico y algunas demandas del departamento de Magdalena.
- **UCP Cartagena:** Demanda de energía del departamento de Bolívar y algunas demandas del departamento de Magdalena.
- **UCP Planeta:** Demanda de energía del departamento de Córdoba.
- **UCP Sinu:** Demanda de energía del departamento de Córdoba, Sucre y algunas demandas del departamento de Bolívar.
- **UCP Tairona:** Demanda de energía del departamento de Guajira, Cesar, Magdalena y algunas demandas del departamento de Bolívar.

El modelo de predicción de Electricaribe cuenta con 3 módulos (ver **Figura 3-3**):

1. Detección y corrección de errores - Datos de Demanda
2. Edición del pronóstico
3. Aplicación

Figura 3-3: Módulos usados por Electricaribe

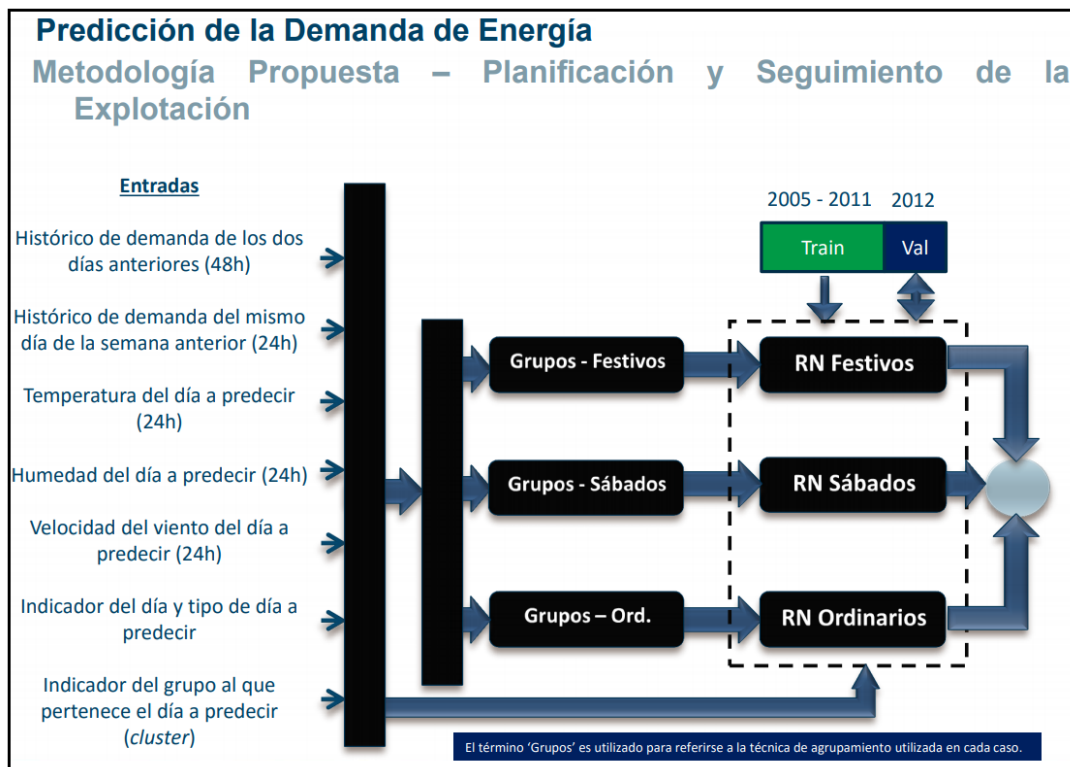
Nombre de la fuente:

https://www.cno.org.co/sites/default/files/documentos/noticias/Pronostico_Codensa_22072016_v1.pdf

En el modelo propuesto por Electricaribe tiene como fuente de información el Supervisory Control And Data Acquisition (SCADA) y para el cálculo de la predicción a definido un grupo por cada tipo de día y para cada grupo utiliza una red neuronal con los siguientes datos de entrada (ver **Figura 3-4**):

- Históricos de demanda de los dos días anteriores (48h).
- Históricos de demanda del mismo día de la semana anterior(24h).
- Temperatura del día a predecir(24h).
- Humedad del día a predecir(24h).
- Velocidad del viento del día a predecir(24h).
- Indicador del día y tipo de día a predecir.
- Indicador del grupo al que pertenece el día a predecir

Figura 3-4: Metodología aplicada por Electricaribe



Nombre de la fuente:

https://www.cno.org.co/sites/default/files/documentos/noticias/Pronostico_Codensa_22072016_v1.pdf

Como ventajas el modelo de Electricaribe tiene:

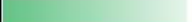
- La actualización automática de la base de datos climatológica: Capacidad para descargar la información climatológica de una base de datos ubicada en la WEB (Electricaribe, 2016).
- La Actualización automática de la base de datos de demanda: Carga, detección y corrección de errores de los datos de demanda de energía extraídos de las fuentes de información de la empresa (C.G.E. M.E.D., SCADA) y otras provenientes de usuarios industriales (PROELECTRICA) (Electricaribe, 2016).

4. Metodología

4.1 Selección de UCP

Los indicadores de demanda de energía en Colombia son publicados por la compañía XM y están conformados por la información de las 35 administradoras de UCP's comerciales y la UCP llamada Usincolgenerador que corresponde al indicador de demanda de Colombia, la cual tiene el 0% de participación en el SIN. Para la investigación se identificarán que UCP's tienen la mayor participación en el SIN, ya que con este método podremos enfocarnos en las que tienen mayor demanda de energía y así poder desarrollar el modelo que se ajustara a la UCP con mayor participación lo cual conlleva a tener un mayor impacto en la disminución de la diferencia con los valores reales y por consiguiente una disminución en el indicador de error con el cual son evaluadas las UCP.

Figura 4-1: Las 10 UCP's con mayor participación en el SIN en agosto de 2018

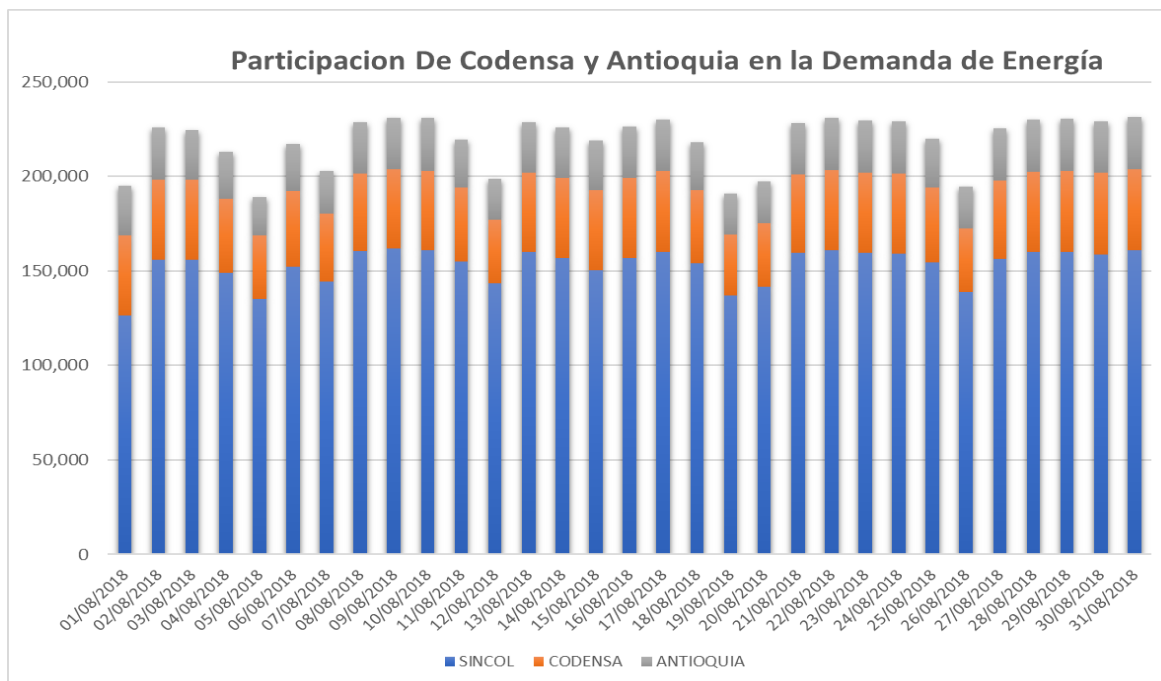
Indicadores de Calidad de los Pronósticos Resumen Mensual agosto 2018			
Nombre MC	% Participación en el SIN		MAPE (%)
Uantioquia		13.29%	 3.0%
Ucodensa		20.68%	 1.2%
Ubaranquilla		8.30%	 1.1%
Utairona		7.09%	 2.9%
Ucali		5.93%	 2.8%
Ucartagena		5.10%	 1.6%
Usinu		4.58%	 1.5%
Usantander		3.89%	 2.2%
Upacifico		3.62%	 3.7%
Uebsa		2.77%	 3.9%

Nombre de la fuente: <https://www.xm.com.co/Paginas/Consumo/indicadores-de-pronosticos-oficiales-de-demanda.aspx>

En la **Figura 4-1** se encuentran 10 de las 36 UCP's con mayor participación en el Sistema Interconectado Nacional, de las cuales observamos que en primer lugar se ubica la ucp Ucodensa con el 20.68% participación, en segundo lugar, se ubica la ucp Uantioquia con el 13.29% participación y en tercer lugar a Ubarranquilla con el 8.30% de participación en el SIN.

De acuerdo a los valores anteriores y como se puede observar en la **Figura 4-2**, se evidencia que el impacto que tienen en la demanda de energía las UCP's Ucodensa y Uantioquia es bastante alta, ya que la suma de sus participaciones es del 33.97%. Por consiguiente, la investigación se dirigirá a disminuir el error porcentual de estas 2 UCP's por medio del modelo propuesto.

Figura 4-2: Participación de Codensa y Antioquia en la demanda de energía del SIN para el mes de agosto de 2018



Nombre de la fuente: Elaborada por Javier H. Velasco

4.2 Revisión del modelo

Después de analizar las técnicas y modelo utilizados por las empresas Codensa y Electricaribe, se eligió como modelo base para iniciar la investigación los métodos presentado por Codensa, donde sus predicciones parten de modelos estadísticos como se observa en la **Figura 3-2**. A continuación, los 5 métodos utilizados por Codensa:

1. Mínimos Cuadrados con potencia Total Diaria.
2. Mínimos Cuadrados por Hora (5 Datos).
3. Mínimos Cuadrados por Hora (12 Datos).
4. Mediana de los últimos 15 datos.
5. Regresión Múltiple ajuste a funciones radicales.

Como base para el proyecto se implementaron 4 de los 5 métodos usados por Codensa, excepto regresión múltiple con ajuste a funciones radicales

4.3 Extracción de información

Para iniciar con el modelo se debe tener una información previa de energía eléctrica de la UCP a la que se realizará la implementación del modelo, como fuentes primarias y públicas se tienen los indicadores de pronósticos oficiales de demanda que son archivos en Excel (**Figura 4-3**) que se encuentran publicados mensualmente en la sección Consumo en la parte de informes de demanda del portal de XM que es la entidad encargada de operar el Sistema Interconectado Nacional de Colombia. La información que se obtendrá de estos indicadores será la reportada por las fronteras (Demanda Real). Esta información ayudará a generar nuevas predicciones y a validar el modelo y como indicador de evaluación se usará el indicador de error absoluto medio porcentual (MAPE).

Figura 4-3: Indicadores de pronósticos oficiales de demanda de Ucodensa

Ucodensa-Ofi-Ago 2017-Ago 2017.xls [Modo de compatibilidad] - Excel

ArchivoInicioInsertarDiseño de páginaFórmulasDatosRevisarVista¿Qué desea hacer?Compartir

E21234.25

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	UCP	VARIABLE	FECHA	TIPO DIA	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
2	UCODENSA	Demanda_Real	01/08/2017	MARTES	1,234	1,185	1,160	1,180	1,345	1,645	1,708	1,796	1,876	1,911
3	UCODENSA	Demanda_Real	02/08/2017	MIERCOLES	1,245	1,195	1,175	1,197	1,363	1,660	1,716	1,802	1,892	1,915
4	UCODENSA	Demanda_Real	03/08/2017	JUEVES	1,244	1,193	1,173	1,193	1,354	1,653	1,724	1,816	1,911	1,915
5	UCODENSA	Demanda_Real	04/08/2017	VIERNES	1,276	1,221	1,188	1,215	1,379	1,666	1,727	1,813	1,915	1,915
6	UCODENSA	Demanda_Real	05/08/2017	SAALF	1,283	1,221	1,190	1,185	1,245	1,316	1,442	1,616	1,763	1,915
7	UCODENSA	Demanda_Real	06/08/2017	DOALF	1,175	1,110	1,071	1,059	1,064	1,053	1,075	1,199	1,328	1,915
8	UCODENSA	Demanda_Real	07/08/2017	7--AGO	1,101	1,043	1,013	1,002	1,030	1,070	1,078	1,217	1,373	1,915
9	UCODENSA	Demanda_Real	08/08/2017	MADLF	1,150	1,105	1,087	1,118	1,292	1,602	1,667	1,752	1,861	1,915
10	UCODENSA	Demanda_Real	09/08/2017	MIERCOLES	1,246	1,193	1,170	1,194	1,358	1,646	1,719	1,799	1,899	1,915
11	UCODENSA	Demanda_Real	10/08/2017	JUEVES	1,253	1,203	1,180	1,199	1,367	1,673	1,731	1,813	1,904	1,915
12	UCODENSA	Demanda_Real	11/08/2017	VIERNES	1,269	1,214	1,187	1,208	1,371	1,664	1,743	1,842	1,935	1,915
13	UCODENSA	Demanda_Real	12/08/2017	SABADO	1,284	1,217	1,185	1,196	1,261	1,345	1,478	1,657	1,796	1,915
14	UCODENSA	Demanda_Real	13/08/2017	DOMINGO	1,195	1,125	1,090	1,064	1,063	1,065	1,089	1,220	1,361	1,915
15	UCODENSA	Demanda_Real	14/08/2017	LUNES	1,139	1,094	1,073	1,098	1,268	1,576	1,644	1,741	1,866	1,915

Diferencia Real-Pronostico

pronostico

real

Listo

Nombre de la fuente: <https://www.xm.com.co/Paginas/Consumo/indicadores-de-pronosticos-oficiales-de-demanda.aspx>

Después de extraer los archivos de la fuente de origen se procede a analizar la estructura de cada uno de los archivos, iniciamos con el análisis de los indicadores de pronósticos oficiales de demanda que se encuentran en extensión xls. Este es un libro en Excel que contiene 8 hojas, donde la hoja 7 (pronostico) corresponde al pronóstico reportado por la UCP con la variable OFI y la hoja 8 (real) corresponde a la demanda real reportada por las fronteras con la variable Demanda Real. Después de determinar las hojas de Excel, se valida que las hojas de “pronostico” y “real” tengan el mismo número de columnas (UCP, Variable, Fecha, Tipo Dia, P1, P2, ..., P24, TOTAL) para un total de 29 columnas, de las cuales 25 corresponden a valores numéricos, 3 a cadenas y 1 a fecha.

