PREDICCIÓN DEL CONSUMO ENERGÉTICO RESIDENCIAL
IPD-440 MACHINE LEARNING PROFESOR: WERNER CREIXELL

Dayana Hernández Rodríguez
Department of Electronic Engineering
Universidad Técnica Federico Santa María
Valparaíso, 2019
dayana.hernandez@sansano.usm.cl

Andrea Reales Villalba
Department of Electrical Engineering
Universidad Técnica Federico Santa María
Valparaíso, 2019
andrea.reales@sansano.usm.cl

July 14, 2019

ABSTRACT

The prediction of energy consumption allows efficient use of energy for both the user and the electricity company. A dataset is used that records the energy consumption of several houses in London over a period of two and a half years. The recurrent LSTM network that has shown high performance in the prediction of the time series is used. The project contains 8 tests that are based mainly on the prediction of consumption given climatic conditions and given the previous behavior of the consumption time series.

***K*eywords** LSTM · Time Series · Power Consumption

1 Introducción

En la actualidad se han implemetando mayor cantidad de medidores inteligentes ya que estos son mucho más eficientes tanto para los proveedores de electricidad como para los usuarios, debido a que proporcionan información actualizada y pueden llevar la correcta medición del consumo de energía de una vivienda hacia la compañía de eléctricidad. Debido a la cantidad de datos que se obtienen con la instalación de medidores inteligentes, los investigadores en análisis de datos y machine learning han visto un camino abierto para el análisis de estos y la obtención de resultados que pueden servir para lograr una eficiencia energética tanto para el usuario como para la compañía eléctrica. Se han realizado varios trabajos de investigación relacionados con la predicción del consumo energético siendo este un tema de gran relevancia, por lo que es abordado en el presente proyecto. El consumo energético es una serie de tiempo que puede considerarse altamente no lineal y no estacionaria. Las variaciones en los flujos de energía que hacen sumamente difícil hacer una predicción vienen dadas por las condiciones climáticas, el uso de aparatos y equipos, los patrones de ocupación y las preferencias de comodidad de los ocupantes, la eficiencia de los componentes como la iluminación, los aparatos y los sistemas HVAC (calefacción, ventilación y aire acondicionado). Las redes recurrentes LSTM (*Long Short Term Memory networks*) han demostrado tener un alto desempeño en la predicción de series de tiempo por lo que es la herramienta usada en el proyecto para realizar la predicción del consumo energético siendo este nuestro principal objetivo. El dataset usado registra el consumo energético de aproximadamente 5567 casas en Londres por un periódo de dos años y medio. Durante el proyecto fueron fueron realizadas 8 pruebas que son detalladas en el desarrollo del artículo, que se basan fundamentalmente en la predicción del consumo energético dado condiciones climáticas y dado el comportamiento anterior de la serie de tiempo. El artículo está dividido en una sección dedicada a hacer un acercamiento de las redes LSTM, otra sección dedicada a la descripción de las series de tiempo, una sección donde es analizado el dataset usado y por último una sección donde son presentadas las pruebas realizadas y los resultados de estas.

2 Redes Recurrentes LSTM

La red neuronal recurrente (Recurrent neural network, RNN) es un tipo de modelo de aprendizaje profundo (Deep Learning) que se utiliza principalmente para el análisis de datos secuenciales (predicción de datos de series de tiempo). La redes recurrentes poseen un lazo de retroalimentación desde la salida a la entrada, lo que permite una persistencia de los datos, lo cual se muestra en la Figura 1.

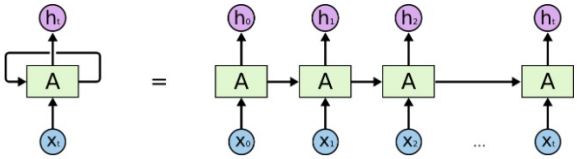


Figure 1: Red Neuronal Recurrente

Las RNN se utilizan en diferentes áreas de aplicación: modelo del lenguaje, traducción automática neuronal, generación de música, predicción de series de tiempo, predicción financiera, etc. Las RNN en la práctica han demostrado problemas con las dependencias largas por lo que se comienzan a usar las LSTM (*Long Short Term Memory networks*), las cuales son capaces de seguir las largas dependencias sin el problema del *vanishing gradient*. Las LSTM son un tipo especial de RNN diseñada especialmente para seguir las largas dependencia de los datos de entrada. La LSTM poseen la misma estructura de cadena de módulos repetitivos que las RNN pero poseen una estructura más complicada en cada uno de los módulos. La Figura 2 muestra la estructura de una LSTM .

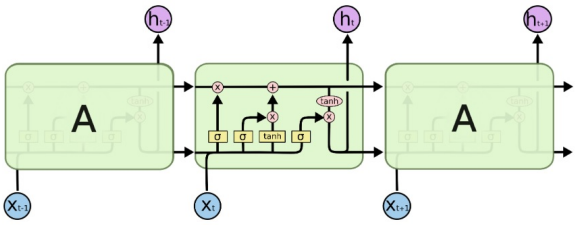


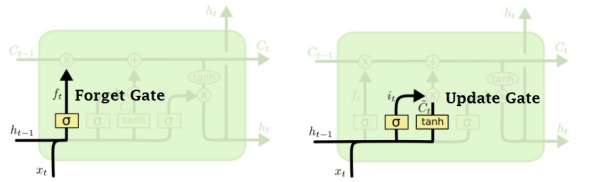
Figure 2: LSTM

Las LSTM siguen un funcionamiento que está basado en 4 etapas(Figuras 3 y Figura 4) y es conformado por 3 Gate o Niveles fundamentales: Nivel de Borrado (*Forget Gate Layer*), Nivel de Actualización (*Update Gate Layer*) y el Nivel de Salida (*Output Gate Layer*). La primera etapa en el funcionamiento de LSTM consiste en decidir qué información va a ser eliminada del estado de la celda. La decisión es tomada usando la función sigmoide en el Nivel de Borrado (*Forget Gate Layer*). La salida genera un número entre 0 y 1 para cada número en el estado de celda anterior. Un 1 representa "mantener completamente el estado" mientras que un 0 representa "deshacerse completamente del estado". La siguiente etapa consiste en decidir qué información nueva se va almacenar en el estado de la celda. El Nivel de Actualización (*Update Gate Layer*) tiene dos partes. Primero, una capa sigmoide llamada capa de puerta de entrada que decide que valores serán actualizados. Segundo, una capa tanh crea un vector de nuevos valores candidatos, que pueden agregarse al estado. Luego las dos partes son combinadas para crear una actualización del estado. Una vez que es decidido que debe ser olvidado y que debe ser actualizado entonces se procede a realizar la operación, siendo esta la tercera etapa. Finalmente, se decide que se entrega como salida. La salida se basa en el estado de celda, pero es una versión filtrada de esta. Primero, se ejecuta una capa sigmoide que decide qué partes del estado de la celda irán a la salida. Luego, el estado de la celda pasa a través de una $tanh$ (valores entre 1 y 1) y es multiplicada por la salida de la puerta sigmoide, de modo que solo son salida de la celda las partes que se deciden.

