Date of publication 10, 2022, date of current version 10, 2022.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Forecasting On Power Consumption Of Tetouan City

Daniel Molina, Dorian Jaramillo[1]
[1] Universidad de Antioquia CO, Departamento Ingeniería de Sistemas

Correos de los autores: Daniel Molina (e-mail: daniel.molinay@udea.edu.co), Dorian Jaramillo (e-mail: : dorian.jaramillo@udea.edu.co))

ABSTRACT La energía determina en gran medida el desarrollo y la sostenibilidad de un país, la forma en la que una sociedad administra la producción de energía, el consumo y su exportación son factores determinantes para su economía; por tanto la predicción del consumo de energía se convierte en un problema de vital importancia, es importante analizar el desbalance que existe entre producción y consumo para la toma de decisiones con el fin de cubrir la demanda requerida para un tiempo o temporada específica. En este trabajo se analiza la base de datos "Power consumption of Tetouan city Data Set"[1], tratándose de un problema de regresión se utilizaran los siguientes modelos: Redes neuronales, Regresión múltiple, Random Forest, Ventana de parzen y Máquinas de soporte vectorial. Los modelos serán aplicados variando algunos parámetros y seleccionando aquel para el cual se obtengan los mejores resultados.

INDEX TERMS Power consumption analysis, Machine learning for forecasting, Power consumption of Tetouan city

I. INTRODUCCIÓN

La demanda de energía crece de forma vertiginosa debido a factores como: crecimiento poblacional, vehículos eléctricos, desarrollo económico, industrialización, centros de datos y minado de criptomonedas.

Aunque el avance tecnológico ha permitido la creación de ciudades inteligentes y dispositivos que autorregulan el consumo de energía, a su vez los nuevos avances vienen con nuevas formas de consumo masivo de energía como:

- Centros de cómputo para el minado de criptomonedas:
  dichos centros de minado en conjunto consumen más energía que países enteros.

- Vehículos eléctricos:
  Los vehículos eléctricos necesitan de estaciones de energía para su carga.

Debido a esto y otros factores se tiene como resultado un incremento anual del consumo de energía en la mayoría de los países desarrollados o en vía de desarrollo. se vuelve trascendente entonces contar con herramientas que permitan predecir el consumo de energía a futuro.

El problema abordado corresponde a un problema de regresión, donde se tratará de predecir el consumo de energía teniendo en cuenta los atributos fecha y hora, temperatura, humedad, velocidad del viento, flujo difuso general, flujo difuso y tres atributos de consumo de las zonas 1, 2 y 3.

Aunque se consideran distintos modelos, se hace énfasis en redes neuronales LSTM que funcionan bastante bien con series temporales, se tomó dicha decisión debido a que el consumo de energía depende mucho de la estación climática o temporadas especiales como navidad.

II. ESTADO DEL ARTE

Se han encontrado autores que abordan el conjunto de datos del consumo de energía de la ciudad marroquí Tetouan. Uno de ellos es Waleed Abdu Zogaan[2] que propone el uso de métodos de Machine Learning como Random Forest, Máquinas de soporte vectorial y redes neuronales artificiales.

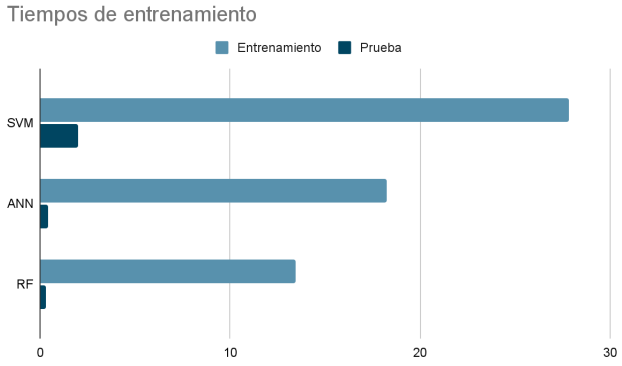
en la siguiente tabla se muestran los resultados obtenidos por el autor para cada uno de los modelos utilizados

Tabla 1 - Resultados para diferentes métricas de error

Model	MAE	MSE	R^2
SVM	0.7073	0.1952	0.9765
ANN	0.8441	13.5639	0.7335
RF	0.2166	0.0987	0.9013

- Teniendo en cuenta dichos resultados se logra evidenciar que el modelo RF obtuvo el error más bajo tanto en MAE como en MSE pero las máquinas de soporte vectorial SVM obtuvieron mejor R^2.