Definido los datos que se extraerán, se diseña y crea en el motor de base de datos mysql, el esquema “DemandaEnergia” que tendrá inicialmente la tabla “ucp” que almacenara los datos de las UCP’s de Colombia y la tabla “datosucp” que almacenara los datos históricos de demanda de energía real y pronosticada de la UCP. La tabla “datosucp” está compuesta por 30 campos (ver **Tabla 4-2**), correspondiente a las 29 columnas del archivo Excel más el campo “dia” que dará la referencia al día de la semana (lunes a Domingo).

Tabla 4-1: Estructura de la tabla datosucp

<i>datosucp</i>	
<i>Nombre</i>	<i>Tipo</i>
Codigoucp	Varchar
Codigovvariable	Varchar
Fecha	Date
dia	Varchar
tipodia	Varchar
P1	Int
P2	Int
...	Int
P23	Int
P24	Int
Total	Int

4.4 Selección de datos

Los estudios realizados por XM, determinaron que existen tipos de días, es decir que el comportamiento de un domingo antes de un lunes festivo es diferente a un domingo antes de un lunes normal. En la **Tabla 4-2** se observan los 42 tipos de días que definió XM.

Tabla 4-2: Tipos de días definidos por XM

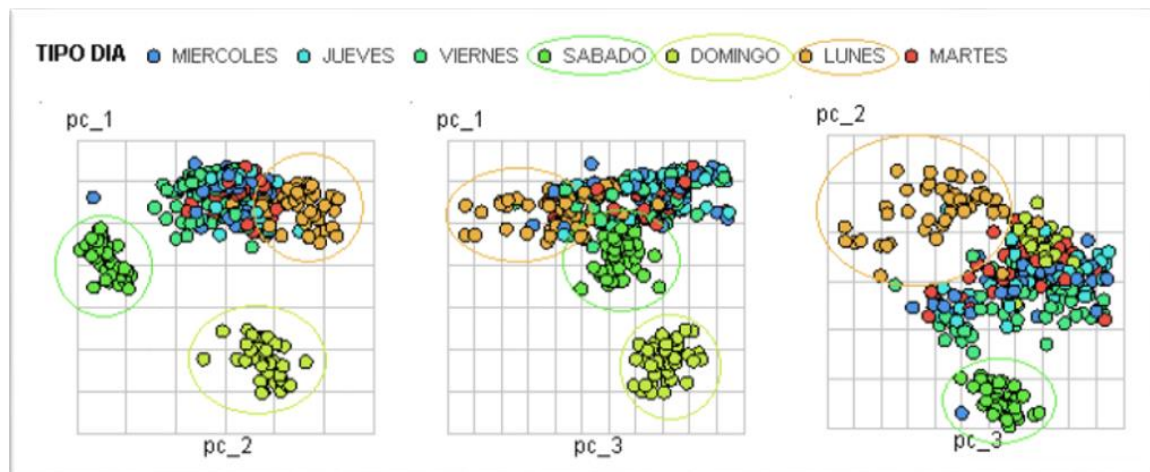
ABREV.	TIPO DE DÍA	ABREV.	TIPO DE DÍA
1-ENE	Primero de Enero	MSS	Miercoles de Semana Santa
1-MAY	Primero de Mayo	JSS	Jueves de Semana Santa
20-JUL	20 De Julio	JUEVES	Jueves
24-DIC	24 de Diciembre	JUVDIC	Jueves Vacaciones de Diciembre
25-DIC	25 de Diciembre	JUVENE	Jueves Vacaciones de Enero
2-ENE	2 de Enero	VIERNES	Viernes
31-DIC	31 de Diciembre	VIVDIC	Viernes Vacaciones de Diciembre
7-AGO	7 de Agosto	VIVENE	Viernes Vacaciones de Enero
8-DIC	8 de Diciembre	VSS	Viernes de Semana Santa
LF	Lunes Festivo	SAALF	Sabado Antes de Lunes Festivo
LFENE	Lunes Festivo de Enero	SAALFENE	Sabado Antes de Lunes Festivo de Enero
LUNES	Lunes	SABADO	Sabado
LUVDIC	Lunes Vacaciones de Diciembre	SAVDIC	Sabado Vacaciones de Diciembre
LUVENE	Lunes Vacaciones de Enero	SAVENE	SabadoVacaciones de Enero
MADLF	Martes Despues de Lunes Festivo	SSS	Sabado de Semana Santa
MARTES	Martes	DOALF	Domingo Antes De Lunes Festivo
MAVDIC	Martes Vacaciones de Diciembre	DOALFENE	Domingo Antes De Lunes Festivo de Enero
MAVENE	Martes Vacaciones de Enero	DOMINGO	Domingo
MIERCOLES	Miercoles	DOVDIC	Domingo Vacaciones de Diciembre
MIVDIC	Miercoles Vacaciones de Diciembre	DOVENE	DomingoVacaciones de Enero
MIVENE	Miercoles Vacaciones de Enero	DSS	Domingo de Semana Santa

Para el desarrollo de la investigación se utilizan los siguientes tipos de días:

LUNES, MARTES, MIERCOLES, JUEVES, VIERNES, SABADO, DOMINGO, LF, MADLF, SAALF, DOALF.

Debido al comportamiento de los tipos de días se realizó un análisis de componente principal (PCA) a la información de la UCP de Codensa y como se observa en la **Figura 4-4** los días lunes, sábado y domingo son fácilmente separables. Caso contrario para los días de martes a viernes los cuales no son separables

Figura 4-4: Análisis de componentes principales de Codensa



Nombre de la fuente: Elaborado por Javier H. Velasco

Para la selección de tipos de Datos se tuvieron en cuenta los tipos de días (*LUNES, MARTES, MIERCOLES, JUEVES, VIERNES, SABADO, DOMINGO, LF, MADLF, SAALF, DOALF.*) Los días se seleccionan usando los siguientes criterios:

- **Muestreo 1:** Se seleccionan los datos de los días similares, es decir aquellos días de las semanas anteriores que son del mismo tipo de día.
- **Muestreo 2:** Se selecciona los días con semanas similares, es decir los días de las semanas que tienen los mismos tipos de días.
LUNES-MARTES-MIERCOLES-JUEVES-VIERNES-SABADO-DOMINGO
LUNES-MARTES-MIERCOLES-JUEVES-VIERNES-SAALF-DOALF

LF-MADLF-MIERCOLES-JUEVES-VIERNES-SABADO-DOMINGO

LF-MADLF-MIERCOLES-JUEVES-VIERNES-SAALF-DOALF

- **Muestreo 3:** Como conclusión al análisis del PCA a los datos de Codensa se creó el muestreo 3, que tiene como finalidad seleccionar solo los datos de los días de martes a viernes y proporcionando un peso al día a que se requiere predecir.

4.5 Identificación de algoritmos y métodos

Para realizar la predicción de la demanda de energía utilizaremos 3 modelos estadístico como es regresión lineal, mediana y ponderación. Para la aplicación de los modelos se definirá el número de datos de entrada y se utilizaran los tres algoritmos del punto 5.4.

- **Método 1 - Mínimos Cuadrados Con Energía total Diaria:** Se pronostica la energía total con respecto a la mediana de los 5 últimos datos y se pondera cada hora con respecto al último dato del día.
- **Método 2 - Mínimos Cuadrados por hora:** Se realizará con los últimos 5 datos, se calcula con regresión lineal por cada hora
- **Método 3 - Mínimos Cuadrados por hora:** Se realiza con los últimos 10 datos, se calcula con regresión lineal por cada hora.
- **Método 4 – Mediana:** Se realiza con los últimos 15 datos, se calcula mediante la mediana para cada hora.

Como resultado de la combinación de los 3 tipos de muestreo con los 4 métodos, se obtendrán 12 predicciones (ver **Tabla 4-3**).

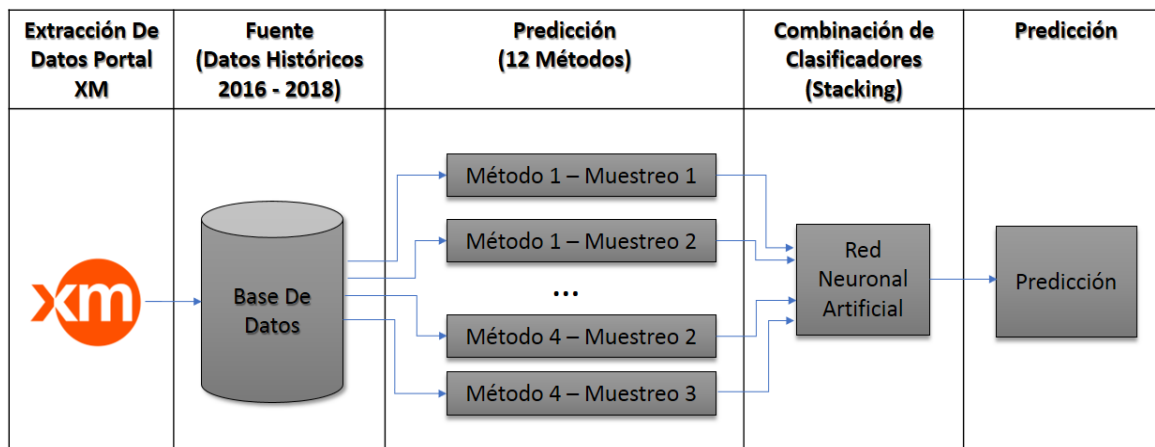
Tabla 4-3: Cruce de métodos con los tipos de muestreo

PREDICCIONES Mst.1	PREDICCIONES Mst.2	PREDICCIONES Mst.3
Método 1 – Muestreo 1	Método 1 – Muestreo 2	Método 1 – Muestreo 3
Método 2 – Muestreo 1	Método 2 – Muestreo 2	Método 2 – Muestreo 3
Método 3 – Muestreo 1	Método 3 – Muestreo 2	Método 3 – Muestreo 3
Método 4 – Muestreo 1	Método 4 – Muestreo 2	Método 4 – Muestreo 3

4.6 Modelo base

La **Figura 4-5** describe el modelo de pronóstico que se implementó. es El proceso comienza con la extracción de datos desde el portal de XM y se procede a almacenar la información de la demanda real y pronosticada por la UCP en un motor de base de datos, luego con el lenguaje de programación Python se realizará el proceso de consulta, cálculo y entrega de los 12 valores de predicción, para ser combinadas y obtener un solo valor de predicción que sea el más cercano al valor de la demanda real por hora.

Figura 4-5: Modelo Base



Nombre de la fuente: Elaborado por Javier H. Velasco

Para la UCP Codensa se descargan los datos disponibles desde julio de 2016 hasta abril de 2018, obteniendo como resultado 22 archivos en formato csv.

Esta información se almacena en la Base de Datos en la tabla “datosucp”. Como resultado del cargue obtenemos un total de 1338 registros, de los cuales 669 Registros corresponde a datos de la “Demanda_Real” y 669 corresponde a la “OFI” (Oficial o reportada por Codensa).

Se realiza la predicción para las semanas que cumplen el criterio de selección del punto 5.4, por lo que se seleccionarán las semanas compuestas por los siguientes tipos de días (LUNES, MARTES, MIERCOLES, JUEVES, VIERNES, SABADO, DOMINGO, LF, MADLF,

SAALF, DOALF). Las semanas entre el 1 de enero de 2017 hasta el 30 de abril de 2018 que cumplen el criterio son 57 semanas. Las semanas seleccionadas inician desde el año 2017, debido a que los métodos estadísticos calculan la predicción a partir de los 5 y hasta los 10 últimos datos históricos similares, por lo que se requieren los datos del 2016 para realizar la predicción.

Las salidas de los 12 modelos descritos anteriormente, son combinadas en una única predicción por medio de una red neuronal artificial que aprende por medio de datos de entrenamiento. A este proceso se le conoce como Stacking.

4.7 Stacking

Para la implementación del modelo de combinaciones de regresiones se seleccionó como algoritmo de aprendizaje la red neuronal artificial backpropagation, como variables complemento se adiciono el tipo de día y la hora, estos dos serán parte de la entrada a la red neuronal con las 12 predicciones del modelo de regresión, para un total de 14 entradas.

De acuerdo a (Hernández Camilo) los datos de entrada de la red deben ser menor a 1, y los valores de cada nodo $\propto l m$ tendrán n conexiones con valores reales entre $\{0,1\}$, los valores de activación de la capa oculta pueden ser cualquier número real que tome valor entre el intervalo $[-1,1]$ (Hernández, 2018). Debido a que se tiene una variable de predicción no numéricas se crearon valores de equivalencia que como se muestra en la

Tabla 4-4: Valor equivalente por tipo de día

TIPO DE DIA	VALOR EQUIVALENTE
LUNES	11
MARTES	12
MIERCOLES	13
JUEVES	14
VIERNES	15
SABADO	16
DOMINGO	17
LF	18
MADLF	19
SAALF	20
DOALF	21

Con los datos de las entradas de la red neuronal artificial en valores numéricos, se procede a realizar la siguiente normalización:

Tendremos como valor único de referencia para normalizar los valores de entrada la variable “divisor”, la cual tendrá un valor de 10.000.

Tipo de día: la ecuación es: $((\text{Tipo de día} \times 100) / \text{divisor})$

Por ejemplo, para el día lunes sería: $(11 \times 100) / 10000 = 0.11$

Hora: la ecuación es: $((\text{Hora} / 24) \times 10000) / \text{divisor}$

Por ejemplo, para la hora 15 sería: $(15 / 24) \times 10000 / 10000 = 0.62$

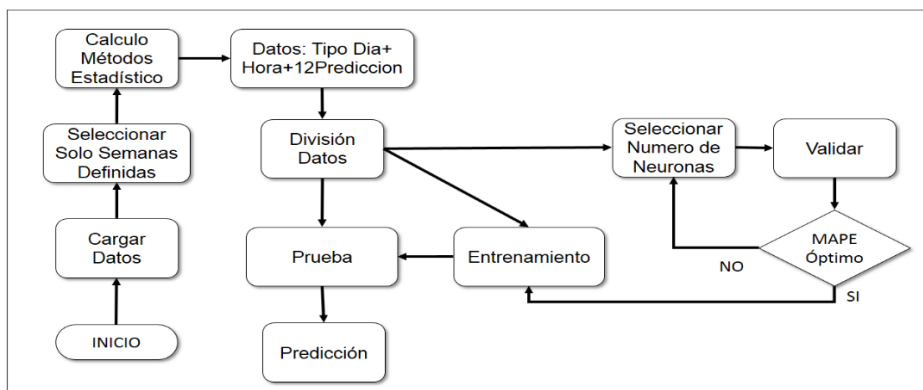
Clasificador: la ecuación es: $(\text{Clasificador N} / \text{divisor})$

Por ejemplo $(2500 / 10000) = 0.25$

4.7.1 Experimentación

Para dimensionar el número de nodos en las capas ocultas. Con la técnica de validación cruzada, se obtendrá una muestra del 80 % de los datos resultantes de las 57 semanas, para luego proceder a aplicar esta muestra entre los 7 pliegues ya definidos, donde cada pliegue tendrá el valor de error obtenido en ese modelo y se promedian para determinar un valor único de error que será referencia para seleccionar el número de nodos para el modelo más óptimo (ver **Figura 4-6**).