Las LSTM han sido ampliamente usada en predicción de series de tiempo, por lo se decide usar este tipo de red neuronal en el presente proyecto.

3 Predicción de Series de Tiempo

Un dataset normal en machine learning es una conjunto de observaciones, que cuando se realiza una predicción todos los datos son tratados de la misma manera. Sin embargo, las series de tiempo son un dataset diferente debido a que le agrega una relación de tiempo a la conjunto de observaciones. Esta dimensión adicional de tiempo entre los

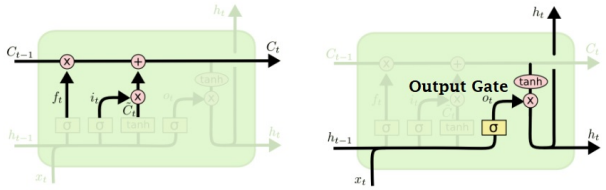


$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

Figure 3: Etapas de Funcionamiento de LSTM



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

$$o_t = \sigma\left(W_o \; [h_{t-1}, x_t] \; + \; b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

Figure 4: Etapas de Funcionamiento de LSTM

datos añade tanto una restricción como una estructura que proporciona una fuente de información adicional. Hacer predicciones sobre el futuro se conoce como hacer una extrapolación en el manejo estadístico clásico de datos de series de tiempo. La predicción de las series de tiempo implica tomar modelos que se ajusten a los datos históricos y usarlos para predecir futuras observaciones. Las redes LSTM pueden ser usadas para diferentes tipos de problemas de predicción de series de tiempo. Existen tres tipos fundamentales de series de tiempo: Univariables, Multivariables y Multi-step, esta última puede venir acompañada por las dos primeras clasificaciones. Las series de tiempo univariables son aquellas que comprenden una sola serie de observaciones, y se requiere un modelo para aprender de la serie de observaciones pasadas para predecir el siguiente valor en la secuencia. Las series de tiempo multivariables es donde existe más de una observación para cada paso de tiempo (*time step*).Existen dos modelos principales de series de tiempo multivariables: *Multiple Input Series* y *Multiple Parallel Series*. El modelo *Multiple Input Series* tiene dos o más series de tiempo de entrada paralelas y una serie de tiempo de salida que depende de las series de tiempo de la entrada. El modelo *Multiple Parallel Series* es aplicado en el caso donde hay múltiples series de tiempo paralelas de entrada y se debe predecir un valor de salida para cada una de esas series de entrada. Las series de tiempo multi-step requieren una predicción de múltiples pasos de tiempo (*time step*) en el futuro, específicamente, estos son problemas donde el horizonte o intervalo de pronóstico es más de un paso de tiempo. Las series de tiempo usadas en el proyecto son univariables y multivariables. Las series de tiempo deben ser ajustadas a la forma de entrada de los modelos LSTM la cual generalmente es de la siguiente forma: [samples, time-steps, features]. Para una mayor profundización en los tipos de series de tiempo y los modelos LSTM que pueden ser usados consultar `https://machinelearningmastery.com/how-to-develop-lstm-models-for-time-series-forecasting/`.

4 Análisis del Dataset Disponible

El análisis del dataset disponible permite hacer una profundización en el comportamiento de cada una de las variables que influyen en el aprendizaje de la red. El dataset usado es "Smart Meter In London", el cuál fue tomado de **Kaggle**, `https://www.kaggle.com/jeanmidev/smart-meters-in-london#acorn_details.csv`. Este dataset contiene el consumo energético registrado por smart meters en Londres desde 11/23/2011 hasta 02/28/2014, al principio solo existen muestras de 11 casas, pero se siguieron instalando más medidores inteligentes en las viviendas hasta llegar a un total de 5.567 casas. El dataset posee varios archivos, pero en el presente proyecto fueron utilizados:

- informations_households.csv: este archivo que contiene toda la información sobre los hogares (ID de la casa, Clasificación socioeconómica (*Acorn*), tarifa) y en qué archivo block.csv.gz se almacenan sus datos.
- halfhourly_dataset.zip: archivo Zip que contiene los archivos de bloque con la medición del medidor inteligente cada media hora. Todos los archivos bloque dentro del zip son unidos en un único archivo nombrado energy1.csv

- daily_dataset.zip: archivo Zip que contiene los archivos de bloque con la información de consumo energético diario (mínimo, máximo, promedio, mediana, suma y estándar). Todos los archivos bloque dentro del zip son unidos en un único archivo nombrado energy.csv
- weather_daily_darksky.csv: archivo que contiene los datos diarios de las variables de clima.
- weather_hourly_darksky.csv: archivo que contiene los datos cada una hora de las variables de clima.

4.1 Clasificación según estudio socio-económico

El archivo **informations_households.csv** posee el registro de la clasificación socioeconómica de cada una de las casas donde es instalado un smart meter para las mediciones. La clasificación socioeconómica se divide en las siguientes categorías.

- Afluentes: son los usuarios residenciales con categoría socio-economica alta.
- Confortable: son los usuarios residenciales con categoría socio-economica media.
- Adversity: son los usuarios residenciales con categoría socio-economica baja.

La Figura 5 muestra el porcentaje asociado a cada clasificación socioeconómica dentro del dataset. Las otras clasificaciones que se pueden apreciar no son detallas por el dataset por lo que no son usadas en el proyecto.

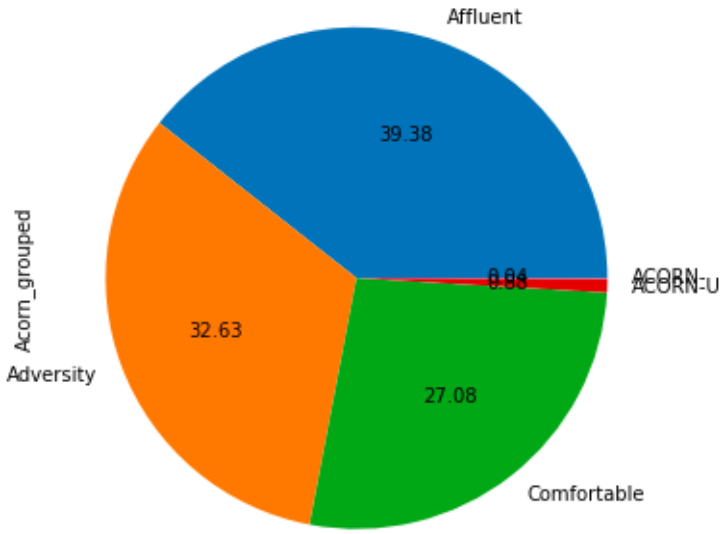


Figure 5: Clasificación según edtudio socio-económico

La Figura 6 muestra el consumo energético según clasificación socio-económica, donde se nota un mayor consumo energético en los Afluentes, y como todas las categorías el consumo energético baja en los meses de verano donde el día es más largo.