Figura 1 - Tiempos de entrenamiento



En la figura anterior se puede evidenciar que el modelo con menor costo durante el entrenamiento es RF.

Los otros autores en cuestión son Abdulwahed Salam y Abdelaaziz El Hibaoui [3], dichos autores proponen un modelo de deep learning haciendo variaciones en las redes neuronales utilizadas, obteniendo como resultado los siguientes datos:

Tabla 2 - Medidas obtenidas por algoritmo

Model	Median RMSE	Ensemble RMSE
DFFNN	7208	6611.7
DFFNN-ResNet	7397	7071.3
CNN	7191.9	7079.9
CNN-ResNet	6874	6365.6
CNN LSTM	6744.1	6645.2
CNN-ResNet LSTM	6429.1	6424.6
DFNN LSTM	6547.5	6944.5
DFNN ResNet LSTM	6941.4	6538
DENSENET	10220	9978
DENSENET LSTM	7443.9	7238.3
EECP-CBL	8146.2	6604.97

III. EXPERIMENTOS

La base de datos que se está analizando representa el consumo de energía en tres zonas de la ciudad de Tetouan [1] ubicada al norte de Marruecos, dicha base de datos cuenta con un total de 52,417 instancias.

El problema a resolver es multivariable, cuenta con un total de 9 variables continuas de las cuales 6 de ellas son independientes y corresponden a los datos de entrada y tres de ellas son dependientes y corresponden a los datos de salida.

Tabla 3 - Variables de entrada del modelo

ATRIBUTOS	DESCRIPCIÓN
DateTime	Representa la fecha y la hora en la que fueron tomadas las muestras. (La recolección fue realizada cada 10 minutos)
Temperature	Temperatura de la ciudad de Tetouan.
Humidity	Humedad de la ciudad Tetouan.
Wind Speed	Velocidad del viento de la ciudad de Tetouan.
general diffuse flows	Flujo difuso general
diffuse flows	Flujo difuso

Tabla 4 - Variables de salida del modelo

VARIABLE	DESCRIPCIÓN
power consumption of zone 1 of Tetouan city	Consumo de energía para la zona 1
power consumption of zone 2 of Tetouan city	Consumo de energía para la zona 2
power consumption of zone 3 of Tetouan city	Consumo de energía para la zona 3

La metodología de validación elegida es la "validación cruzada" (K-Fold), se descarta la validación simple porque los datos destinados para pruebas no son tenidos en cuenta para el entrenamiento, prescindiendo de datos importantes para mejorar el modelo, de igual forma se descarta K-Fold estratificado debido a que nuestro problema no corresponde a un problema de clases desbalanceado, y por último se descarta "K-Fold leave one out" debido a su excesivo coste durante el entrenamiento.

El numero de particiones usadas para la validación cruzada fue de tres, se optó por este valor con el fin de tener un equilibrio entre el tiempo que toma el entrenamiento y la eficiencia del modelo resultante; dicha configuración fue aplicada para todos los algoritmos usados durante el entrenamiento.

El problema en cuestión es un problema de regresión por tanto no se presenta el inconveniente de las clases desbalanceadas.

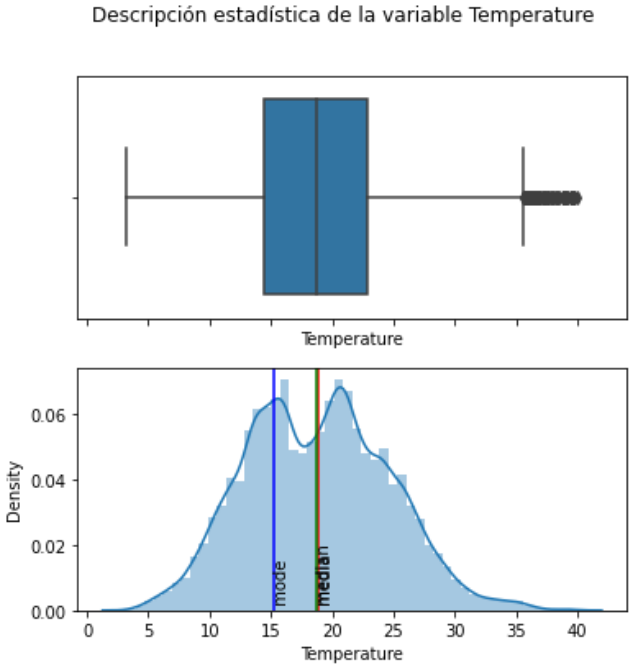
A continuación se describen algunas medidas descriptivas de las características que se usaran para el entrenamiento de los modelos.