Figura 4-6: Modelo de Validación cruzada



Nombre de la fuente: Elaborado por Javier H. Velasco

Para determinar el costo de la red neuronal artificial, se realizará por medio de la suma de todos los errores y de esta forma se medirá el rendimiento. Con el número de iteraciones o número de veces que se modificaron los pesos se realizara el refinamiento y se encontrara la solución más óptima (ver **Figura 4-7**).

Figura 4-7: Determinación de nodos por capa

Capas Ocultas: 1 - 1 - 2									
costo: 920.72492	valor: 0.08143	iteración: 175	función: 184	gradiente: 184					
Capas Ocultas: 3 - 1 - 1									
costo: 1102.04585	valor: 0.08808	iteración: 573	función: 658	gradiente: 658					
Capas Ocultas: 3 - 7 - 3									
costo: 2292.89991	valor: 0.06898	iteración: 15196	función: 16535	gradiente: 16535					
Capas Ocultas: 4 - 12 - 5									
costo: 2876.88278	valor: 0.05786	iteración: 25225	función: 26314	gradiente: 26314					
Capas Ocultas: 7 - 14 - 11									
costo: 3289.54599	valor: 0.04423	iteración: 62262	función: 63956	gradiente: 63945					

4.7.2 Entrenamiento (Train)

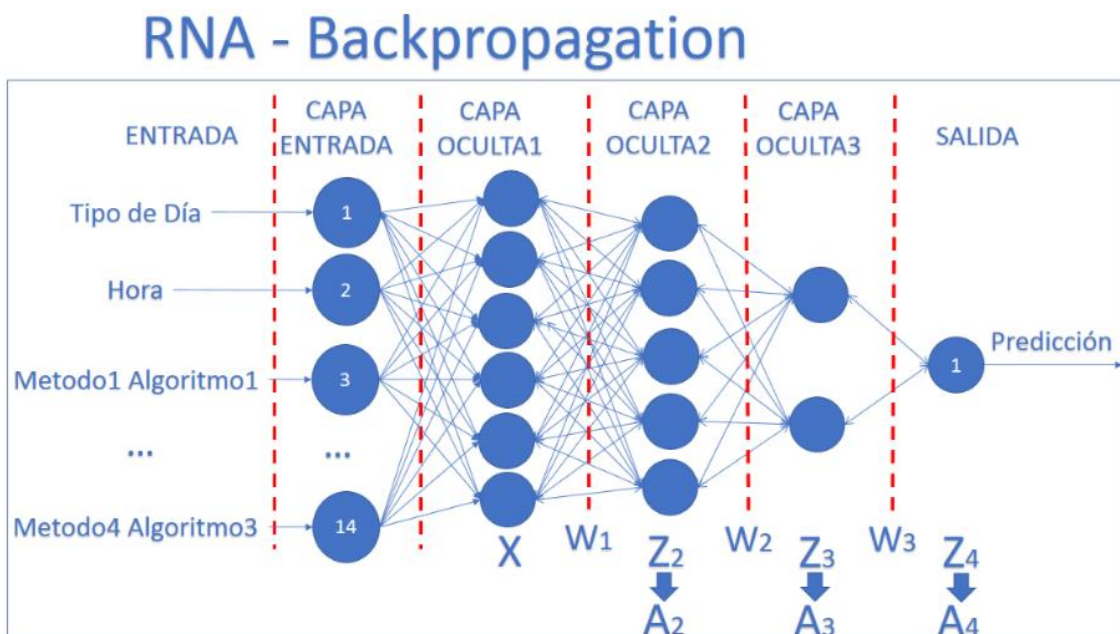
El proceso de aprendizaje o entrenamiento de la red neuronal busca los pesos que permiten desarrollar la predicción, los pesos se van modificando de manera proporcional al error que se produce entre la salida real de la red y la salida esperada, durante este proceso mediante iteraciones se refina la solución y se encuentra el conjunto de pesos sinápticos que minimizan (o maximizan) la función hasta alcanzar el punto óptimo de la red neuronal (Bertona, 2005).

El modelo SILVCAS tiene como datos fuente los reportes históricos de la UCP, pero estos reportes solo son publicados los 10 primeros días hábiles del mes, por consiguiente, en la tercera semana del mes publican el reporte, pero solo se podría iniciar la predicción desde la cuarta semana del mes con los últimos datos publicados.

Con la información y las variables definidas se procede a realizar la predicción de los 12 métodos con la configuración de la red neuronal que se observa en la **Figura 4-8**, obteniendo como resultado un archivo llamado "Datos.csv" el cual contiene los campos

(fecha, mes, día, tipodia, hora, metodo1, ...metodo12, valor ucp, valor real). El número de registro que tendrá el archivo csv, es el resultado de tener 57 semanas, donde cada semana a predecir está conformada por los 7 días independiente del tipo de día y sus respectivas 24 horas, es decir por cada semana tenemos 7 días y cada día 24 horas, por lo que tenemos 168 registros, como la predicción se realizó para 57 semanas, entonces tenemos que multiplicar las 57 semanas por los 168 registros lo que nos da un total de 9576 registros.

Figura 4-8: Red neuronal artificial con backpropagation



Nombre de la fuente: Elaborado por Javier H. Velasco

4.7.3 Prueba (Test)

El 20 % de los datos se utilizaron para realizar las pruebas, utilizando como referencia el valor real menos el valor de la predicción (ver Ecuación (4.1)) y de la propuesta de la UCP (ver Ecuación (4.2)). Cada valor hace referencia a cada tipo de día (ver **Figura 4-9**).

$$\text{Mape Silvcas} = \text{diferencia entre real} - \text{predicción} \quad (4.1)$$

$$\text{Mape Ucp} = \text{diferencia entre Ucp} - \text{real} \quad (4.2)$$

Figura 4-9: Valores de diferencia por tipo de día

RNA: 74671	Ucp: 76845
Días ['LUNES', 'MARTES', 'MIERCOLES', 'JUEVES', 'VIERNES', 'SABADO', 'DOMINGO', 'LF', 'MADLF', 'SAALF', 'DOALF']	
Silvcas [9785, 8360, 11002, 10276, 9915, 6076, 7822, 2438, 3295, 2785, 2917]	
RealUCP [9900, 12868, 12965, 10144, 9606, 5802, 6868, 1845, 2682, 1631, 2534]	

4.8 Minería de texto

En el momento las redes sociales son tendencia, ya que la mayoría de la gente registra sus opiniones y comentarios. Estas fuentes de información son utilizadas para analizar los datos, promocionar productos o realizar estudios de mercado (Urrutia). Para el desarrollo del proyecto nos enfocamos en 2 Twitter que nos proporcionan características cualitativas y cuantitativas que aportan para el análisis de las demandas reales en determinados lapsos de tiempo. Para realizar la minería de datos a estos Twitter se seguirá un proceso que está conformado por tres etapas (Morales, 2014):

- **Determinación de los objetivos:** Se delimitan los objetivos que se desean conseguir.
- **Preprocesamiento de los datos:** Se realiza un proceso de selección, limpieza, enriquecimiento, reducción y transformación de las bases de datos.
- **Análisis de los resultados:** Verifica si los resultados obtenidos son coherentes.

En la actualidad Twitter ha colocado a la disposición del público APIs gratuitos que han propiciado la creación de software de terceros, permitiendo de esta manera conectarse y manejar datos de la aplicación. Para el desarrollo del proyecto utilizaremos la librería Tweepy de Python para acceder a Twitter. El API de Twitter limita el acceso a la aplicación entre 150 o 350 solicitudes por hora dependiendo si registramos o no la aplicación en el apartado de desarrolladores de Twitter (Ortiz Folgan, 2015).

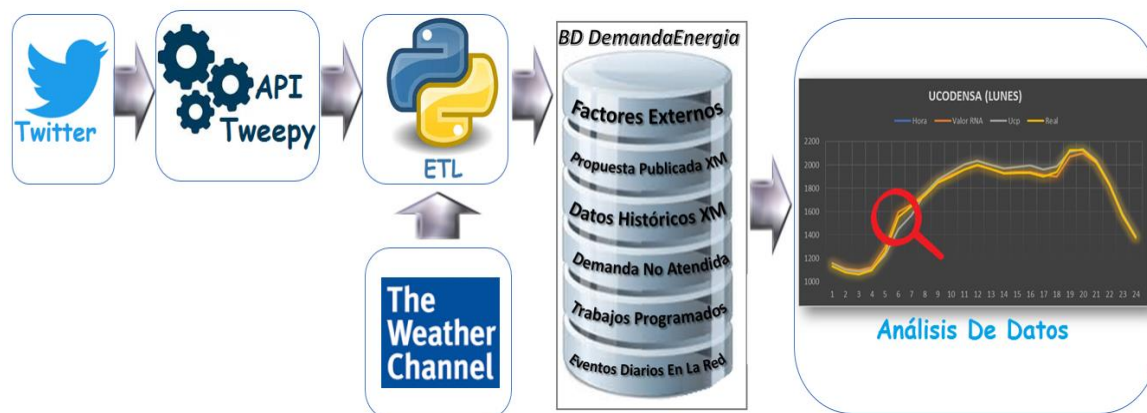
Dada la cobertura que tiene Codensa nos enfocaremos en Bogotá ya que la mayor carga instalada se encuentra en la ciudad. Para ello el desarrollo del proyecto tendrá como fuentes de datos los nombres de usuario “SectorMovilidad”, “IDEAMColombia”, “ELTIEMPO”, ya que después de observar el comportamiento y analizar los datos de estos dos usuarios podemos recuperar información que está relacionada con el comportamiento de la demanda de energía.

Con el API ya instalado y configurado, se procede a realizar las peticiones a Twitter, por medio de las librerías de comunicación del api con Python “import tweepy”. La cual extraerá la información de la publicación de twitter en una estructura, donde solo capturan los datos de los siguientes campos para ser almacenados:

- Id
- created_at
- full_text
- in_reply_to_status_id

El modelo diseñado para minería de Texto para extracción y análisis de la red social Twitter es el siguiente:

Figura 4-10: Modelo de ETL y análisis de datos para minería de texto



Nombre de la fuente: Elaborado por Javier H. Velasco

El clima es uno de los factores externos que afecta la demanda de energía, pero en el momento no existen fuentes de datos históricos de clima que almacenen por hora y que la información suministrada sea confiable, por este motivo el modelo de predicción no puede tener la variable clima dentro de las entradas de entrenamiento, con el objetivo de análisis de las demandas se programó una aplicación en python que descargue de un proveedor de pronósticos del tiempo en línea las características del clima y la temperatura (**Tabla 4-5**).

Tabla 4-5: Registro por rango de minutos de temperatura y clima

FECHA	CIUDAD	HORA ACTUALIZACION	TEMPERATURA	CLIMA
20/09/2018	Bogotá - Bogotá	actualizado a las 13:07 COT	13°	Soleado
20/09/2018	Bogotá - Bogotá	actualizado a las 13:17 COT	13°	Parcialmente nublado
20/09/2018	Bogotá - Bogotá	actualizado a las 13:27 COT	13°	Soleado
20/09/2018	Bogotá - Bogotá	actualizado a las 13:37 COT	13°	Soleado
20/09/2018	Bogotá - Bogotá	actualizado a las 13:47 COT	13°	Soleado
20/09/2018	Bogotá - Bogotá	actualizado a las 13:57 COT	13°	Soleado
20/09/2018	Bogotá - Bogotá	actualizado a las 14:07 COT	13°	Soleado
20/09/2018	Bogotá - Bogotá	actualizado a las 14:20 COT	13°	Soleado
20/09/2018	Bogotá - Bogotá	actualizado a las 14:34 COT	13°	Soleado
20/09/2018	Bogotá - Bogotá	actualizado a las 14:44 COT	13°	Soleado
20/09/2018	Bogotá - Bogotá	actualizado a las 14:54 COT	13°	Soleado
20/09/2018	Bogotá - Bogotá	actualizado a las 15:04 COT	12°	Parcialmente nublado
20/09/2018	Bogotá - Bogotá	actualizado a las 15:13 COT	12°	Parcialmente nublado
20/09/2018	Bogotá - Bogotá	actualizado a las 15:23 COT	12°	Parcialmente nublado
20/09/2018	Bogotá - Bogotá	actualizado a las 15:33 COT	12°	Parcialmente nublado
20/09/2018	Bogotá - Bogotá	actualizado a las 15:34 COT	12°	Parcialmente nublado
20/09/2018	Bogotá - Bogotá	actualizado a las 15:46 COT	12°	Parcialmente nublado
20/09/2018	Bogotá - Bogotá	actualizado a las 15:47 COT	11°	Parcialmente nublado

Como resultado de la investigación se creó el modelo de predicción de demanda de energía eléctrica para UCP's en Colombia bajo el método de Stacking y Minería de texto llamado **SILVCAS**. Ver **Figura 4-11**.

Figura 4-11: Modelo de predicción de demanda de energía - SILVCAS

Nombre de la fuente: Elaborado por Javier H. Velasco

5. Resultados

Los resultados obtenidos de la validación cruzada y el entrenamiento es:

Entrenamiento:

costo: 1616.18091 valor: 0.05997 iteración: 2016 función: 2192
gradiente: 2192

Para calcular si el modelo SILVCAS o la UCP tiene menor valores de diferencia, se inició con los datos de prueba que corresponden a 1848 registros (77 tipos de días ver **Tabla 5-1**).