4.2 Dataset de consumo de energía

Los archivos de consumo energéticos usados son **halfhourly_dataset.zip** y **daily_dataset.zip**, los cuales permitieron obtener los dataset energy1.csv y energy.csv.

energy.csv

El dataset contiene los datos de consumo de energía diarios medidos por los smart meters de de todas las casas con una resolución diaria. Los parámetros son los siguientes: LCLid, corresponde al número del smart meter, energy_sum, a la suma de la energía por día. La Figura 7 muestra el dataset energy.csv.

El comportamiento del consumo de energía diario de cada casa puede ser apreciado en la Figura 8, lo cual muestra el comportamiento no lineal de la serie de tiempo.

energy1.csv

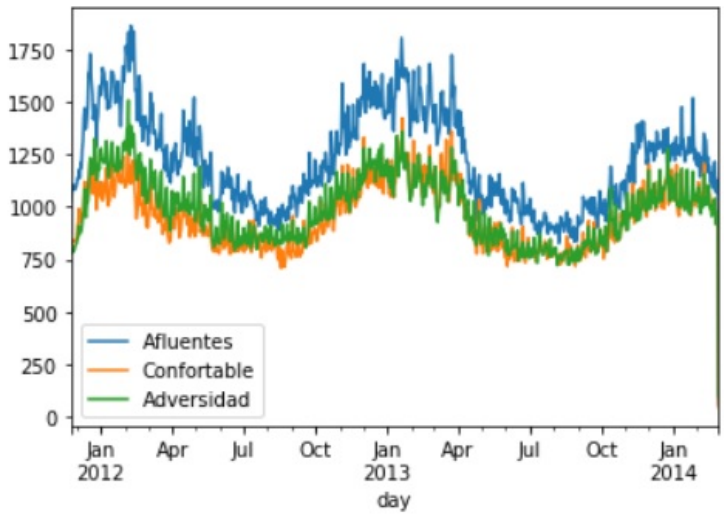


Figure 6: Consumo de energía según clasificación socio-económica

day	LCLid	energy_sum
2012-10-12	MAC000002	7.098
2012-10-13	MAC000002	11.087
2012-10-14	MAC000002	13.223
2012-10-15	MAC000002	10.257
2012-10-16	MAC000002	9.769

Figure 7: Datos diarios del consumo de energía de cada casa

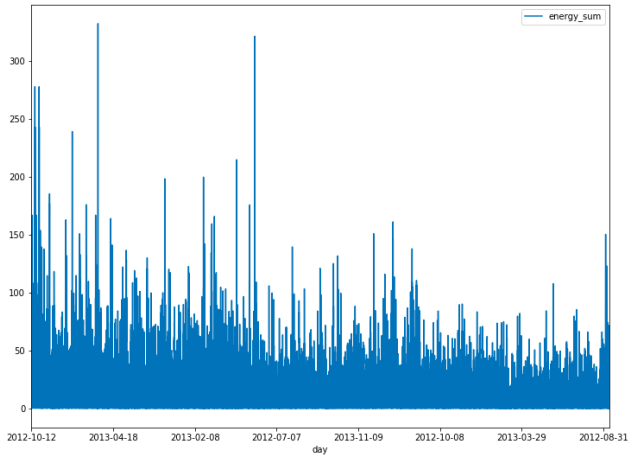


Figure 8: Comportamiento del consumo de energía diario de cada casa

El dataset contiene los datos de consumo de energía con una resolución de media hora medidos por los smart meters de cada casa. Los parámetros son los siguientes: LCLid, corresponde al número del smart meter, tstp, el tiempo cada media hora con su respectiva fecha y energy(kWh/hh) es el consumo de energía. La Figura 9 muestra la forma del dataset energy1.csv.

LCLid	tstp	energy(kWh/hh)
MAC004828	2014-02-27 22:00:00.0000000	0
MAC004828	2014-02-27 22:30:00.0000000	0.001
MAC004828	2014-02-27 23:00:00.0000000	0.047
MAC004828	2014-02-27 23:30:00.0000000	0.008
MAC004828	2014-02-28 00:00:00.0000000	0

Figure 9: Datos cada media hora del consumo de energía de cada casa

4.3 Dataset de clima

Los datos de clima durante el registro del consumo energético del dataset, son dados por el creador de este. Los archivos son **weather_daily_darksky.csv** con resolución diaria y **weather_hourly_darksky.csv** con resolución cada una hora.

weather_daily_darksky.csv

Los datos de clima que utilizados en el proyecto son: temperatureMax, windspeed, humidity, como se aprecia en la Figura 10.

day	temperatureMax	windSpeed	humidity
2011-11-11	11.96	3.88	0.95
2011-12-11	8.59	3.94	0.88
2011-12-27	10.33	3.54	0.74
2011-12-02	8.07	3.00	0.87
2011-12-24	8.22	4.46	0.80

Figure 10: Datos diarios del clima

Los parámetros que presentan mayor influencia sobre la variable objetivo se deben tener en cuenta, para hacer que la red sea capaz de generalizar. El comportamiento de las variables principales de clima durante el intervalo de tiempo recogido por el dataset puede verse en la Figura 11.

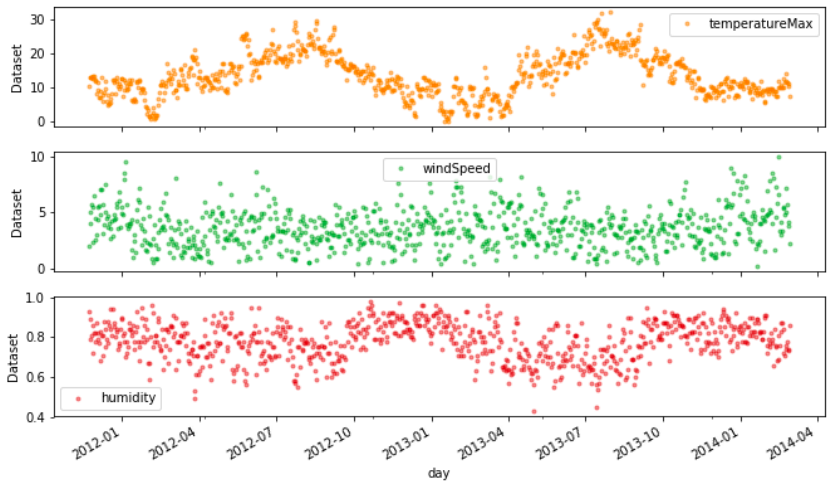


Figure 11: Comportamiento de los datos diarios del clima

weather_hourly_darksky.csv
Los datos de clima que utilizados en el proyecto son: temperatureMax, windspeed, humidity, pero en este caso con una resolución cada una hora como se aprecia en la Figura 12.