Tabla 5 - Análisis de los características

VARIABLE	RANGO	MEDIA	MODA	MEDIANA	INTERCUARTIL
Temperatura	83.46	68.25	85.9	69.86	8.48
Humedad	83.46	68.25	85.9	69.86	23.09
Velocidad del viento	64.433	1.69	0.08	0.09	4.837
Flujo difuso general	1162.99	182.96	0.05	05.03	319.53
Flujo difuso	935.98	75.02	0.11	4.45	100.87

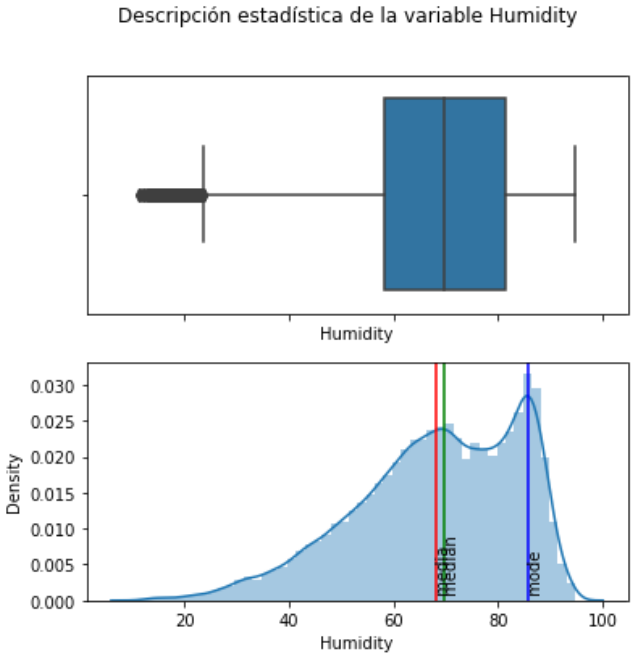
En las siguientes figuras se muestran los diagramas de caja de bigotes y la FDP de las variables Temperatura, Humedad y velocidad del viento.

Figura 2: Caja de bigotes y FDP de la variable temperatura



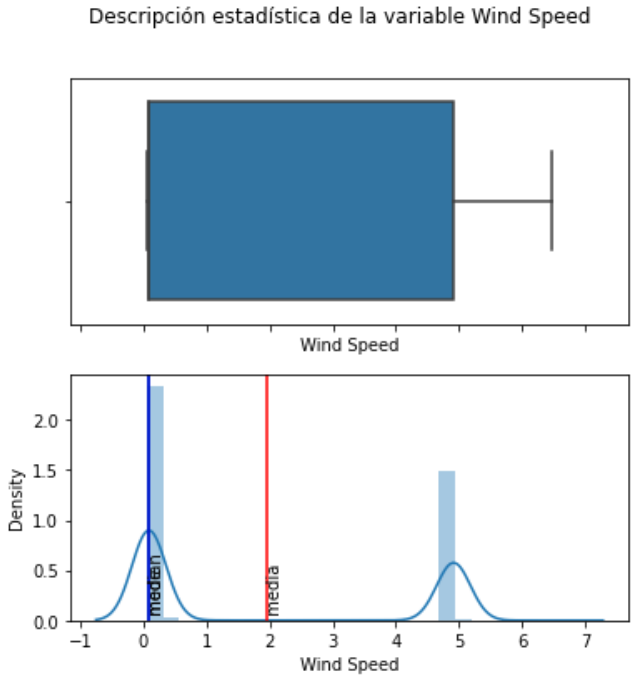
Para la variable temperatura la distribución es bimodal, en este caso no se presentan outliers representativos.