Tabla 5-1: Cantidad de muestras por tipo de día

TIPO DE DIA	CANTIDAD DE LA MUESTRA
LUNES	9
MARTES	9
MIERCOLES	11
JUEVES	11
VIERNES	11
SABADO	9
DOMINGO	9
LF	2
MADLF	2
SAALF	2
DOALF	2
TOTAL	77

Después de ser tratados por la red neuronal cada una de las semanas, se obtuvieron la predicción y se confrontaron contra los valores de la UCP, luego se procede a sumar los

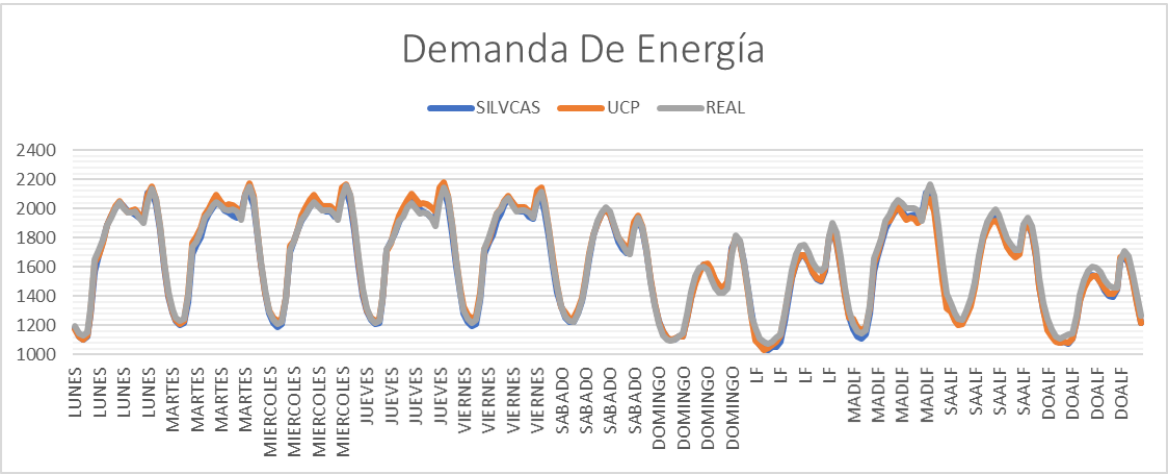
valores absolutos de las diferencias de cada hora para cada tipo de día y como resultado se obtienen los datos que se observan en la **Tabla 5-2**:

Tabla 5-2: Predicción del modelo SILVCAS vs UCP Codensa

TIPO DE DIA	PREDICCIÓN		VALOR MENOR
	MODELO SILVCAS	UCP	
LUNES	8408	8368	UCP
MARTES	7908	11213	SILVCAS
MIERCOLES	9336	10450	SILVCAS
JUEVES	9653	8277	UCP
VIERNES	9735	8122	UCP
SABADO	4452	4635	SILVCAS
DOMINGO	6230	5512	UCP
LF	2894	1845	UCP
MADLF	3345	2682	UCP
SAALF	2231	3125	SILVCAS
DOALF	2148	2815	SILVCAS
TOTAL	66340	67044	

En la anterior tabla se obtuvieron en general que, de los 11 tipos de días analizados, 5 de los 11, el modelo SILVCAS obtuvo una diferencia menor en la suma total de los tipos de días con un valor de 704 MWh. Para realizar un análisis más detallado se tomará 24 horas correspondiente a un tipo de día (ver **Figura 5-1**).

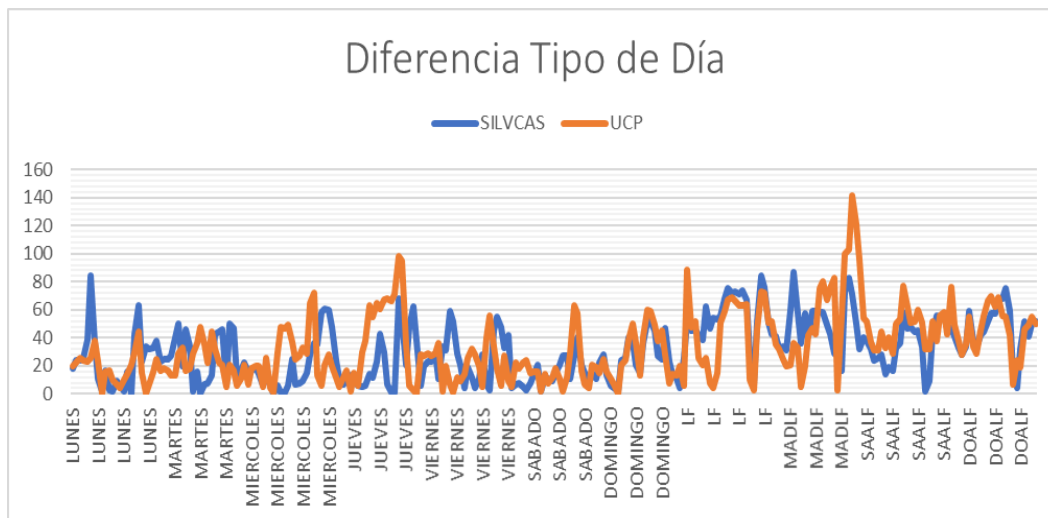
Figura 5-1: Demanda de energía



Nombre de la fuente: Elaborado por Javier H. Velasco

En la **Figura 5-2** la gráfica muestra la diferencia que existe con la demanda real, entre el modelo SILVCAS y la UCP. Para tener un buen ajuste la gráfica debe tender al valor cero y cada pico que se observa por día debe ser analizado.

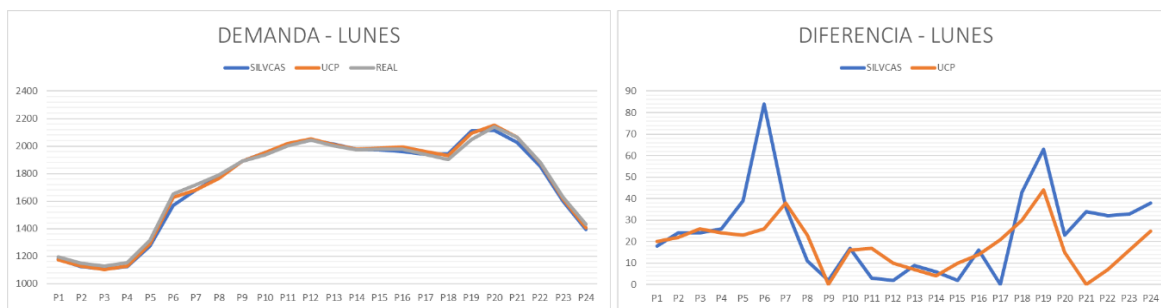
Figura 5-2: Diferencia por tipo de día



Nombre de la fuente: Elaborado por Javier H. Velasco

En la **Figura 5-3** correspondiente al día lunes el valor de la diferencia de la demanda de energía diaria es de 585 Mw-día para el modelo SILVCAS y de 438 Mw-día para la UCP, teniendo una mejor predicción por parte de la UCP. De la gráfica podemos deducir que las horas P16 y P19 tiene un desfase alto que hace que la diferencia sea mayor a la de la UCP.

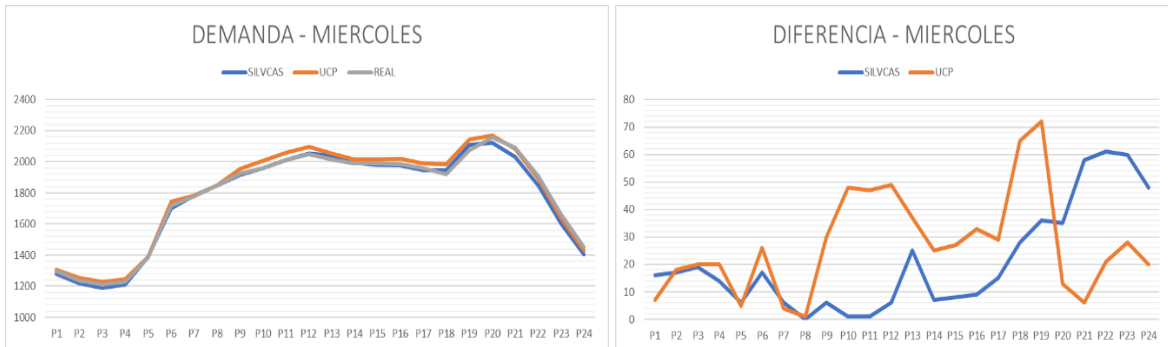
Figura 5-3: Demanda y diferencia día lunes



Nombre de la fuente: Elaborado por Javier H. Velasco

En la **Figura 5-4** correspondiente al día miércoles el valor de la diferencia de la demanda de energía diaria es de 499 Mw-día para el modelo SILVCAS y de 651 Mw-día para la UCP, teniendo una mejor predicción por parte del modelo SILVCAS. De la gráfica podemos deducir que en el modelo SILVCAS las horas de la noche la demanda de energía es atípica por lo que el sistema no logra ajustarse y tiene una tendencia hacia arriba.

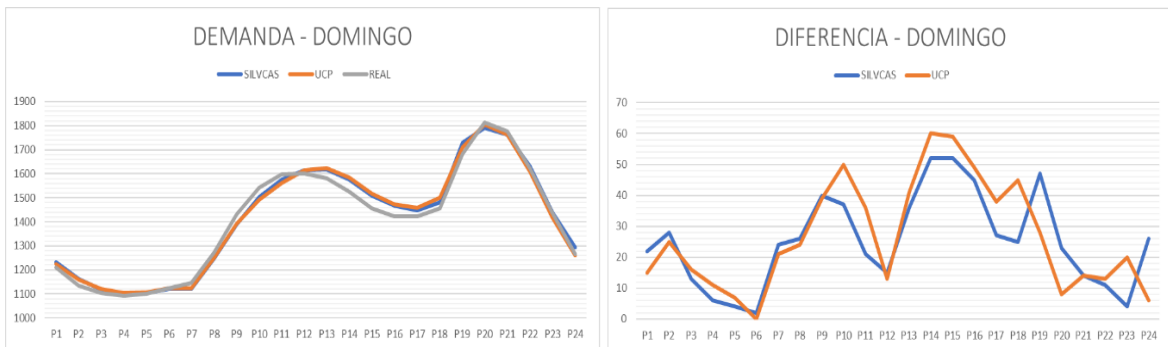
Figura 5-4: Demanda y diferencia día miércoles



Nombre de la fuente: Elaborado por Javier H. Velasco

En la **Figura 5-5** correspondiente al día domingo el valor de la diferencia de la demanda de energía diaria es de 600 Mw-día para el modelo SILVCAS y de 638 Mw-día para la UCP, teniendo una mejor predicción por parte del modelo SILVCAS. De la gráfica podemos deducir que las dos predicciones en las horas de la mañana y la tarde el desfase es alto, se tiene la hipótesis que es consecuencia a factores externos que desplazaron la curva de demanda de energía.

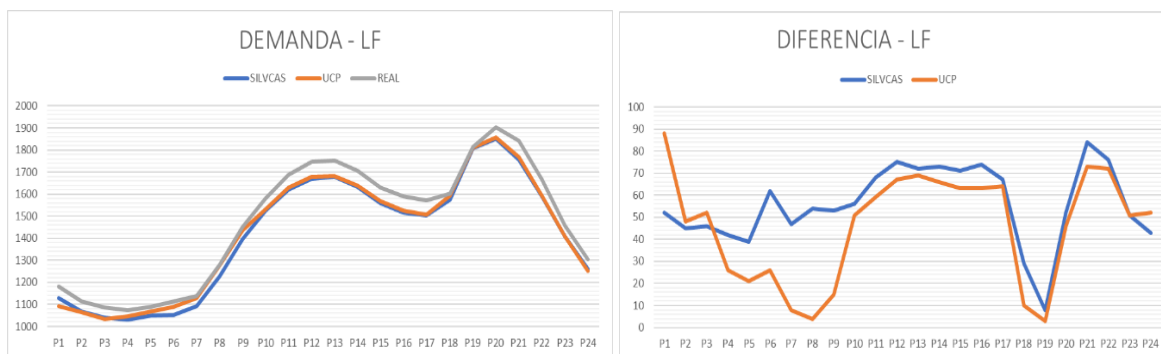
Figura 5-5: Demanda y diferencia día domingo



Nombre de la fuente: Elaborado por Javier H. Velasco

En la **Figura 5-6** correspondiente al día lunes festivo (LF) el valor de la diferencia de la demanda de energía diaria es de 1339 Mw-día para el modelo SILVCAS y de 1097 Mw-día para la UCP, teniendo una mejor predicción por parte de la UCP. De la gráfica podemos deducir que en la franja horaria entre las P10 y P17 las dos predicciones tienen un desfase similar y el comportamiento atípico puede ser a factores externos

Figura 5-6: Demanda y diferencia día lunes festivo

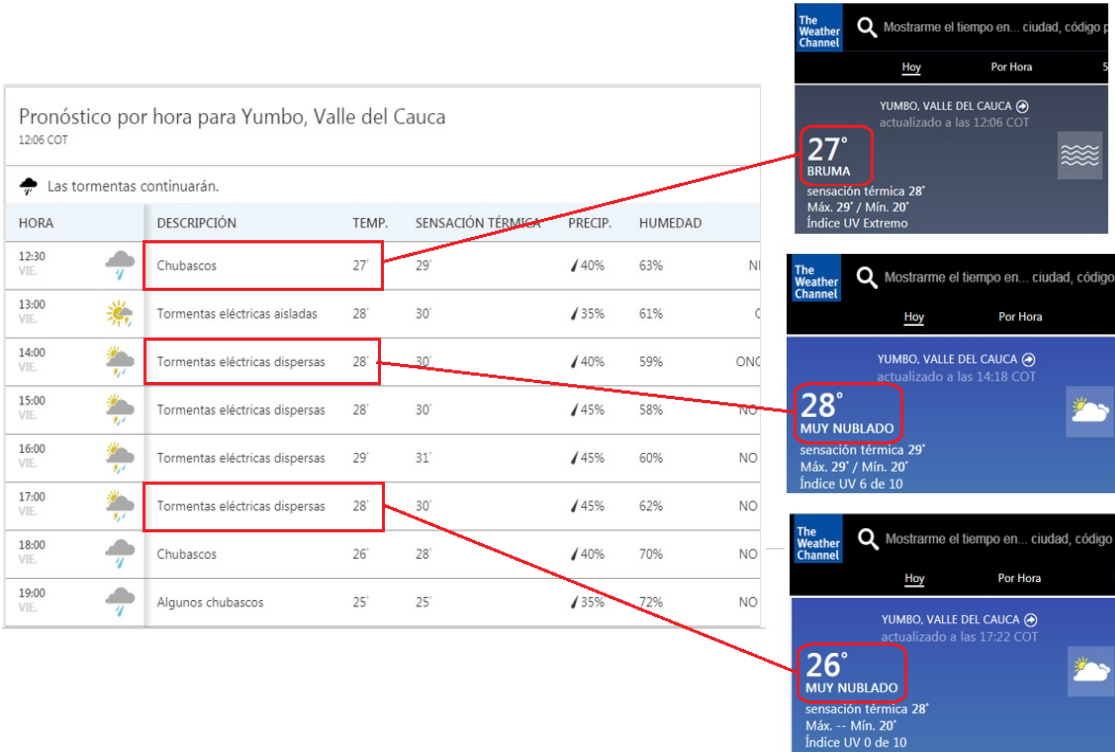


Nombre de la fuente: Elaborado por Javier H. Velasco

5.1 Extracción de información de redes sociales

La Selección del proveedor de pronósticos del tiempo en línea se realizó por medio pruebas y analizando el tiempo de actualización del mismo y comparando contra las predicciones realizadas por el mismo proveedor. El proveedor de pronóstico con un buen porcentaje de predicción fue Weather.com por lo que se selección y se realizaron algunas pruebas, como se observa en la **Figura 5-7** que se realizaron validaciones de la predicción entre las 12 y las 19 horas, como resultado obtuvimos que la predicción del clima para ese día no fue exactamente como lo pronosticaba el proveedor, pero el sistema entre 10 a 13 minutos actualiza su sistema. Debido a este hallazgo no es conveniente agregar la variable de clima a la red neuronal artificial, ya que el clima es muy difícil de predecir y haría que nuestro sistema tuviera un error mayor.