time	temperature	windSpeed	humidity
2011-11-11 00:00:00	10.24	2.77	0.91
2011-11-11 01:00:00	9.76	2.95	0.94
2011-11-11 02:00:00	9.46	3.17	0.96
2011-11-11 03:00:00	9.23	3.25	0.96
2011-11-11 04:00:00	9.26	3.70	1.00

Figure 12: Datos cada hora del clima

Matriz de correlación
La matriz de correlación de la Figura 13 indica como se encuentran relacionados el consumo de energía diario, la temperatura, la velocidad del viento y la humedad. Como se aprecia, la temperatura y el consumo energético tienen una relación inversamente proporcional, mientras que las variables de velocidad del viento y la humedad tienen una correlación positiva.

	energy_sum	temperatureMax	windSpeed	humidity
energy_sum	1.000000	-0.390428	0.170150	0.251849
temperatureMax	-0.390428	1.000000	-0.154118	-0.405068
windSpeed	0.170150	-0.154118	1.000000	-0.040131
humidity	0.251849	-0.405068	-0.040131	1.000000

Figure 13: Matriz de correlación del consumo de energía y el clima diario

5 Escenarios implementados

Durante el desarrollo del presente proyecto se realizaron varias pruebas que están divididas en dos grandes grupos la predicción del consumo energético basado en el comportamiento de variables externas(serie de tiempo multivariable), y la predicción del consumo energético basado en su comportamiento pasado (serie de tiempo univariable). La Tabla 1 detalla cada una de las pruebas realizadas. El dataset usado posee varias archivos que fueron combinados en función de la predicción realizada cada prueba.

Pruebas	Nombre de la Prueba	Descripción de la Prueba
Prueba1	Consumo Energético dado Datos de Clima de una casa	Se realiza una predicción del consumo energético dado los datos de clima temperatura, velocidad del viento y humedad, de una casa usando el dataset de resolución de una hora
Prueba2	Consumo Energético agregado dado Datos de Clima resolucion hora	Se realiza una predicción del consumo energético agregado usando el consumo de 11 casas dado los datos de clima: temperatura, velocidad del viento y humedad, usando el dataset de resolución de una hora
Prueba3	Consumo Energético agregado dado Datos de Clima resolución diaria	Se realiza una predicción del consumo energético agregado usando el consumo de 11 casas dado los datos de clima: temperatura, velocidad del viento y humedad, usando el dataset de resolución de un día
Prueba4	Consumo Energético agregado dado consumo energético de 11 casas resolución media hora	Se realiza la predicción del consumo agregado dado el consumo energético de 11 casas usando el dataset con una resolución de media hora
Prueba5	Consumo Energético agregado dado consumo energético de 11 casas resolución hora	Se realiza la predicción del consumo agregado dado el consumo energético de 11 casas usando el dataset con una resolución de una hora
Prueba6	Consumo agregado resolución media hora	Se realiza una predicción del consumo agregado dado el consumo agregado pasado usando el dataset con resolución de media hora
Prueba7	Consumo agregado resolución hora	Se realiza una predicción del consumo agregado dado el consumo agregado pasado usando el dataset con resolución de media hora
Prueba8	Consumo agregado por Acorn resolución diaria	Se realiza una predicción del consumo agregado de cada Acorn dado el consumo agregado pasado de cada Acorn usando el dataset con resolución diaria

Table 1: Pruebas de Predicción de Consumo Energético

5.1 Prueba1: Consumo Energético dado Datos de Clima de una casa

Para la realización de la Prueba 1, se tomó solamente una casa del dataset energy1.csv, la cual contaba con datos del consumo energético con resolución cada media hora durante la mayor parte del tiempo que el dataset cubre, por lo que se podia contar con mayor cantidad de datos. Por otra parte se tomó el dataset de datos de clima con una resolución de una hora, tomando solamente las variables de Temperatura, Velocidad del Viento y Humedad. El dataset conformado para la prueba se muestra en la Figura 14. La relación de las variables de clima y el consumo energético se puede apreciar en la Figura 15, donde se evidencia una relación inversa con la temperatura siendo el consumo menor en los meses donde la temperatura es más alta.

En la prueba se realiza una predicción del consumo energético cada una hora de una casa dado los datos de clima, temperatura, velocidad del viento y humedad, con la resolución de una hora. Debido a que se usan diferentes variables

tstp	energy(kWh/hh)	temperature	windSpeed	humidity
2012-10-13 00:00:00	0.263	8.78	2.28	0.84
2012-10-13 01:00:00	0.275	8.27	1.81	0.87
2012-10-13 02:00:00	0.211	7.87	1.95	0.89
2012-10-13 03:00:00	0.161	7.89	1.83	0.93
2012-10-13 04:00:00	0.167	7.74	1.38	0.90