Figura 3: Caja de bigotes y FDP de la variable humedad



Para la variable humedad la distribución es asimétrica negativa y hay una gran cantidad de outliers a la izquierda, representando humedades extremadamente bajas, posible consecuencia de un aumento en la temperatura.

Figura 4: Caja de bigotes y FDP de la variable velocidad del viento



Para la variable velocidad del viento el tipo de distribución es bimodal, con una moda muy cerca de "0" y otra moda muy cerca de "5". En este caso no cuenta con outliers.

Medidas de desempeño

Entre las medidas de desempeño usadas se encuentran:

R^2: usada con el fin de determinar el porcentaje de varianza que puede explicar el modelo (medida principal de desempeño).

MAPE: se usa MAPE en lugar de MAE con el fin de obtener una medida en porcentaje del error dado por el modelo.

Intervalo de confianza: dado por la desviación estándar del MAPE, nos ayuda a identificar qué tanta variación existe entre las medidas de error del modelo.

Duración del entrenamiento: Indica la cantidad de minutos que tomó el modelo en entrenar completamente dado un conjunto específico de parámetros.

IV. MODELOS DE APRENDIZAJE

Por tratarse de un problema de regresión, en el cual se trata de predecir el consumo de energía en tres zonas de la ciudad de Tetouan, la base de datos será abordada por los modelos de aprendizaje regresión múltiple, ventana de parzen, random forest, redes neuronales artificiales y regresión por vectores de soporte con kernel RBF

Regresión multiple

Tabla 6 - Resultados entrenamiento para regresión múltiple

ZONA	R^2	MAPE	INTERVALO DE CONFIANZA
1	0.523610	0.125662	0.012095
2	0.327768	0.145760	0.013714
3	-1.964040	0.391489	0.238354

El modelo de regresión múltiple en específico no contaba con hiperparametros, por tanto al no tener variaciones en los resultados de entrenamiento se tiene que los mejores resultados de R^2 para cada una de las zonas fueron:

Tabla 7 - Mejor resultado de entrenamiento para regresión

	Zona 1	Zona 2	Zona 3
R^2	0.52	0.32	-1.96

De dichos resultados se infiere que el modelo con mayor porcentaje de varianza explicada es el modelo de la zona 1.

Ventana de parzen

Tabla 8 - Resultados entrenamiento para ventana de parzen

ZONA	H	R^2	MAPE	I.C	DURACIÓN
1	1	0.507809	0.129746	0.015187	20.3779796
	2	0.214197	0.167071	0.010042	214.460829
	3	0.038492	0.184848	0.011217	206.747262
	5	0.195756	0.195756	0.011346	202.737921
2	1	0.459666	0.140867	0.000608	20158059
	2	0.097628	0.189485	0.031842	213.262206
	3	-0.115388	0.212893	0.043641	208.042370
	5	-0.255256	0.226903	0.048833	209.088500
3	1	-2.48606	0.414835	0.28658	204.211580

		4			
	2	-1.942152	0.409319	0.230391	207.747027
	3	-1.825999	0.403060	0.213083	201.266425
	5	-1.771506	0.399480	0.202722	200.833213

Analizando los hiperparametros se observa que el mejor resultado de R2 se obtiene cuando H es igual a 1:

Tabla 9 - Mejores resultados para ventana de parzen

ZONA	H	R2
1	1	0.50
2	1	0.45
3	1	-2.48

Random Forest

Tabla 10 - Resultados usando random forest

ZONA	**MAX-CAR**	**N-ESTIMADOR**	**R^2**	**MAPE**	**I-C**	**DURACIÓN**
1	2.0	5	0.950375	0.034587	0.000522	0.014
	3.0	5	0.963139	0.029403	0.000186	0.018
	4.0	5	0.965296	0.028320	0.000139	0.022
	5.0	5	0.967137	0.027528	0.000240	0.024
	2.0	10	0.961768	0.030733	0.000140	0.029
	3.0	10	0.968551	0.027576	0.000215	0.036
	4.0	10	0.970446	0.026341	0.000355	0.041
	5.0	10	0.971673	0.025699	0.000322	0.048
	2.0	15	0.965332	0.029459	0.000140	0.041372
	3.0	15	0.971065	0.026544	0.000227	0.052927
	4.0	15	0.972195	0.025527	0.000338	0.063782
	5.0	15	0.973404	0.024978	0.000073	0.072232
	2.0	20	0.967088	0.028638	0.000381	0.056357
	3.0	20	0.971107	0.026456	0.000139	0.071063
	4.0	20	0.973366	0.025058	0.000184	0.083253
	5.0	20	0.973956	0.024669	0.000130	0.096852
2	2.0	5	0.953145	0.038705	0.000515	0.014051
	3.0	5	0.962816	0.033793	0.000072	0.017682
	4.0	5	0.967443	0.031125	0.000418	0.021346
	5.0	5	0.969107	0.030184	0.000176	0.025059
	2.0	10	0.962952	0.034602	0.000692	0.028541
	3.0	10	0.970313	0.030375	0.000152	0.034979
	4.0	10	0.972724	0.028672	0.000089	0.041454
	5.0	10	0.973706	0.027990	0.000200	0.048595
	2.0	15	0.967351	0.032872	0.000259	0.043113
	3.0	15	0.972802	0.029362	0.000194	0.051396
	4.0	15	0.974256	0.027871	0.000244	0.061788
	5.0	15	0.974991	0.027386	0.000420	0.072163
	2.0	20	0.967677	0.032652	0.000075	0.054899
	3.0	20	0.973644	0.028810	0.000179	0.068217
	4.0	20	0.975382	0.027442	0.000026	0.081570
	5.0	20	0.976194	0.026663	0.000063	0.093834
3	2.0	5	0.967523	0.044938	0.002150	0.014
	3.0	5	0.976558	0.037859	0.000219	0.017

	4.0	5	0.978745	0.035328	0.000420	0.021
	5.0	5	0.980440	0.033861	0.000417	0.024
	2.0	10	0.975204	0.040381	0.001051	0.028
	3.0	10	0.980331	0.034599	0.000415	0.034
	4.0	10	0.982418	0.032714	0.000501	0.041
	5.0	10	0.983030	0.031957	0.000264	0.047
	2.0	15	0.978017	0.037860	0.000747	0.040
	3.0	15	0.981776	0.033687	0.000542	0.051
	4.0	15	0.983006	0.031930	0.000079	0.060
	5.0	15	0.983805	0.031070	0.000089	0.070
	2.0	20	0.978611	0.037317	0.000203	0.057
	3.0	20	0.982646	0.033044	0.000282	0.073
	4.0	20	0.983944	0.031326	0.000235	0.084
	5.0	20	0.984315	0.030803	0.000371	0.100

Con base en los resultados se observa que la mejor configuración de hiperparametros por zona utilizando random forest corresponde a:

Tabla 11 - Mejores resultados con Random Forest

Zona	**MAX-CAR**	**N-ESTIMADOR**	**R^2**
1	5	20	0.973956
2	5	20	0.976194
3	5	20	0.984315

Máquina de soporte vectorial

Tabla 12 - Resultados usando SVM

Zona	**KERNEL**	**GAMMA**	**R^2**	**MAPE**	**I-C**	**DURACIÓN**
1	RBF	Auto	0.287317	0.156596	0.001114	5.737755
	RBF	Scalar	0.287669	0.156565	0.001054	5.916860
	Lineal	Auto	0.627	0.107	0.000	2.130
			081	648	731	983
	Lineal	Escalar	0.626872	0.107682	0.000142	2.131243
	Poly	Auto	0.140145	0.159721	0.001615	3.050227
	Poly	Escalar	0.240263	0.159724	0.000641	3.027026
2	RBF	Auto	0.332531	0.167998	0.001421	5.886628
	RBF	Escalar	0.332424	0.168060	0.000568	5.932536
	Lineal	Auto	0.586628	0.128606	0.001422	2.12.8587
	Lineal	Escalar	0.586539	0.128620	0.001110	2.127127
	Poly	Auto	0.269322	0.176996	0.001037	2.296549
	Poly	Escalar	0.269201	0.176997	0.001098	2.592007
3	RBF	Auto	0.158605	0.246720	0.000517	5.951777
	RBF	Escalar	0.158588	0.246695	0.000844	5.944983
	Lineal	Auto	0.538520	0.187995	0.007033	2.099771
	Lineal	Escalar	0.538536	0.187960	0.001000	2.239953
	Poly	Auto	0.186008	0.248367	0.001946	2.391817
	Poly	Escalar	0.186110	0.248369	0.000574	2.298353