Figura 5-7: Datos del clima



Nombre de la fuente: Elaborado por Javier H. Velasco

En internet existen varios proveedores que pronostican el clima. Entre ellos tenemos el Instituto de Hidrología, Meteorología y Estudios Ambientales de Colombia, conocido como el IDEAM que en su sección de Datos Abiertos provee los datos climáticos de las ciudades de Colombia. Para validar los datos de este proveedor se confrontó los datos descargados de esa fecha contra la información del proveedor Weather.com la cual se desarrolló un código que constantemente está conectado a la página de Weather y almacena los datos en un archivo excel para proceder a almacenarlo en la Base de Datos.

En la **Tabla 5-3** tenemos los datos abiertos del IDEAM para el 20 de septiembre de 2018 entre las horas de 11 hasta 18 horas, en la **Tabla 5-4** tenemos los datos descargados de la página Weather.com los cuales tienen la fecha, hora de actualización, la temperatura y el clima. Como resultado de la comparación obtuvimos que el IDEAM tiene temperaturas más altas y el pronóstico no es acertado, en los datos de Weather.com hay más datos y la temperatura como el pronóstico son actualizados aproximadamente entre 10 a 12 minutos.

Debido a las ventajas que tiene el modelo de extraer la información y ser almacenada en la Base de Datos se puede utilizar para analizar en las curvas de demanda pasadas.

Tabla 5-3: Datos del clima (Datos abiertos del IDEAM)

DATOS ABIERTOS IDEAM			
Fecha	Hora	Temperatura	Pronóstico
20/09/2018	11:00	20.7	Parcialmente nublado
20/09/2018	12:00	21.7	Parcialmente nublado
20/09/2018	13:00	22.6	Parcialmente nublado
20/09/2018	14:00	22	Parcialmente nublado
20/09/2018	15:00	21.4	Parcialmente nublado
20/09/2018	16:00	20.7	Parcialmente nublado
20/09/2018	17:00	18	Parcialmente nublado
20/09/2018	18:00	15.3	Parcialmente nublado

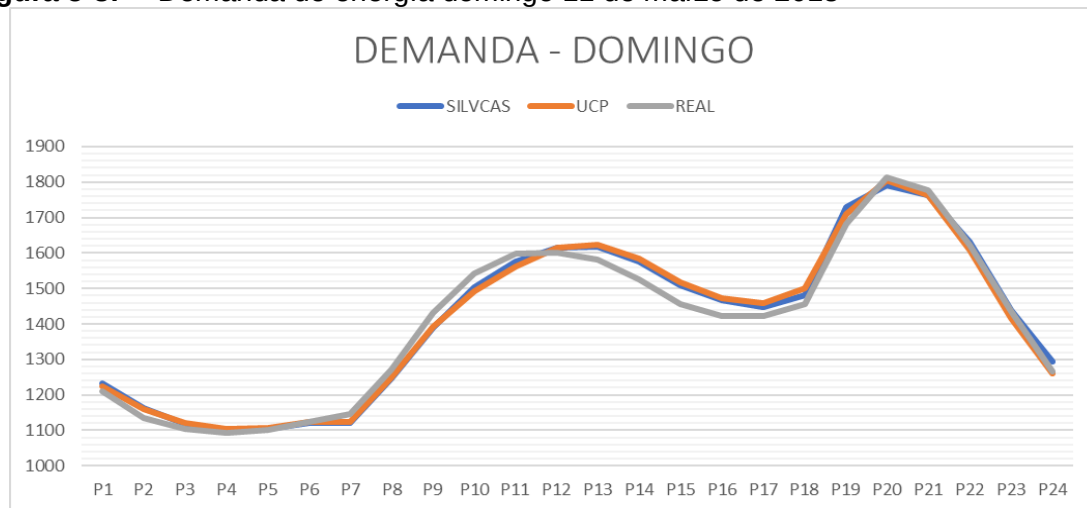
Tabla 5-4: Datos del clima (Weather.com)

DATOS WEATHER.COM				
Fecha	Actualizacion	Hora	Temperatura	Pronostico
20/09/2018	actualizado a las 12:08 COT	12:00	12°	Parcialmente nublado
20/09/2018	actualizado a las 12:21 COT	12:00	12°	Parcialmente nublado
20/09/2018	actualizado a las 12:30 COT	12:00	12°	Parcialmente nublado
20/09/2018	actualizado a las 12:41 COT	12:00	12°	Parcialmente nublado
20/09/2018	actualizado a las 12:54 COT	12:00	12°	Parcialmente nublado
20/09/2018	actualizado a las 13:07 COT	13:00	13°	Soleado
20/09/2018	actualizado a las 13:17 COT	13:00	13°	Parcialmente nublado
20/09/2018	actualizado a las 13:27 COT	13:00	13°	Soleado
20/09/2018	actualizado a las 13:37 COT	13:00	13°	Soleado
20/09/2018	actualizado a las 13:47 COT	13:00	13°	Soleado
20/09/2018	actualizado a las 13:57 COT	13:00	13°	Soleado
20/09/2018	actualizado a las 14:07 COT	14:00	13°	Soleado
20/09/2018	actualizado a las 14:20 COT	14:00	13°	Soleado
20/09/2018	actualizado a las 14:34 COT	14:00	13°	Soleado
20/09/2018	actualizado a las 14:44 COT	14:00	13°	Soleado
20/09/2018	actualizado a las 14:54 COT	14:00	13°	Soleado
20/09/2018	actualizado a las 15:04 COT	15:00	12°	Parcialmente nublado
20/09/2018	actualizado a las 15:13 COT	15:00	12°	Parcialmente nublado
20/09/2018	actualizado a las 15:23 COT	15:00	12°	Parcialmente nublado
20/09/2018	actualizado a las 15:33 COT	15:00	12°	Parcialmente nublado
20/09/2018	actualizado a las 15:34 COT	15:00	12°	Parcialmente nublado
20/09/2018	actualizado a las 15:46 COT	15:00	12°	Parcialmente nublado
20/09/2018	actualizado a las 15:47 COT	15:00	11°	Parcialmente nublado
20/09/2018	actualizado a las 15:57 COT	15:00	11°	Parcialmente nublado
20/09/2018	actualizado a las 15:59 COT	15:00	11°	Parcialmente nublado

5.2 Análisis de datos

Con la información obtenida en la fase de prueba, procedemos a realizar el análisis al día domingo que corresponde al 11 de marzo de 2018 (ver **Figura 5-8**). Al revisar las curvas del modelo SILVCAS, de la UCP y la REAL, se observa que la curva de color gris(Real) tiene un desfase entre las horas de P9 hasta P18.

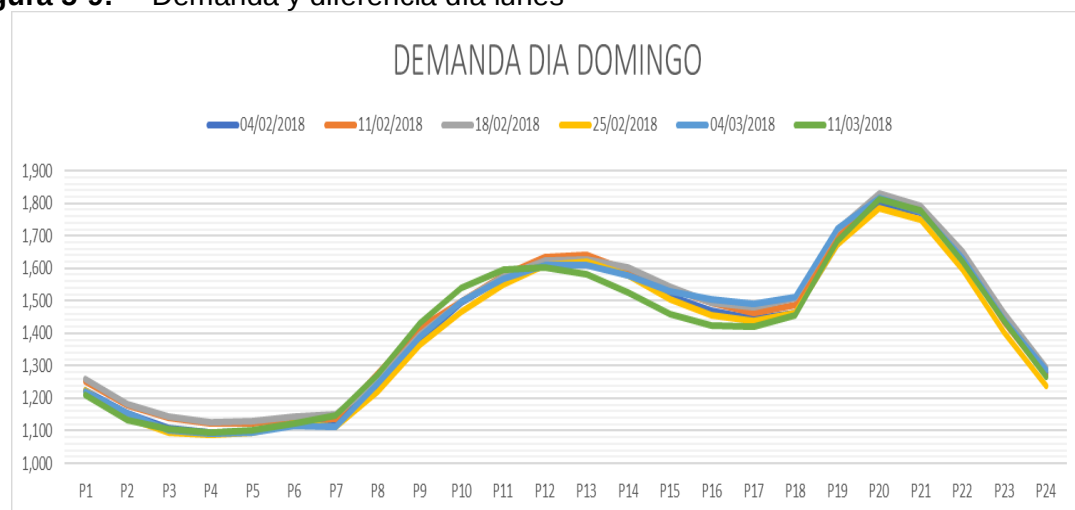
Figura 5-8: Demanda de energía domingo 11 de marzo de 2018



Nombre de la fuente: Elaborado por Javier H. Velasco

Para determinar si es un valor típico de la curva REAL, realizaremos el gráfico con los cinco domingos anteriores al 11 de marzo de 2018 (ver **Figura 5-9**).

Figura 5-9: Demanda y diferencia día lunes



Nombre de la fuente: Elaborado por Javier H. Velasco

Como podemos observar en la Figura 50 la curva verde (11/03/2018) se encuentra desfasada por lo que este valor no es típico, por lo que se procede a analizar las posibles causas de estos valores.

Las redes sociales son una gran fuente de información por lo que realizaremos la consulta a los datos almacenados para la fecha, en este caso el 11 de marzo de 2018. Como resultado de la búsqueda tenemos que ese día se llevaron a cabo las elecciones de Congreso (ver **Figura 5-10** y **Figura 5-11**), por lo que esto puede ser un factor que influyó en que la curva REAL se desfasaba.

Figura 5-10: Publicación del IDEAM - Elecciones al congreso 11 de marzo 2018

IDEAMColombia @IDEAMColombia

Mar10/2018 [#Atención](#) Este es el pronóstico del tiempo para las principales ciudades del país, correspondiente al día de las elecciones legislativas del 11 de marzo.

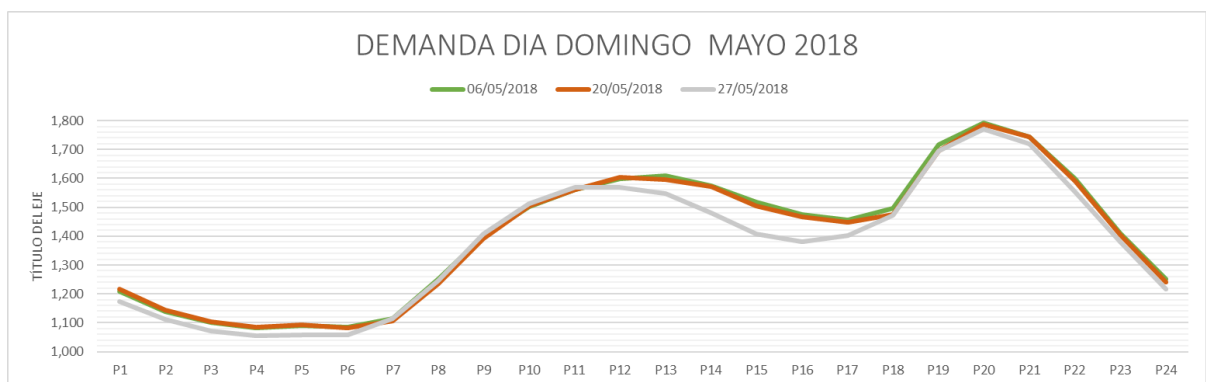
PRONÓSTICO DEL TIEMPO ELECCIONES MARZO 2018		
CIUDADES CAPITALES DEL ORIENTE DEL PAÍS		
TIPO	NUBOSIDAD	PROBABILIDAD DE LLUVIA
Mañana	Cielo parcialmente Nublado	Sin probabilidad
Tarde	Cielo mayormente nublado	Lloviznas

Nombre de la fuente: Twitter del @IDEAMColombia

Figura 5-11: Publicación de SectorMovilidad - elecciones del 11 de marzo de 2018

Nombre de la fuente: Twitter de @SectorMovilidad

Para confirmar que la curva fue afectada por un factor externo como son las elecciones, se realizó el análisis a las elecciones de presidente que se llevaron a cabo el día domingo 27 de mayo de 2018.

Figura 5-12: Demanda día domingo de los meses de febrero y marzo

Nombre de la fuente: Elaborado por Javier H. Velasco

Como se observa en la **Figura 5-12**, la curva de color gris (27/05/2018) se encuentra desfasada de las curvas de los dos domingos anteriores, por lo que concluimos que las elecciones son factores que influyen en la demanda de energía.

Con minería de texto confirmamos el evento (ver **Figura 5-13**)

Figura 5-13: Publicación de SectorMovilidad - elecciones del 27 de mayo de 2018



Nombre de la fuente: Twitter de @SectorMovilidad

5.3 Ejecución del modelo en otra UCP

Se ejecutó el modelo con los datos de la UCP Uantioquia, para observar el comportamiento del modelo con otra UCP, a continuación, los resultados obtenidos.

Se extraen desde el portal de XM los archivos desde julio de 2016 hasta abril de 2018 correspondientes a Uantioquia, y se procede al cargue en la Base de Datos y se ejecuta para las mismas 57 semanas entre el 1 de enero de 2017 hasta abril de 2018 que cumplen los criterios.

El cálculo de los 12 métodos para las 57 semanas es de 9576 Registros y serán guardados en un archivo llamado "Datos.csv"

El archivo Datos.csv tiene los datos de entrenamiento(Train) y de prueba(Test), por lo que, del total de datos obtenidos de los 12 clasificadores, dividiremos en 80% para entrenamiento y 20% para prueba, los datos de entrenamiento y de prueba deben ser normalizados para realizar el aprendizaje a la red neuronal artificial.

Se ejecutó la fase del modelo de la red neuronal y se inició con la división del archivo de los 12 clasificadores en 70 % para entrenamiento y 30 % para validación y prueba. Como costo de la función de error y el número de iteraciones tenemos:

ENTRENAMIENTO: 70 % y PRUEBA: 30 %

Capas Ocultas: 3 - 3 - 3

costo: 1252.20071 valor: 0.0937 iteración: 2964 función: 3309 gradiente: 3297

En el proceso de validación del modelo, se obtuvieron buenos resultados, ya que en cada iteración el indicador de error siempre fue menor.