Figure 14: Dataset usado en la Prueba 1

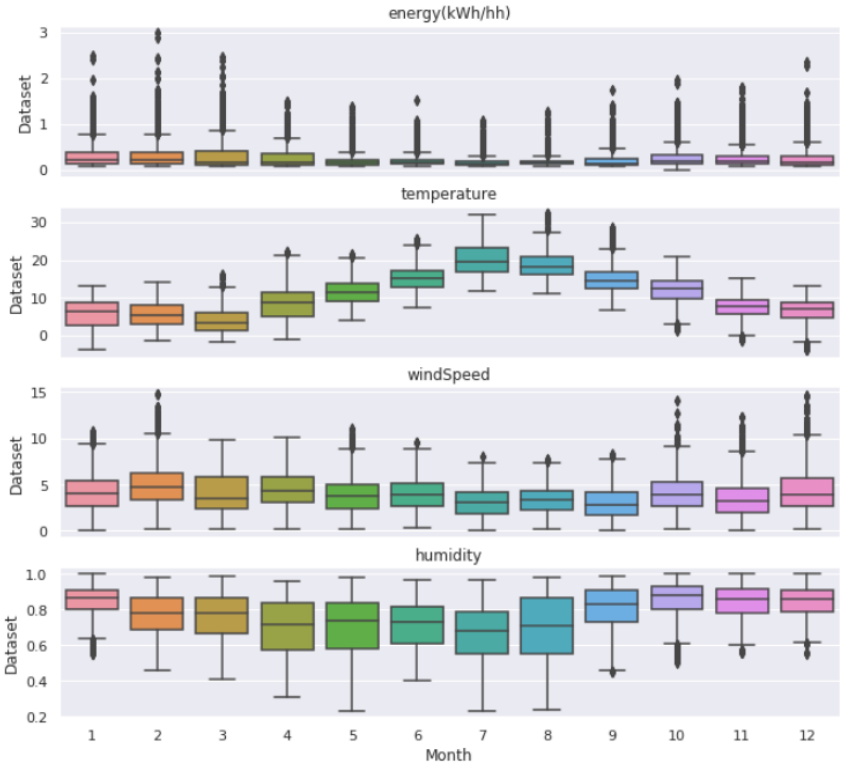


Figure 15: Relación entre las variables del dataset de la Prueba 1

en el entrenamiento y solamente se predice una de las variables específicamente, la variable de consumo energéticos, se dice que es una serie de tiempo multivariable de la forma de series de entrada múltiples. La red LSTM es capaz de hacer una buena predicción, con un coeficiente de determinación del 0.89, que muestra una medida estadística de qué tan bien las predicciones de regresión se aproximan a los puntos de datos reales. Los parámetros de configuración de la prueba se muestran en la Tabla 2.

También se realizó una prueba donde solamente es entrenada la red con los datos de clima y se realizó una predicción del consumo energético, logrando un error de 0.012 y un 0.80 en el coeficiente de determinación. La Figura 16 muestra la predicción del consumo energético con datos de clima y consumo energético anterior como datos de entrada y la Figura 17 muestra la predicción del consumo energético solamente con los datos de clima como entrada.

Durante esta prueba se usaron varios tipos de LSTM, lo cual queda registrado en la Tabla ,

Parámetros del diseño	Valores
Datos de Entrenamiento	16895
Datos de Test	7241
look_back	20
Neuronas capa LSTM	100
Optimizador	Adam
Loss	mse
Epochs	200
batch_size	100

Table 2: Parámetros usados en la Prueba1

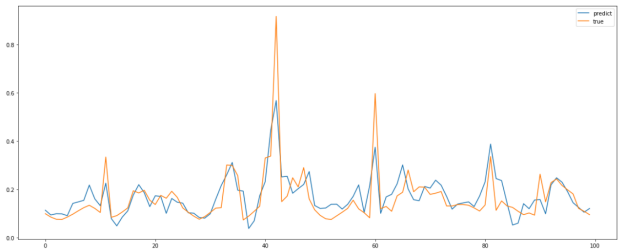


Figure 16: Predicción consumo energético dado datos de clima y consumo anterior

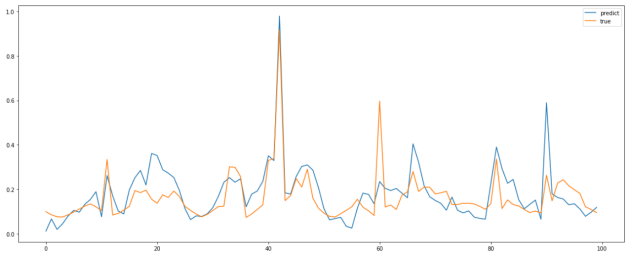


Figure 17: Predicción consumo energético dado datos de clima

Modelo LSTM	**RMSE**	**Coeficiente de Determinación**
LSTM Vainilla	0.082	0.890
Stacked LSTM	0.085	0.883
Bidirectional LSTM	0.138	0.693
CNN LSTM	0.134	0.709
ConvLSTM	0.096	0.850

Table 3: Comparación de resultados entre diferentes modelos LSTM

5.2 Prueba2: Consumo Energético agregado dado Datos de Clima resolución hora

En la presente prueba se realiza una predicción del consumo energético agregado de 11 casas dado los datos de clima con una resolución de una hora. Las 11 casas son usadas debido a que son las casas que poseen mayor cantidad de datos, contando con mediciones desde el 23/11/2011. La Figura 18 muestra el dataset usado para el aprendizaje de la red recurrente. El dataset fue dividido en un 80 % para entrenamiento y un 20% para el test.

	consumo_agregado	temperature	windSpeed	humidity
2011-11-23 13:00:00	2.473	9.87	3.12	0.82
2011-11-23 14:00:00	3.183	10.36	3.26	0.84
2011-11-23 15:00:00	2.484	10.09	3.04	0.85
2011-11-23 16:00:00	1.624	9.84	2.95	0.86
2011-11-23 17:00:00	2.745	9.14	2.58	0.92

Figure 18: Dataset usado en la Prueba 2

El entrenamiento de la red LSTM multivariable fue realizado con los parámetros presentados en la Tabla 4. Los resultados muestran un error de 0.408, y un coeficiente de determinación de 0.60.

Además fue realizada una prueba donde la red solamente es entrenada con los datos de clima y se realizó una predicción del consumo energético, obteniendo un error de 0.596 y un 0.37 en el coeficiente de determinación. La Figura 19

Parámetros del diseño	Valores
Datos de Entrenamiento	15876
Datos de Test	3970
look_back	200
Neuronas capa LSTM	200
Optimizador	Adam
Loss	mse
Epochs	30
batch_size	100

Table 4: Parámetros usados en la Prueba2

muestra la predicción del consumo energético dado los datos de clima y consumo energético anterior y la Figura 20 muestra la predicción del consumo energético solamente con los datos de clima como entrada.

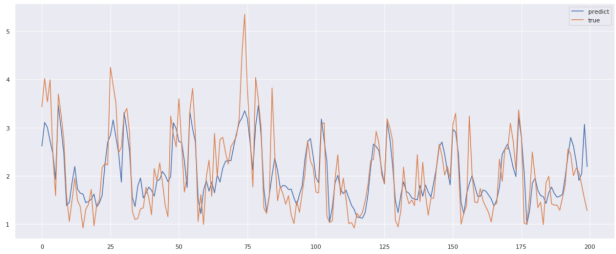


Figure 19: Predicción consumo energético dado datos de clima y consumo anterior

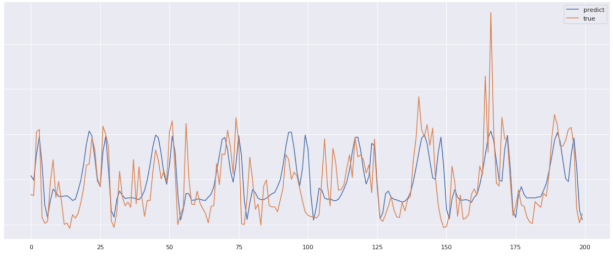


Figure 20: Predicción consumo energético dado datos de clima

5.3 Prueba3: Consumo Energético agregado dado Datos de Clima resolución diaria

En la Prueba 3 se realiza una predicción del consumo energético agregado usando el consumo de 11 casas dado los datos de clima: temperatura, velocidad del viento y humedad, usando el dataset de resolución de un día. Las 11 casas seleccionadas para calcular el consumo agregado, fueron las casas que poseen lecturas desde el primer día registrado en el dataset. El dataset usado en la prueba se muestra en la Figura 21. El dataset fue dividivo en un 70% para el entrenamiento y un 30 % para el test. Los datos fueron normalizados en una escala de [0,1] para lograr un mejor entrenamiento en la red.

day	consumo_agregado	temperatureMax	windSpeed	humidity
2011-11-23	78.968	10.36	2.04	0.93
2011-11-24	120.029	12.93	4.04	0.89
2011-11-25	121.263	13.03	5.02	0.79
2011-11-26	125.614	12.96	5.75	0.81
2011-11-27	131.695	13.54	5.48	0.72

Figure 21: Dataset usado en la Prueba 3

La Tabla 5 muestra la configuración de los principales parámetros que fueron usados en el entrenamiento de la red LSTM.

