



**Pronóstico de generación de potencia eléctrica en paneles solares con técnicas de aprendizaje automático**

John Edison Orrego Muñoz

Bryan Andrés García Villa

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor

Antonio Jesús Tamayo Herrera, Magíster (MSc) en Ingeniería

Universidad de Antioquia

Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2022

---

Cita

García Villa y Orrego Muñoz [1]

---

**Referencia**

Estilo IEEE (2020)

- [1] B.A. García Villa y J. E. Orrego Muñoz, “Pronostico de generación de potencia eléctrica en paneles solares con técnicas de aprendizaje automático”, Trabajo de grado especialización, Especialización en Analítica y Ciencia de Datos, Universidad de Antioquia, Medellín, Antioquia, Colombia, 2022.
- 



Especialización en Analítica y Ciencia de Datos, Cohorte III.



Elija un elemento.

**Repositorio Institucional:** <http://bibliotecadigital.udea.edu.co>

Universidad de Antioquia - [www.udea.edu.co](http://www.udea.edu.co)

**Rector:** John Jairo Arboleda Céspedes.

**Decano/Director:** Jesús Francisco Vargas Bonilla.

**Jefe departamento:** Diego José Luis Botía Valderrama.

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

## **Dedicatoria**

Este documento se dedica a nuestras familias, asesor y amigos, pues sin su apoyo no se habría logrado realizar esta monografía. La bendición de nuestras familias a diario y a lo largo de la vida que protege y nos lleva por el mejor camino. A nuestros docentes por darnos grandes enseñanzas académicas y a nuestros amigos por permitirnos aprender más de la vida a su lado. Todo esto es posible gracias ustedes.

## **Agradecimientos**

Primeramente, damos gracias a la Universidad de Antioquia por habernos permitido formarnos en ella en una educación superior, a todas las personas que estuvieron en este proceso, fueron ustedes los que nos ayudaron para conseguir este logro, gracias a nuestros padres, que son nuestros motor y apoyo en la vida para nunca desistir de nuestros sueños.

Este documento es muy especial para nosotros y esperamos que perdure por muchos años, por todas las personas que tuvieron el interés de verla, agradecemos con todo nuestro ser por haberla leído.

## CONTENIDO

RESUMEN.....	8
ABSTRACT .....	9
I. INTRODUCCIÓN .....	10
II. PLANTEAMIENTO DEL PROBLEMA .....	13
III. OBJETIVOS.....	14
IV. MARCO TEÓRICO.....	15
V. METODOLOGÍA .....	18
VI RESULTADOS.....	22
VII. CONCLUSIONES.....	33
REFERENCIAS .....	35

## LISTA DE TABLAS

Tabla 1.Resultado de métricas iniciales de modelos clásicos. ....	24
Tabla 2. Resultado de métricas totales de modelos clásicos. ....	25
Tabla 3.Resultado de métricas iniciales para planta 1 .....	27
Tabla 4.Resultado de métricas totales planta 1 .....	27
Tabla 5.Experimentos Perceptrón. ....	28
Tabla 6.RMSE de modelos clásicos y deep learning. ....	31
Tabla 7.Mejores resultados planta 1.....	32

## LISTA DE FIGURAS

Ilustración 1. Panorama de potencia instalada en Colombia, Abril 2022.Fuente: PARATEC [2]	10
Ilustración 2. Distribución horaria de potencia eléctrica media de las dos plantas fotoeléctricas .	11
Ilustración 3.Promedios Mensuales de irradiación global media recibida en superficie para principales .....	15
Ilustración 4.Aportes hídricos, precio de bolsa y capacidad de embalses en Colombia 2010-2022. Fuente: Elaboración propia y Datos de XM [13] .....	16
Ilustración 5.Matriz de correlación de características para predicción de potencia.....	20
Ilustración 6.Modelo KNN para generación potencia eléctrica .....	22
Ilustración 7.Modelo de Regresión Lineal para generación potencia eléctrica.....	23
Ilustración 8. Modelo de Random Forest para generación potencia eléctrica. ....	23
Ilustración 9. Modelo de Gradient Boosting para generación potencia eléctrica.....	24
Ilustración 10.Modelo KNN para generación planta 1. ....	25
Ilustración 11.Modelo Random Forest para generación planta 1.....	26
Ilustración 12.Modelo Gradient Boosting para generación planta 1.....	26
Ilustración 13.Modelo SVR para generación planta 1. ....	26
Ilustración 14.Modelo de Regresión Lineal para generación planta 1. ....	26
Ilustración 15.Modelo MLP para generación de potencia eléctrica.....	28
Ilustración 16.Modelo MLP para generación - Planta 1. ....	29
Ilustración 17.Modelo LSTM para generación de potencia.....	29
Ilustración 18. Modelo LSTM para generación - Planta 1.....	30
Ilustración 19.Modelo Convolutacional para generación de potencia. ....	30
Ilustración 20.Modelo Convolutacional para generación - Planta 1. ....	31