#PLIEGUE: 1 - MAPE SILVCAS: 4.53 - MAPE UCP: 6.048

#PLIEGUE: 2 - MAPE SILVCAS: 4.421 - MAPE UCP: 6.093

#PLIEGUE: 3 - MAPE SILVCAS: 4.237 - MAPE UCP: 5.639

#PLIEGUE: 4 - MAPE SILVCAS: 4.538 - MAPE UCP: 6.078

#PLIEGUE: 5 - MAPE SILVCAS: 4.449 - MAPE UCP: 6.262

#PLIEGUE: 6 - MAPE SILVCAS: 4.298 - MAPE UCP: 5.8

#PLIEGUE: 7 - MAPE SILVCAS: 4.313 - MAPE UCP: 6.038

Después de ser tratados por la red neuronal obtuvieron la predicción y se confrontaron contra los valores de la UCP, luego se procede a sumar los valores absolutos de las diferencias de cada hora para cada tipo de día y como resultado se obtienen los datos que se observan en la **Tabla 5-5**:

Tabla 5-5: Valores de predicciones de SILVCAS y Uantioquia

TIPO DE DIA	PREDICCIÓN		VALOR MENOR
	MODELO SILVCAS	UCP	
LUNES	12897	18022	SILVCAS
MARTES	11583	17658	SILVCAS
MIERCOLES	16358	20112	SILVCAS
JUEVES	14794	20557	SILVCAS
VIERNES	17047	21670	SILVCAS
SABADO	11793	15548	SILVCAS
DOMINGO	11157	14526	SILVCAS
LF	3695	5016	SILVCAS
MADLF	3569	4363	SILVCAS
SAALF	8599	4790	UCP
DOALF	8088	4820	UCP
TOTAL	119580	147082	

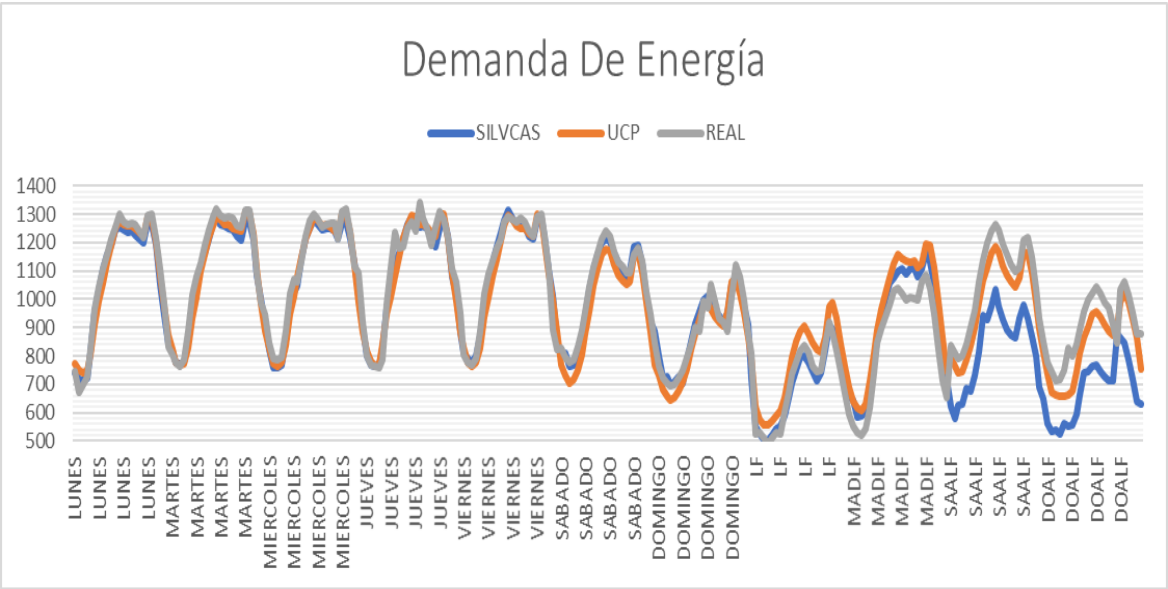
En la **Tabla 5-6** se tienen los datos correspondientes al último día (tipo de día) con el valor acumulado de las diferencias por MWh en cada día.

Tabla 5-6: Valor acumulado por diferencias en tipos de día

TIPO DE DIA	MODELO SILVCAS	UANTIOQUIA	DIFERENCIA (Mw-día)
LUNES	529	579	50
MARTES	609	628	19
MIERCOLES	610	573	37
JUEVES	773	896	123
VIERNES	403	710	307
SABADO	375	1352	977
DOMINGO	727	943	216
LF	513	1702	1189
MALF	1815	2567	752
SAALF	5540	1397	4143
DOALF	5341	1728	3613

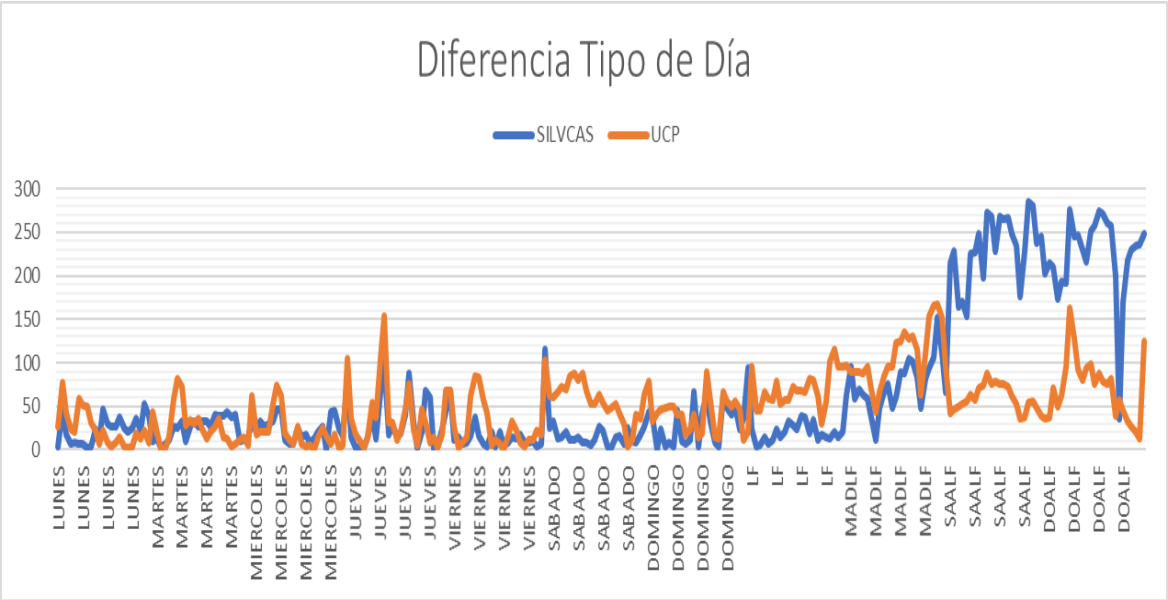
Como se observa en la **Figura 5-14** el modelo tiene buena predicción, de los 11 tipos de días 8 los predice bien el modelo.

Figura 5-14: Demanda de energía

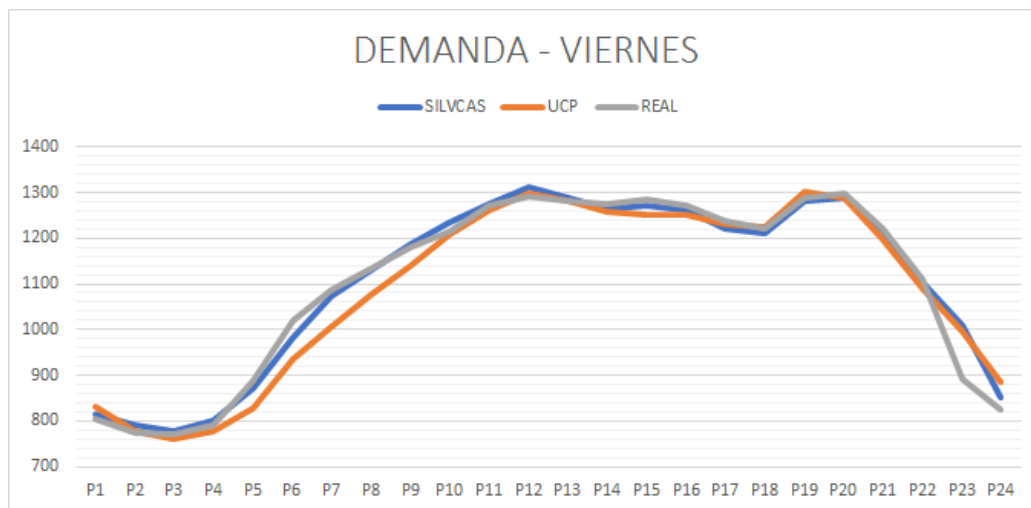


Nombre de la fuente: Elaborado por Javier H. Velasco

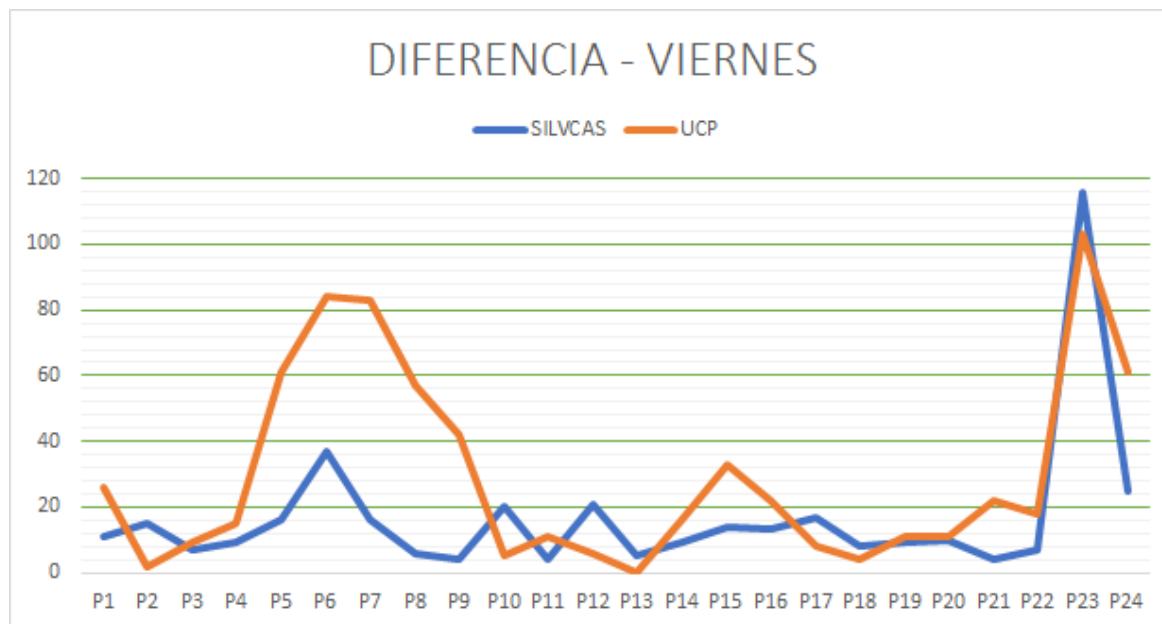
Figura 5-15: Diferencia por tipo de día



Nombre de la fuente: Elaborado por Javier H. Velasco

Figura 5-16: Demanda de energía día viernes

Nombre de la fuente: Elaborado por Javier H. Velasco

Figura 5-17: Diferencia entre SILVCAS y la UCP para el día viernes

Nombre de la fuente: Elaborado por Javier H. Velasco

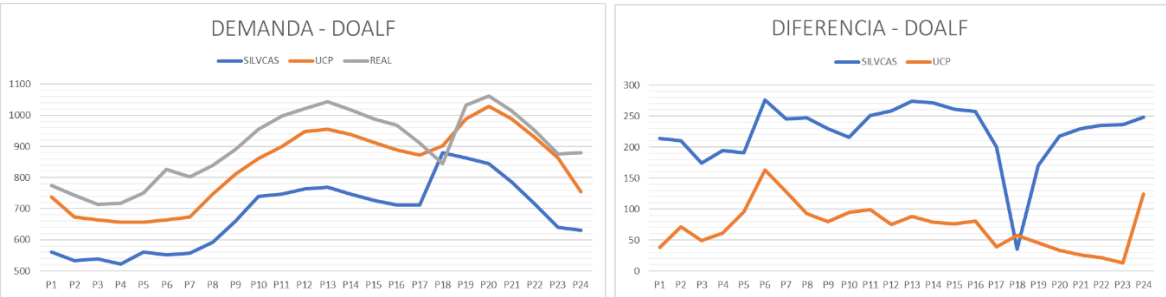
Al analizar la **Figura 5-17** se observa un pico en la hora 23 donde el modelo SILVCAS tiene un valor de 116 MWh de diferencia Mientras que la UCP un valor de 103 MWh, al analizar la figura 60, en la hora 23 la demanda real presenta una disminución posiblemente a un valor atípico y causante de la diferencia. Si se revisa la curva en su totalidad el modelo

SILVCAS se ajusta mejor a la demanda real, por lo que sus valores de diferencia se encuentran en un 12.5 en promedio, caso contrario la UCP que tiene valores promedio de 26.4, para el cálculo se excluyó la hora 23. Como resultado de la **Tabla 5-7** podemos concluir que al aplicar el modelo SILVAS se mejora la predicción y disminuye un 403 Media la generación planificada para la UCP Uantioquia.