Los resultados de la Prueba 3 se muestran en la Figura 22, como se puede apreciar estos resultados son pobres debido a la poca cantidad de datos con los que se cuenta, pero se ve como la red LSTM es capaz de reproducir el comportamiento de la serie de tiempo de una forma aceptable aún cuando existen pocos datos.

Parámetros del diseño	Valores
Datos de Entrenamiento	580
Datos de Test	249
look_back	8
Neuronas capa LSTM	100
Optimizador	Adam
Loss	mae
Epochs	400
batch_size	60

Table 5: Parámetros usados en la Prueba3

Figure 22: Resultados Prueba 3

5.4 Prueba4: Consumo Energético agregado dado consumo energético de 11 casas resolución media hora

En esta prueba se realiza la predicción del consumo energético agregado de 11 casas dado los datos de consumo de energía cada media hora de las mismas. La Figura 23 muestra el dataset usado para el aprendizaje de la red LSTM.

tstp	house1	house2	house3	house4	house5	house6	house7	house8	house9	house10	house11	consumo_agregado
2011-11-23 13:30:00	0.110	0.075	0.055	0.938	0.201	0.049	0.01	0.250	0.176	0.151	0.458	2.473
2011-11-23 14:00:00	0.244	0.064	0.144	1.650	0.185	0.048	0.00	0.163	0.092	0.059	0.534	3.183
2011-11-23 14:30:00	0.241	0.318	0.106	0.752	0.248	0.048	0.00	0.177	0.093	0.088	0.353	2.424
2011-11-23 15:00:00	0.287	0.185	0.116	0.713	0.486	0.049	0.00	0.162	0.158	0.086	0.242	2.484
2011-11-23 15:30:00	0.296	0.121	0.139	0.261	0.281	0.048	0.00	0.162	0.142	0.116	0.324	1.890

Figure 23: Dataset utilizado en la Prueba 4

Los parámetros de aprendizaje para la LSTM estan dados en la Tabla 6.

Se utilizó el optimizador RMSProp porque proporcionó un mejor resultado, pero se hicieron pruebas con otros como: Adragrad, Adadelta y Adam, este último dió resultados parecidos al que usamos, pero el tiempo de entrenamiento era más lento. Además se hicieron pruebas con Dropout de 0.1, 0.2, 0.3, en donde para todos los casos el error incrementaba. Las unidades en la capa LSTM se tomaron de 60, pero entrenamos con 100 y 200, para 200 unidades la red se saturaba y demoraba demasiado en cada epoch, debido a la cantidad de datos se tomó un look_back de 300.

La normalización de los datos se hizo entre los rangos [0,1] y se separó el dataset de un 70% para el entrenamiento y un 30 % para el test, obteniendo como resultado final un error mse de 0.018.

Parámetros del diseño	Valores
Datos de Entrenamiento	27765
Datos de Test	11900
look_back	300
Unidades capa LSTM	60
Optimizador	RMSProp
Loss	mse
Epochs	50
batch_size	100

Table 6: Parámetros usados en la Prueba4

La Figura 24 muestra la predicción del consumo agregado para 11 casa cada media hora.

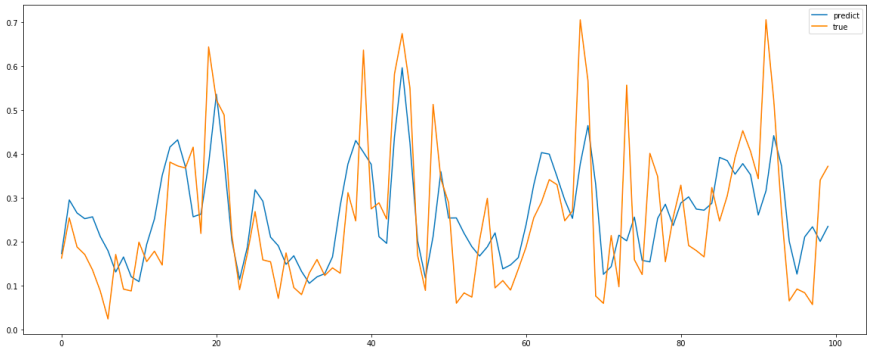


Figure 24: Predicción del consumo agregado para 11 casa cada media hora

5.5 Prueba5: Consumo Energético agregado dado consumo energético de 11 casas resolución hora

En esta prueba se realiza la predicción del consumo energético agregado de 11 casas dado los datos de consumo de energía cada hora de las mismas. La Figura 25 muestra el dataset usado para el aprendizaje de la red LSTM.

tstp	house1	house2	house3	house4	house5	house6	house7	house8	house9	house10	house11	consumo_agregado
2011-11-23 13:30:00	0.110	0.075	0.055	0.938	0.201	0.049	0.01	0.250	0.176	0.151	0.458	2.473
2011-11-23 14:00:00	0.244	0.064	0.144	1.650	0.185	0.048	0.00	0.163	0.092	0.059	0.534	3.183
2011-11-23 14:30:00	0.241	0.318	0.106	0.752	0.248	0.048	0.00	0.177	0.093	0.088	0.353	2.424
2011-11-23 15:00:00	0.287	0.185	0.116	0.713	0.486	0.049	0.00	0.162	0.158	0.086	0.242	2.484
2011-11-23 15:30:00	0.296	0.121	0.139	0.261	0.281	0.048	0.00	0.162	0.142	0.116	0.324	1.890

Figure 25: Descripción de dataset utilizado para la prueba 5

Los parámetros de aprendizaje para la LSTM estan dados en la Tabla 7:

Se hicieron pruebas con Dropout de 0.1, 0.2, 0.3, en donde para todos los casos el error incrementaba. Las unidades en la capa LSTM se tomaron de 130, pero entrenamos con 50, 100 y 200, para 50 unidades el error estaba muy alto mientras que para 200 el error fuE cercano al que obtuvimos finalmente.