Tabla 13 - Mejores resultados usando SVM

ZONA	**KERNEL**	**GAMMA**	**R^2**
1	Linear	Auto	0.627081
2	Linear	Auto	0.586628
3	Linear	Escalar	0.868536

Se observa que en general al aplicar “Linear” como kernel y “Auto” como Gamma se obtienen los mejores resultados..

Redes neuronales

Tabla 14 - Mejores resultados usando Redes neuronales

ZONA	F. ACTIVACIÓN	NEURONAS	CAPAS OCULTAS	R^2
1	Relu	8.0	3.0	0.167610
2	Relu	8.0	3.0	0.063973
3	Relu	8.0	3.0	-2.272303

En general se observa un resultado poco óptimo para R^2 con el uso de redes neuronales, sin embargo se encuentra que usando "Relu" como función de activación, 8 neuronas y 3 capas ocultas se obtiene el mejor resultado posible.

V. SELECCIÓN Y EXTRACCIÓN DE CARACTERÍSTICAS

Debido a que el problema es de regresión se selecciona el coeficiente de pearson como medida de correlación.

Tabla 15 - Tabla de correlación entre las variables de entrada y salida

Caracteristica	Zona 1	Zona 2	Zona 3
Temperature	0.44	0.38	0.49
Humidity	-0.29	-0.29	-0.23
Wind Speed	0.17	0.15	0.28
general diffuse flows	0.19	0.16	0.06
diffuse flows	0.08	0.04	-0.04
Month	-0.00	0.32	-0.23
Day	0.03	0.05	0.00
DayWeek	-0.07	-0.12	0.00
Hour	0.73	0.67	0.45

Analizando los datos de la tabla anterior, las características candidatas a ser eliminadas son:

- diffuse flows
- general diffuse flows
- Month
- Day
- DayWeek

Se dejarán las otras características porque en todos los casos su coeficiente es mayor a 0.1, por lo tanto se considera que tienen una influencia significativa en el modelo.

Selección de características - Método de búsqueda secuencial ascendente

Entre los criterios utilizados para el algoritmo de selección se encuentran:

1. direction = "forward"
2. scoring = "r2"
3. n_features_to_select = 4

Se seleccionaron 4 características debido a que en el análisis de correlación las características más relevantes fueron : Temperature, Humidity, Wind Speed y Hour.

Se seleccionó "r2" como criterio de evaluación porque este fue el que se usó previamente en los entrenamientos sin selección de características y también porque es una excelente métrica para medir la variación explicada por el modelo.

Tabla 16 - Mejora R2 para regresión lineal múltiple con método de búsqueda secuencial ascendente

Zona	R2	R2 con selección	% de mejora
Zona 1	**0.52**	**0.64**	**11%**
Zona 2	**0.33**	**0.59**	**26%**
Zona 3	**-1.96**	**0.58**	**254%**

La predicción del modelo mejora de manera significativa para la zona 2 y 3, llegando incluso a un 254% en la zona 3.

Tabla 17 - Cambio de R2 para random forest con método de búsqueda secuencial ascendente

Zona	R2	R2 con selección	% de mejora
Zona 1	**0.97**	**0.96**	**-1%**
Zona 2	**0.97**	**0.95**	**-2%**
Zona 3	**0.98**	**0.96**	**-2%**

con base en la tabla se logra observar que para el caso de random forest el rendimiento del modelo empeoró un poco usando selección de características, en promedio empeoró un 1%.

Tabla 18 - Mejora R2 para SVM con método de búsqueda secuencial ascendente

Zona	R2	R2 con selección	% de mejora
Zona 1	**0.63**	**0.63**	**0%**
Zona 2	**0.59**	**0.59**	**0%**
Zona 3	**0.54**	**0.54**	**0%**

usando selección de características para SVM, se observó que el porcentaje de mejora es nulo.