Tabla 5-7: Valor acumulado de diferencias para el día viernes de Uantioquia

PREDICCIÓN	VALOR (Mw-día)
Modelo SILVCAS	403
UCP UAntioquia	710
Diferencia	307

Figura 5-18: Demanda y diferencia para el día DOALF de Uantioquia



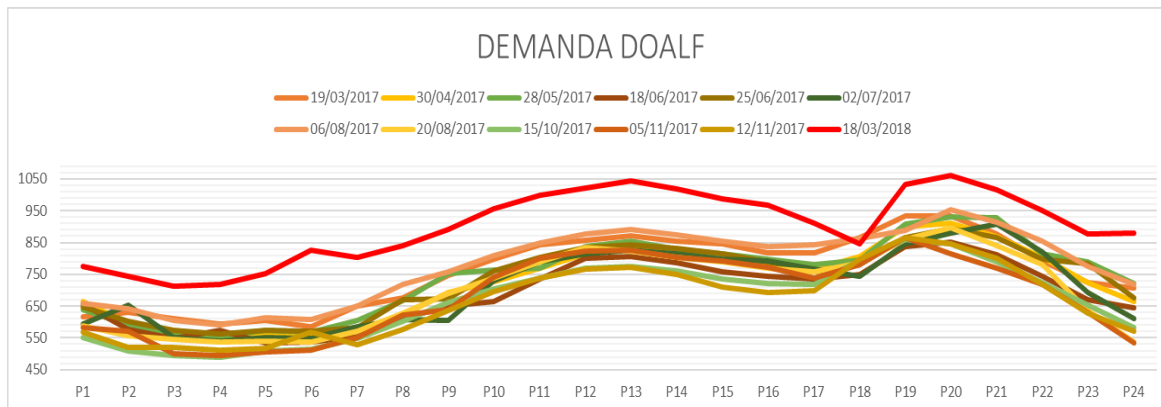
Nombre de la fuente: Elaborado por Javier H. Velasco

Los valores de diferencia entre lo real y pronosticado para el día DOALF se encuentra en la **Tabla 5-8:**

Tabla 5-8: Valor acumulado de diferencias para el día doalf de Uantioquia

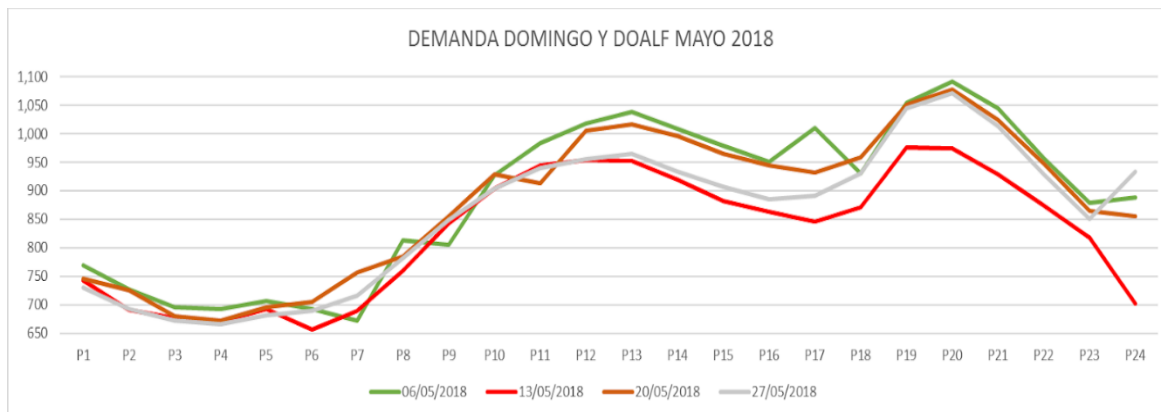
PREDICCIÓN	VALOR (Mw-día)
Modelo SILVCAS	5341
UCP UAntioquia	1728
Diferencia	3613

Debido a las diferencias de los días SAALF Y DOALF, se procedió a realizar el análisis al día DOALF y se encontró lo siguiente:

Figura 5-19: Demanda del tipo de día DOALF en Uantioquia

Nombre de la fuente: Elaborado por Javier H. Velasco

Como se observa en la **Figura 5-19** el día 18 de marzo de 2018 (fecha a predecir), presenta valores atípicos con valores de demanda demasiado altos, por consiguiente, los valores que tiene el modelo se encuentran dentro de los valores con que la red fue entrenada. Según la **Figura 5-20** que corresponde al mes de mayo de 2018, los días DOALF están presentando demandas altas casi similares a un DOMINGO normal.

Figura 5-20: Demanda de energía de la UCP Uantioquia para los días DOMMINGO y DOALF en mayo 2018

Nombre de la fuente: Elaborado por Javier H. Velasco

El entrenamiento se realizó en un computador de escritorio con Intel Core i3 y 4 GB de RAM, corriendo en Windows 8. El código fuente se encuentra disponible en github en: https://github.com/javiervelasco/RNA_Backpropagation

5.4 Prototipo

Como prototipo se desarrolló una aplicación web llamada WEBSILVCAS con la siguiente estructura:

- ❖ Framework: Django.
- ❖ Lógica: Python 3.6.2
- ❖ motor de base de datos: MYSQL.

WEBSILVCAS tiene el siguiente mapa de sitio:

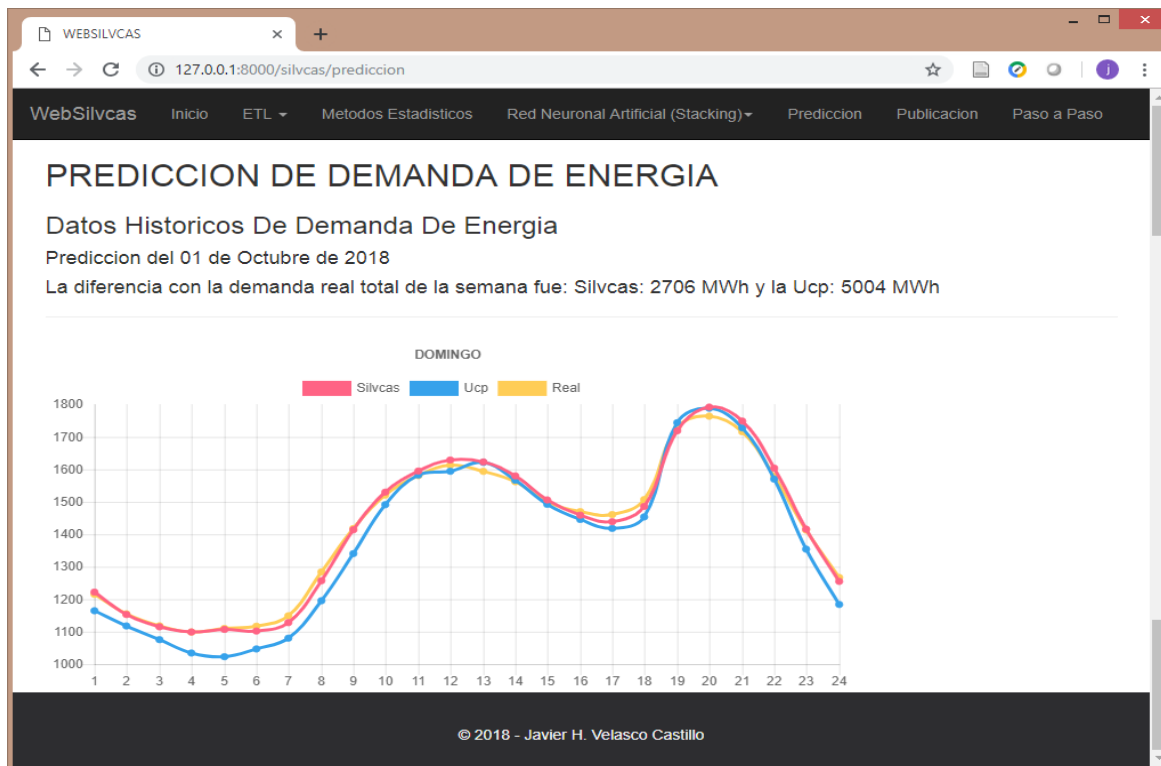
- Inicio
- ETL
- Métodos Estadísticos
- Red Neuronal Artificial (Stacking)
- Predicción
- Publicación
- Paso a Paso

Figura 5-21: Interfaz de inicio de la aplicación web.



Para validar si el modelo SILVCAS es mejor que el utilizado por Codensa, se realizara en el aplicativo WEBSILVCAS la predicción de la primera semana del mes de octubre con el entrenamiento de los últimos datos del mes de septiembre.

Figura 5-22: Interfaz de predicción de la primera semana de octubre de 2018



Después de ejecutar la aplicación WBSILVCAS para la primera semana de octubre de 2018 como se observa en la **Figura 5-22** obtenemos que Codensa presenta mayor diferencia después de aplicar las ecuaciones (4.1) y (4.2). En la **Tabla 5-9** se tiene los valores totales de diferencia de la primera semana de octubre de 2018.

Tabla 5-9: Diferencia en MWh en la primera semana de octubre de 2018

Predicción 01/10/2018	
Silvcas	2706 MWh
Ucp Codensa	5004 Mwh
Diferencia (Semana)	2298 MWh

En el portal de XM <https://www.xm.com.co/Paginas/Mercado-de-energia/precio-de-bolsa-y-escasez.aspx> se obtienen los precios de bolsa con que se negocia diariamente (ver **Figura 5-23**) y además en el portal se puede obtener el precio de bolsa por hora (ver **Figura 5-24**)

Figura 5-23: Precio de bolsa y escasez

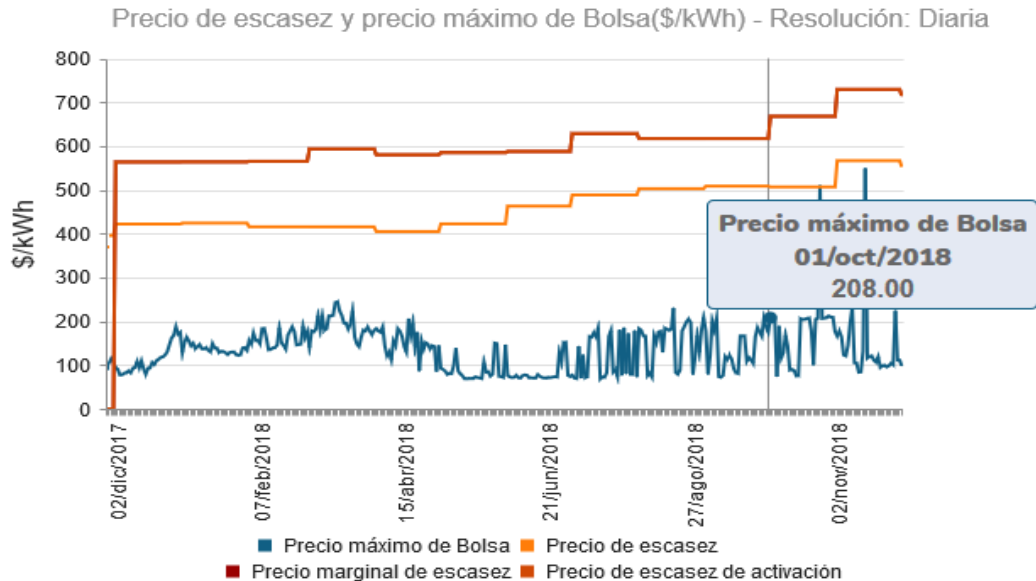


Figura 5-24: Precio de bolsa por hora de la primera semana de octubre de 2018

Precio Bolsa Nacional (\$/kWh) 2018						
Fecha	0	1	22	23	Maximo	Promedio
2018-10-01	118.00	118.00	177.99	139.00	208.00	162.17
2018-10-02	160.44	160.44	161.94	160.44	209.94	173.12
2018-10-03	94.19	94.19	154.19	134.19	209.19	156.25
2018-10-04	72.81	72.81	72.81	72.81	72.81	72.81
2018-10-05	79.36	79.36	140.60	110.10	190.09	129.86
2018-10-06	106.74	77.35	115.74	106.74	115.97	105.03
2018-10-07	87.58	87.58	128.18	87.58	128.18	98.71
Promedio					162.03	128.28

Teniendo los valores del precio del kWh para la primera semana de octubre de 2018 se procede a calcular el valor promedio en bolsa por la diferencia que se tiene entre el modelo SILVCAS y la predicción de Codensa. Como resultado se obtiene que si se utilizara el

modelo SILVCAS el valor aproximado de ahorro para el operador sería de \$294.787.440 (ver **Tabla 5-10**).

Tabla 5-10: Cálculo del valor en pesos de la diferencia

Predicción 01/10/2018	
Diferencia	2298 MWh
En KWh	2.298.000 KWh
Valor promedio Bolsa	\$128.28
Valor Semana	\$294.787.440

5.5 Discusión

Debido a que en el proyecto se desarrolló con la información pública que se encuentra en el portal de XM y que solo es publicada los diez primeros días hábiles del mes y solo en este punto es posible descargar los datos para realizar la actualización al sistema. Se concluyó que los resultados obtenidos son buenos, dado que tenemos predicciones con valores menores a la UCP de comparación que para este caso es Ucodensa. Para obtener mejores resultados, se realizó una simulación con datos más recientes, es decir con datos de la semana anterior, lo que dio como resultado una disminución mayor en el indicador de error y se puede constatar en la **Tabla 5-11**:

Tabla 5-11: Predicciones con 4 y 1 semana de datos históricos.

TIPO DE DIA	PREDICCIÓN CON 4 SEMANAS		PREDICCIÓN CON 1 SEMANA	
	MODELO SILVCAS	UCP	MODELO SILVCAS	UCP
LUNES	8408	8368	7366	8368
MARTES	7908	11213	6939	11213
MIÉRCOLES	9336	10450	8571	10450
JUEVES	9653	8277	9096	8277
VIERNES	9735	8122	7963	8122
SABADO	4452	4635	3876	4635
DOMINGO	6230	5512	5175	5512
LF	2894	1845	2020	1845
MADLF	3345	2682	3444	2682
SAALF	2231	3125	1594	3125
DOALF	2148	2815	1978	2815
TOTAL	66340	67044	58022	67044

Si realizamos la resta de los valores totales del cuadro anterior obtenemos que el
Valor UCP – Valor Modelo SILVCAS (1 y 4 semanas): Diferencia en Energía (MWh)

Como resultados obtenemos que con datos del mes anterior la diferencia de energía disminuiría es de 291MWh equivalente a una mejora de 3.12%

67044(UCP) – 66340(MODELO): 704 MWh

Con datos de la semana anterior el modelo mejora considerablemente y pasa de 291MWh a 9022MWh es decir el modelo mejora de un 3.12% a un 13.46%

67044(UCP) – 58022(MODELO): 9022 MWh

En los anexos se encuentran los resultados obtenidos de las 4 semanas anteriores y la semana anterior a partir del modelo desarrollado.

6. Conclusiones y trabajos futuros

La presente investigación tuvo como objetivo desarrollar un modelo de predicción de demanda de energía eléctrica que fusionó los métodos estadísticos con el modelo stacking teniendo como base una red neuronal artificial que tiene como entrada los métodos estadísticos y como salida un valor de predicción más ajustado al valor de la demanda real, además se acoplo un modelo de minería de texto que apoya en la identificación de posibles valores atípicos que se presentan en los valores de la demanda real histórica.

A continuación, se analizan los resultados obtenidos a partir de conclusiones y los posibles trabajos futuros que pueden continuar desarrollándose como resultados de esta investigación.

6.1 Conclusiones

En el momento existen varios métodos usados para realizar el pronóstico de la demanda de energía, entre ellos tenemos los métodos estadísticos, que tienen como ventajas que son de fácil implementación, no consumen recursos y son fáciles de entender y analizar, estos métodos estadísticos han tenido un aceptable resultado al realizar las predicciones. También entre otros métodos utilizados tenemos las redes neuronales artificiales, las cuales consumen más recursos que otros métodos, pero tiene la ventaja encontrar nuevos patrones, partiendo de la información que se provea en la etapa de entrenamiento. Todos estos métodos siempre son supervisados por un analista que determinara si el resultado se encuentra dentro de los parámetros de los valores de la semana a pronosticar.