La normalización de los datos se hizo entre los rangos [0,1] y se separó el dataset de un 70% para el entrenamiento y un 30 % para el test, obteniendo como resultado final un error mse de 0.026.

La Figura 26 muestra la predicción del consumo agregado para 11 casas cada una hora.

Parámetros del diseño	Valores
Datos de Entrenamiento	13853
Datos de Test	5955
look_back	200
Unidades capa LSTM	130
Optimizador	RMSProp
Loss	mse
Epochs	50
batch_size	100

Table 7: Parámetros usados en la Prueba5

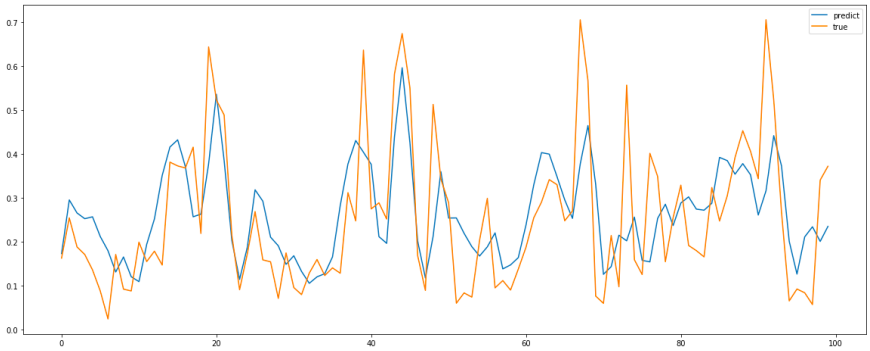


Figure 26: Predicción del consumo agregado para 11 casas cada una hora

5.6 Prueba6: Consumo agregado resolución media hora

Durante la Prueba 6 se realizó una predicción del consumo agregado dado el consumo agregado anterior usando el dataset con resolución cada media hora. La red LSTM es Univariable en este caso, debido a que solamente tiene como entrada la variable de consumo energético y debe ser capaz de predecir el comportamiento futuro de la serie de tiempo dado el comportamiento pasado de esta. Los datos fueron diferenciados y escalados entre [0,1] usando el MinMaxScaler de sklearn, para lograr un mejor aprendizaje de la red. La Figura 27 muestra el dataset usado para la presente prueba.

La Red LSTM durante la prueba fue entrenada con los parámtros que se muestran en la Tabla **??**. Como se puede apreciar en la Tabla **??**, los datos de entrenamiento varían en función de la cantidad de horas que se quieren predecir, se hicieron pruebas para predecir 12 horas, 24 horas, 48 horas, una semana (168 horas) y un mes (672 horas), cabe aclarar que debido a que la resolución es de media hora, para predecir 12 horas se necesita predecir 12*2 step de tiempo. La Figura 28 muestra la predicción de una semana obtenida usando la red LSTM

Parámetros del diseño	Valores
Datos de Entrenamiento	39665
Datos de Test	12,24,48,168, 672
look_back	24
Neuronas capa LSTM	10
Optimizador	Adam
Loss	mae
Epochs	7
batch_size	1

Table 8: Parámetros usados en la Prueba7

5.7 Prueba7: Consumo agregado resolución hora

Durante la Prueba 7 se realizó una predicción del consumo agregado dado el consumo agregado anterior usando el dataset con resolución de una hora. La red LSTM es Univariable en este caso, debido a que solamente tiene como

tstp	consumo_agregado
2011-11-23 13:30:00	2.473
2011-11-23 14:00:00	3.183
2011-11-23 14:30:00	2.424
2011-11-23 15:00:00	2.484
2011-11-23 15:30:00	1.890

Figure 27: Dataset usado en la Prueba 6

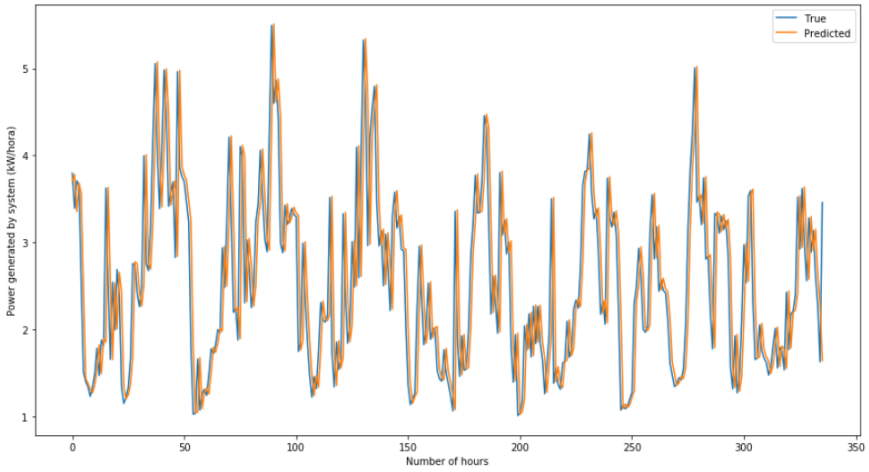


Figure 28: Predicción de una semana con una resolución cada media hora

entrada la variable de consumo energético y debe ser capaz de predecir el comportamiento futuro de la serie de tiempo dado el comportamiento pasado de esta. El dataset usado en la prueba se muestra en la siguiente Figura 29. El dataset fue dividido en un 70% para el aprendizaje y un 30% para el test. Los datos fueron escalados para el entrenamiento usando MinMaxScaler de sklearn entre [0,1], mostrando el aprendizaje un mejoramiento considerable en comparación cuando es entrenado en la escala original de los datos.

Los datos de configuración de la Red LSTM durante la prueba se muestran en la Tabla 9. Como se puede apreciar en la Tabla 9, los datos de entrenamiento varían en función de la cantidad de horas que se quieren predecir, se hicieron pruebas para predecir 12 horas, 24 horas, 48 horas, una semana (168 horas) y un mes (672 horas). La Figura 30 muestra la predicción de una semana obtenida usando la red LSTM

La Tabla 10 muestra una comparación del error entre las pruebas 6 y 7 respectivamente, las cuales tienen la diferencia de tener diferentes resoluciones de datos, media hora y una hora respectivamente. Se muestra que de forma general existe un menor error en las predicciones que poseen una resolución de media hora, debido a que la red posee mayor cantidad de valores para aprender.

tstp	consumo_agregado
2011-11-23 13:00:00	2.473
2011-11-23 14:00:00	3.183
2011-11-23 15:00:00	2.484
2011-11-23 16:00:00	1.624
2011-11-23 17:00:00	2.745

Figure 29: Dataset usado en la Prueba 7

Parámetros del diseño	Valores
Datos de Entrenamiento	19835
Datos de Test	12,24,48,168, 672
look_back	24
Neuronas capa LSTM	100
Optimizador	Adam
Loss	mae
Epochs	7
batch_size	1