Extracción de características - PCA

Entre los tres mejores modelos sin usar ninguna técnica de selección o extracción de características se encuentran:

- Random Forest
- SVM
- Regresión Múltiple.

A continuación se aplicará el método PCA a cada uno de ellos y se observará su porcentaje de mejora

Tabla 19 - Mejora R2 para Regresión múltiple usando PCA

Zona	R2	R2 con PCA	% de mejora
Zona 1	**0.52**	**0.54**	**2%**
Zona 2	**0.33**	**0.57**	**24%**
Zona 3	**-1.96**	**0.23**	**220%**

con base en la tabla se observa que usando PCA se mejoró significativamente el rendimiento en la predicción, especialmente para la zona 2 y 3, obteniendo un 220% de mejora en la zona 3.

Tabla 20 - Mejora R2 para Random forest usando PCA

Zona	R2	R2 con PCA	% de mejora
Zona 1	**0.97**	**0.93**	**-4%**
Zona 2	**0.97**	**0.93**	**-4%**
Zona 3	**0.98**	**0.89**	**-9%**

Con base en la tabla anterior se logra evidenciar que aplicando PCA a Random Forest no se obtiene ningún tipo de mejora, el cambio en el rendimiento es negativo.

Tabla 21 - Mejora R2 para SVM usando PCA

Zona	R2	R2 con PCA	% de mejora
Zona 1	**0.63**	**0.53**	**-l0%**
Zona 2	**0.59**	**0.56**	**-3%**
Zona 3	**0.54**	**0.18**	**-36%**

Con base en el resultado anterior se logra observar que aplicando PCA para SVM el rendimiento empeora significativamente en especial para la predicción de la zona 3.

Análisis de resultados aplicando selección de características

En general se observa que al aplicar selección de características solo mejora el modelo de regresión logística, para el caso de SVM y Random forest su mejora en el rendimiento es nula o incluso un poco negativa.

Análisis de resultados aplicando PCA

En general se observa que al aplicar PCA solo mejora el modelo de regresión logística, para el caso de SVM y Random forest su rendimiento empeora.

Comparativa resultados con los artículos consultados

Haciendo un contraste entre la tabla 1 y las tablas que representan los mejores resultados para los modelos abordados durante el entrenamiento, se logra evidenciar la siguiente diferencia

Tabla 22 - Comparación resultados de entrenamiento artículo actual y otros autores.

Modelo	R2 otro autor	R2 actual	Diferencia
SVM	**0.9765**	**0.6940**	**0.2825**
ANN	**0.7335**	**-0.6802**	**1.4137**
RF	**0.9013**	**0.9781**	**-0.0768**

de la tabla anterior se logra evidenciar que para el autor del artículo citado en el trabajo presente, su mejor resultado lo obtuvo aplicando máquinas de soporte vectorial y su peor resultado lo obtuvo aplicando redes neuronales, en el caso del presente artículo el mejor resultado se obtuvo aplicando random forest, de todos los resultados obtenidos, random forest es el más favorable con un R2 de 0.97.

VI. CONCLUSIÓN

El problema en cuestión es bastante interesante ya que es de gran envergadura por ser la energía y su administración algo inherente a una sociedad entera. En base al estado del arte se tiene cierta predisposición a redes neuronales, sin embargo se encuentra que Random Forest nos dio los mejores resultados con un porcentaje de variabilidad explicada R^2 de 0.97, un porcentaje de error MAPE de 0.03 y un I.C de 0.0001

REFERENCES

[1] Machine Learning Repository "Power consumption of Tetouan city Data Set" Dic 2018 https://archive.ics.uci.edu/ml/datasets/Power+consumption+of+Tetouan+city

[2] W. A. Zogaan, "Power Consumption prediction using Random Forest model", Vol 7, pp 329 - 341, https://kalaharijournals.com/resources/Special_Issue_April_May_43.pdf

[3] A. Salam, A. El Hibaoui, "Energy consumption prediction model with deep inception residual network inspiration and LSTM", May. 2021 https://www.sciencedirect.com/science/article/pii/S03784
75421001774

[4] Daniel Molina, Dorian Jaramillo "Video sustentación" https://drive.google.com/file/d/1Tt1KTU_gPIK0Cx9uK__vwgGAFzjpshNO/view?usp=sharing