El modelo SILVCAS es el resultado de la fusión de los métodos estadísticos y las redes neuronales para realizar las predicciones, la cual mostró un mejor desempeño, ya que al momento de realizar las predicciones a los 77 tipos de días que fueron utilizados para la

prueba se disminuyó el MAPE como se observó en la sección de resultados, donde las validaciones cruzadas realizadas a la UCP's Ucodensa, obtuvieron sobre el modelo SILVCAS valores por debajo de las predicciones de las UCP's. Por lo que esta investigación dio respuesta a la pregunta de investigación planteada en la definición del problema que orientó este trabajo y además alcanzó parte de su objetivo general, al disminuir el indicador de error con que los administradores de las UCP's en Colombia son evaluados.

La otra parte del objetivo general contemplada es que el modelo pueda ser utilizada por otra UCP de Colombia, por lo que dentro de la sección de resultados se ejecutó el modelo a la UCP Uantioquia, obteniendo mejores resultados, ya que al proceder a analizar las validaciones cruzadas de los datos de prueba, el MAPE del modelo SILVCAS presentó valores por debajo de las predicciones de la UCP, es decir que al promediar la diferencia entre el valor MAPE del modelo y de la UCP el valor los 7 Pliegos en promedio se encontró en una disminución de 1.6% del MAPE.

La investigación demostró que la temperatura al ser una variable que afecta la demanda de energía, no puede ser tomada como variable de entrada a la red neuronal artificial, debido a que los pronósticos de clima realizados por los diferentes proveedores, no almacenan la información histórica por hora y actualizada, además no los pronósticos revisados el porcentaje de acierto fue bajo, causa principal que afectaría la predicción negativamente, generando ruido y haciendo que las curvas presentan desfases.

La minería de texto en la investigación fue enfocada como herramienta de análisis ya que con esta técnica logramos extraer información relevante de la red social Twitter, ya que se encontraron patrones de uso de palabras y usuarios claves que ayudaron a analizar las posibles causas que generan comportamientos atípicos en las curvas de las demandas de energía.

El realizar la simulación a la UCP Ucodensa con la información de la semana anterior, es decir que la ventana de tiempo de datos fuentes se desplazó de 4 semanas a 1 semana, obteniendo como resultado un valor de 13% en la disminución de la diferencia entre el valor pronosticado por la UCP y el valor de la demanda real. Por lo que podemos concluir

que al tener valores actualizados el modelo tiene un mejor ajuste y el valor MAPE puede reducir considerablemente.

La recuperación de información desde Twitter no se incluyó durante el entrenamiento del modelo, se usó como fuente para explicar datos atípicos, se mostró que la mayoría de fallas grandes en las predicciones están relacionadas con eventos que son reportados en Twitter.

Los datos del clima y precipitaciones no se pudieron obtener de ninguna fuente confiable y las variaciones entre reportes eran muy grandes para considerar que dicha información podría ayudar en las otras partes del proceso.

6.2 Trabajos futuros

Como continuación de este trabajo de tesis se presentan algunos trabajos futuros que pueden dar continuidad a esta investigación y que por estar fuera del alcance de esta investigación no fueron tratados con suficiente profundidad. Pero que serán base para complementar y mejorar el modelo:

Aplicar el modelo con información de los últimos días previo a la realización del reporte del pronóstico de la demanda. Esta información deberá ser suministrada por el administrador de la UCP a partir de los datos extraídos del SCADA y de los medidores principales y de respaldos que se encuentran instalados en las fronteras comerciales.

Implementar el modelo de predicción para la fase de minería de texto, ya que durante la investigación se identificaron que existen días atípicos similares, como los días de elecciones presidenciales y de Congreso en Colombia

Ampliar el alcance del modelo para los otros tipos de días que no se tuvieron en cuenta para el desarrollo de esta investigación.

Bibliografía

- Agarwal, A. S. (2016). Opinion mining of news headlines using SentiWordNet. Symposium on Colossal Data Analysis and Networking (CDAN). Obtenido de <https://doi.org/10.1109/CDAN.2016.7570949>
- Álvarez, P. L. (s.f.). Estimación De La Velocidad Del Viento Mediante Redes Neuronales.
- Arencibia R, J. R. (2012). Predicción de resistencia a fármacos del VIH utilizando multclasificadores. Revista Cubana de Ciencias Informáticas. Obtenido de [https://rcci.uci.cu/?journal=rcci&page=article&op=view&path\[\]=179](https://rcci.uci.cu/?journal=rcci&page=article&op=view&path[]=179)
- Barrientos, A. F. (s.f.). Un modelo spline para el pronóstico de la demanda de energía eléctrica. Revista Colombiana de Estadística.
- Bertona, L. F. (2005). Entrenamiento de redes neuronales basado en algoritmos evolutivos. Facultad de Ingeniería, Universidad de Buenos Aires. Obtenido de <http://laboratorios.fi.uba.ar/lsi/bertona-tesisingenieriainformatica.pdf>
- Bolaño Cuesta, N. K. (2015). náalisis de la cadena de abastecimiento de la energía eléctrica en Colombia (bachelorThesis). Universidad del Rosario. Obtenido de <http://repository.urosario.edu.co/handle/10336/11646>
- Cárdenas, E. G. (s.f.). *Evaluación de los modelos de predicción de fracaso empresarial en el sector manufacturero colombiano en los años 2010-2014*.
- Carrasquilla-Batista, A. C.-R.-M.-E.-B. (2016). Regresión lineal simple y múltiple: aplicación en la predicción de variables naturales relacionadas con el crecimiento microalgal. Revista Tecnología en Marcha. Obtenido de <https://doi.org/10.18845/tm.v29i8.2983>
- Codensa. (julio de 2016). *Metodología pronóstico oficial de demanda* . Obtenido de https://www.cno.org.co/sites/default/files/documentos/noticias/Pronostico_Codensa_22072016_v1.pdf
- Electricaribe. (julio de 2016). *Elaboración de pronósticos*. Obtenido de <https://www.cno.org.co/sites/default/files/documentos/noticias/Presentaci%C3%B3n%20Pronosticos%20ECA%202016.pdf>
- Estrella, M. S. (s.f.). Un imperativo moral: la enseñanza de la estadística no puede dejarse al azar. 2014. I Encuentro Colombiano de educación. Obtenido de <http://funes.uniandes.edu.co/6530/1/Estrella2014UnimperativoECEE.pdf>
- Fernández de Mesa Bustelo, M. (julio de 2016). Análisis y mejora de la predicción de la demanda eléctrica en periodos de alto ECM. Obtenido de <http://oa.upm.es/43165/>

- Fernández, D. &. (s.f.). estadística descriptiva y análisis de datos. Facultad Regional Mendoza, Universidad Tecnológica Nacional. Obtenido de <http://www1.frm.utn.edu.ar/estadistica/documentos/ed&ad.pdf>
- Franco, C. J. (2012). Micromundo para simular un mercado eléctrico de corto plazo. Cuadernos de Economía, 31(58), 229-256.
- García Ubilla, A. D. (s.f.). Análisis de datos y búsqueda de patrones en aplicaciones médicas. Facultad de ciencias físicas y matemáticas,.
- GEO_tutoriales. (s.f.). *Error Porcentual Absoluto Medio (MAPE) en un Pronóstico de Demanda*. Obtenido de 2015: <https://www.gestiondeoperaciones.net/proyeccion-de-demanda/error-porcentual->
- Hernández, M. &. (2018). *Redes neuronales para clasificación: una aplicación al caso de riesgos laborales en Colombia*. Obtenido de <http://repository.javeriana.edu.co/handle/10554/37845>
- Lang, J.-P. S. (s.f.). *Validación cruzada*. Obtenido de <https://ingsistemastelesup.files.wordpress.com/2017/03/validacion-cruzada.pdf>
- Lozano, I. F. (2017). Aplicación de la red neuronal artificial feedforward backpropagation para la predicción de demanda de energía eléctrica en la empresa eléctrica riobamba s.a. Escuela superior politécnica de chimborazo. Obtenido de <http://dspace.esPOCH.edu.ec/bitstream/123456789/7606/1/20T00925.PDF>
- Madero, R. A. (2006). Estadística para pediatras (I): conceptos generales. Madrid, España: Sección de Bioestadística. Hospital Universitario la Paz. Obtenido de http://appswl.elsevier.es/watermark/ctl_servlet?_f=10&pidet_articulo=80000216&pidet_usuario=0&pcontactid=&pidet_revista=51&ty=89&accion=L&origen=apcc&continuada&web=www.apcontinuada.com&lan=es&fichero=v4n5a216pdf001.pdf&anuncioPdf=ERROR_publi_pdf
- Mariñelarena-Dondena, L. E. (2017). Extracción de conocimiento con técnicas de minería de textos aplicadas a la psicología. Revista Argentina de Ciencias del Comportamiento.
- Maté, A. P. (2016). A hybrid integrated architecture for energy consumption prediction. Future Generation Computer Systems. Obtenido de <https://doi.org/10.1016/j.future.2016.03.020>
- Matich, D. J. (2001). Redes Neuronales: Conceptos Básicos y Aplicaciones. Facultad Regional Rosario, Universidad Tecnológica Nacional. Obtenido de

https://www.frro.utn.edu.ar/repositorio/catedras/quimica/5_anio/orientadora1/monografias/matich-redesneuronales.pdf

- Mir-Montserrat, D. (2015). Analítica de datos de Twitter. Universitat Autònoma de Barcelona. Obtenido de https://ddd.uab.cat/pub/trerecpro/2015/hdl_2072_252402/PFC_DaniMirMontserrat.pdf
- Morales, S. M. (2014). Análisis de Información Proveniente de Redes Sociales como Twitter. Universidad Carlos III. Obtenido de https://e-archivo.uc3m.es/bitstream/handle/10016/22918/TFG_Soledad_Mart%C3%ADn_Morales.pdf
- Ortiz Folgan, Á. L. (2015). Estudio de tendencias diarias en Twitter. Universidad Complutense Madrid. Obtenido de <https://eprints.ucm.es/33137/1/Estudio%20de%20Tendencias%20Diarias%20en%20Twitter.pdf>
- Ortiz, D. A. (2014). Aplicación de redes neuronales artificiales en el pronóstico de la demanda eléctrica a corto plazo en el SNI. Universidad Politécnica Salesiana. Obtenido de <http://dspace.ups.edu.ec/bitstream/123456789/6672/6/UPS-KT00835.pdf>
- Pérez-Planells, L. D.-C. (2015). Análisis de métodos de validación cruzada para la obtención robusta. Revista de Teledetección. Obtenido de <https://doi.org/10.4995/raet.2015.4153>
- Prieto, A. B. (2015). Desarrollo de un modelo de caracterización financiera del sector comercio en Ecuador, período 2002-2012. Universidad Andina. Obtenido de <http://repositorio.uasb.edu.ec/bitstream/10644/4521/1/T1642-MFGR-Tulcanaza-Desarrollo.pdf>
- Ramón Sierra, R. (s.f.). Capítulo 3 – Pronósticos. Universidad de las Américas Puebla. Obtenido de http://catarina.udlap.mx/u_dl_a/tales/documentos/lii/sierra_r_r/capitulo3.pdf
- Ren, Y. &. (2017). Sentiment analysis of Internet performance data. IEEE 3rd Information Technology and Mechatronics Engineering. Obtenido de <https://doi.org/10.1109/ITOEC.2017.8122373>
- Reyes S., J. M. (2014). Desarrollo de un meta-modelo de predicción para el diagnóstico de infección cervical por Chlamydia trachomatis en mujeres de Colombia. Universidad

- Nacional de Colombia. Obtenido de Repositorio institucional UN (masters). : <http://bdigital.unal.edu.co/46126/>
- Rodríguez, O. (s.f.). *Validación Cruzada y Remuestreo*. Obtenido de http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Presentaci
- Rueda, V. M. (2001). Avances recientes en la predicción de la demanda de electricidad usando modelos no lineales. *DYNA*, 78(167), 36-43.
- Santillán Arroyo, M. (s.f.). Unidad IV Compras e Inventarios - Pronóstico. Universidad Autónoma del estado de Hidalgo. Obtenido de http://cvonline.uaeh.edu.mx/Cursos/Lic_virt/Mercadotecnia/DMKT022-2/Unidad%203/32_lec_pronosticos.pdf
- Sarmiento, H. O. (2008). Inteligencia artificial en pronóstico de demanda de energía eléctrica: una aplicación en optimización de recursos energéticos. *Revista Colombiana de Tecnologías de Avanzada*. Obtenido de http://www.unipamplona.edu.co/unipamplona/portallG/home_40/recursos/02_v07_12/revista_12/16112011/v12_15.pdf
- Soler Flores, F. &. (2014). Investigación - Estimación de sucesos raros mediante Redes Bayesianas. *Revista de Investigación "Pensamiento Matemático"*. Obtenido de http://www2.caminos.upm.es/Departamentos/matematicas/revistapm/revista_impr esa/vol_IV_num_1/inv_redes_bayes.pdf
- Tabares Muñoz, J. F. (2014). Comparación de técnicas estadísticas de pronóstico para la demanda de energía eléctrica. *Revista Ingeniería Industrial*. Obtenido de <http://revistas.ubiobio.cl/index.php/RI/article/view/2>
- Urrutia, A. &. (s.f.). *Análisis de Información de Redes Sociales (Twitter)*. Obtenido de http://sedici.unlp.edu.ar/bitstream/handle/10915/61597/Documento_completo.pdf-PDFA.pdf?sequence=1
- Valencia Cárdenas, M. &. (2013). Un modelo dinámico bayesiano para pronóstico de energía diaria. *Revista Ingeniería Industrial*. Obtenido de <https://dialnet.unirioja.es/servlet/articulo?codigo=4734871>
- Vásquez Macedo, R. R. (2013). *Pronóstico de la demanda de energía eléctrica para Bolivia Aplicación de inteligencia artificial*. Obtenido de <http://www.inesad.edu.bo/bcde2013/papers/BCDE2013-100.pdf>
- Vicentin Masaro, J. (2016). *Predicción de precios mediante modelización multivariada de series de tiempo. Una aplicación al sector lácteo argentino*. Obtenido de <https://rdu.unc.edu.ar/handle/11086/3391>

Wood, M. (2013). *Making Statistical Methods More Useful: Some Suggestions From a Case Study*. SAGE Open. Obtenido de <https://doi.org/10.1177/2158244013476873>

XM. (2017). *Informe de Operación del SIN y Administración del Mercado 2017*. Obtenido de http://informesanuales.xm.com.co/2017/SitePages/operacion/Informe_Operacion_SIN_2017.pdf