Table 9: Parámetros usados en la Prueba7

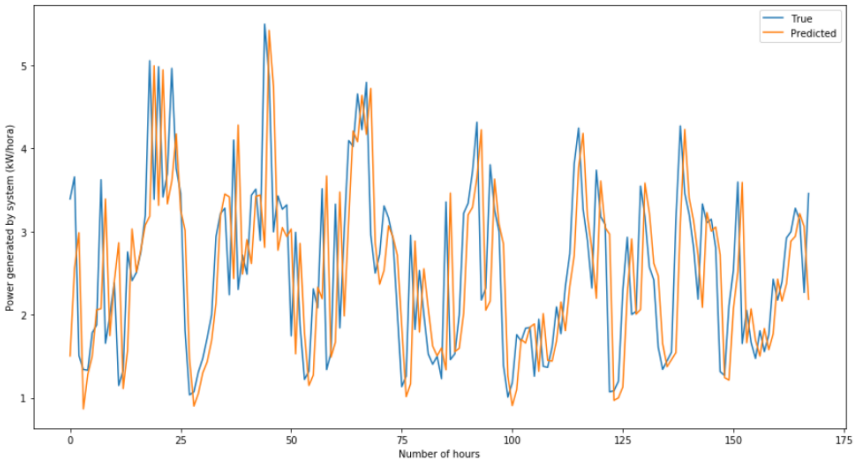


Figure 30: Predicción de una semana con una resolución de una hora

Tiempo	Resolución 1 hora	Resolución media hora
	Cantidad de datos=19848	Cantidad de datos=39665
	MAE%	
12 horaas	14.94	2.3
24 horas	12.14	20.25
48 horas	25.77	18.52
1 semana	29.93	21.11
1 mes	29.32	21.4

Table 10: Comparación de resultados Prueba 6 y Prueba 7

5.8 Prueba8: Consumo agregado por Acorn resolución diaria

Durante la Prueba 8 se toma en cuenta la clasificación socioeconómica (Afluentes, Confortable, Adversidad) que es representada a través del *Acorn* que nos brinda los dataset principales usados en el proyecto. Para obtener el consumo agregado por cada *Acorn* se usó el consumo energético diario de 100 casas de cada una de las clasificaciones socioeconómica, en las fechas comprendidas entre el 23/11/2011 y el 28/02/2014. La Figura 31muestra el dataset usado para el entrenamiento de la red LSTM.

day	Afluentes	Confortable	Adversidad
2011-11-23	1021.759001	794.949	777.872
2011-11-24	1037.756001	820.013	785.898
2011-11-25	1067.520001	828.463	788.827
2011-11-26	1098.287001	839.799	782.273
2011-11-27	1105.152001	837.558	801.556

Figure 31: Dataset del consumo agregado según la clasificación socioeconómica

El dataset solamente cuenta con 829 datos, debido a que es el consumo agregado con una resolución diaria. Se realizó una separación del dataset de un 70% para el entrenamiento y un 30 % para el test. Como se puede apreciar existen pocos datos para el entrenamiento de una LSTM, pero se usó 400 epoch que permitieron un entrenamiento satisfactorio, además los datos fueron normalizados entre 0 y 1, lo cual permitió un mejor resultado. La Tabla 11 muestra los principales parámetros configurados. Debido a que la red LSTM recibe tres parámetros para la predicción se dice que una red LSTM Multivariable y además realiza una predicción para cada una de sus variables de entrada lo que permite una salida multiple paralela. En la Figura 32 y la Figura 33 podemos ver la predicción realizada para la clasificación de afluentes y adversidad respectivamente.

Parámetros del diseño	Valores
Datos de Entrenamiento	580
Datos de Test	249
look_back	8
Neuronas capa LSTM	100
Optimizador	Adam
Loss	mae
Epochs	400
batch_size	50

Table 11: Parámetros usados en la Prueba8

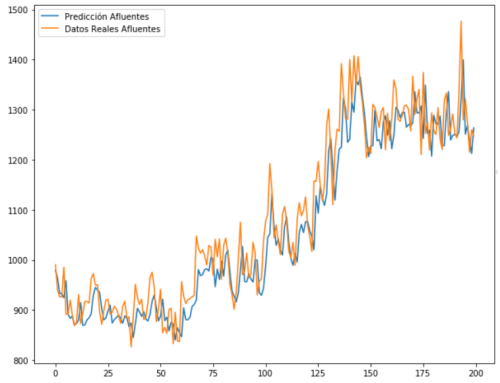


Figure 32: Predicción consumo agregado de la clasificación Afluentes

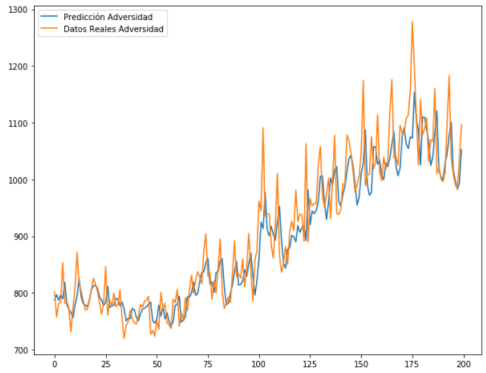


Figure 33: Predicción consumo agregado de la clasificación Adversidad

6 Conclusiones

El proyecto ha sido útil para ampliar los conocimientos aprendidos a través del ramo, debido a la gran investigación involucrada para lograr nuestros objetivos en el transcurso de este proyecto. Fueron estudiados de forma teórica y práctica las redes LSTM y los diferentes tipos de series de tiempo. Además se ha realizado un análisis del comportamiento del consumo energético para hacer una predicción de este. Las pruebas realizadas permitieron corroborar el buen funcionamiento de la redes LSTM para hacer una predicción de las series de tiempo. Se obtuvieron buenos resultados en la predicción de consumo energético dado los datos de clima, y el consumo energético pasado.

References

[1] https://machinelearningmastery.com/how-to-develop-lstm-models-for-time-series-forecasting/

[2] https://machinelearningmastery.com/time-series-forecasting/

[3] https://www.kaggle.com/jeanmidev/smart-meters-in-london#acorn_details.csv

[4] https://towardsdatascience.com/predicting-stock-price-with-lstm-13af86a74944

[5] https://github.com/ShashwatArghode/Wind-Energy-Prediction-using-LSTM

[6] https://eprints.ucm.es/49444/1/2018-MIGUEL%20CABEZON%20Memoria.pdf

[7] http://colah.github.io/posts/2015-08-Understanding-LSTMs/

[8] https://github.com/omerbsezer/LSTM_RNN_Tutorials_with_Demo

[9] https://github.com/ShashwatArghode/Wind-Energy-Prediction-using-LSTM