## SIGLAS, ACRÓNIMOS Y ABREVIATURAS

<b>FERNC.</b>	Fuente de energía renovables no convencionales
<b>LSTM</b>	Long Short Term Memory
<b>KNN</b>	k-nearest neighbors
<b>SVR</b>	Support vector regression
<b>MLP</b>	Multilayer perceptron
<b>RBF</b>	Radial basis function
<b>RMSE</b>	Root mean square error
<b>FV</b>	Fotovoltaico

---

## RESUMEN

El siguiente documento muestra los resultados de la implementación de diferentes modelos de aprendizaje automático, clásicos y profundos, para el pronóstico de generación de potencia eléctrica en FERNC específicamente de plantas de energía solar ubicadas en la India. Los datos fueron extraídos de dos plantas y corresponden a 34 días. Para este pronóstico de generación de potencia eléctrica se tuvieron en cuenta como características datos de lecturas de sensores en la planta y como salida la generación de energía visto desde el inversor en AC y DC; además el problema fue abordado desde la perspectiva de las series de tiempo.

El trabajo realizado permite predecir la cantidad de energía que será generada por un par de días, lo que conlleva a una mejor gestión de las redes y sistemas de potencia, algo que se ha convertido en un requisito indispensable con el auge de las nuevas tecnologías y los nuevos reglamentos implementados en Colombia.

***Palabras clave*** — **Energía solar, potencia eléctrica, machine learning, deep learning.**



---

## ABSTRACT

The following paper shows the results of the implementation of different machine learning models, classical and deep, for forecasting electric power generation in FERNC specifically from solar power plants located in India. The data were extracted from two plants and correspond to 34 days. For this electric power generation forecast, data from sensor readings in the plant were taken into account as characteristics and as output the power generation seen from the inverter in AC and DC; besides not approaching the basic learning problem as inputs and output only, but to handle it as a time series.

The work performed allows predicting the amount of energy that will be generated for a couple of days, which leads to a better management of networks and power systems, something that has become an indispensable requirement with the rise of new technologies and the new regulations implemented in Colombia.

***Keywords* — Solar energy, electrical power, machine learning, deep learning.**

## I. INTRODUCCIÓN

Más del 80% de la energía consumida en el mundo proviene de recursos no renovables donde sus dependencias claramente son el petróleo, carbón, gas natural, entre otros [1]. Debido a que la cantidad de estos recursos, aunque sean abundantes, tiene un límite, han aparecido tecnologías que aprovechan los recursos naturales de índole renovable que apoyan a la disminución de emisiones de efecto invernadero.

TIPO	FUENTE DE GENERACIÓN	CAPACIDAD EFECTIVA NETA (MW)	PORCENTAJE DE PARTICIPACIÓN
Despachada centralmente	Hidráulica	11,043	62.66%
	Térmica	5,157	29.26%
No despachada centralmente	Hidráulica	902	5.12%
	Térmica	176	1.00%
	Eólica	18	0.10%
	Solar	135	0.77%
	Cogenerador	193	1.09%
<b>Total</b>		<b>17,624</b>	

Ilustración 1. Panorama de potencia instalada en Colombia, Abril 2022. Fuente: PARATEC [2]

Actualmente en Colombia, para el año 2022, se tiene de capacidad efectiva neta de potencia instalada un poco más de 17 GW, donde solo el 0.77% es generación solar y el 63% representa generación hidráulica, las cuales son muy susceptibles a los fenómenos del niño y niña, siendo este primero un evento crítico para la matriz de generación del país debido a la sequía que se puede producir en los embalses de las principales plantas hidráulicas del país. Por tanto, es de vital importancia diversificar la matriz de generación del país, que sea confiable y firme.

Algunas barreras que enfrenta el país para la implementación y aprovechamiento de estas FERNC son:

- Alto costo de inversión por kWp.
- Tener una alta seguridad.
- Obtener bajos costos.

Mientras no sea posible implementar a gran escala sistemas de generación de energía solar existen otras alternativas como los autogeneradores. La autogeneración entendida como *‘Aquella actividad realizada por personas naturales o jurídicas que producen energía eléctrica principalmente, para atender sus propias necesidades’*[3], describe a los autogeneradores o personas que tienen la posibilidad por ley de entregar excedentes de energía a la red, y asimismo, si garantizan energía firme adicional a la que requieren para respaldar su demanda, pueden acceder al pago de cargos por confiabilidad.

Por tanto, es primordial la utilización de técnicas de análisis de información para obtener pronósticos confiables con los que estos generadores tengan una proyección disponible para venta de excedentes o como herramienta para eventos desfavorables contra la planta generadora ocasionados por el tipo de generación que es ‘Solar’. El comportamiento de estos sistemas describe una curva de generación no constante durante el día, provocando que su energía firme generada no sea el pico máximo si no el promedio de esta curva.

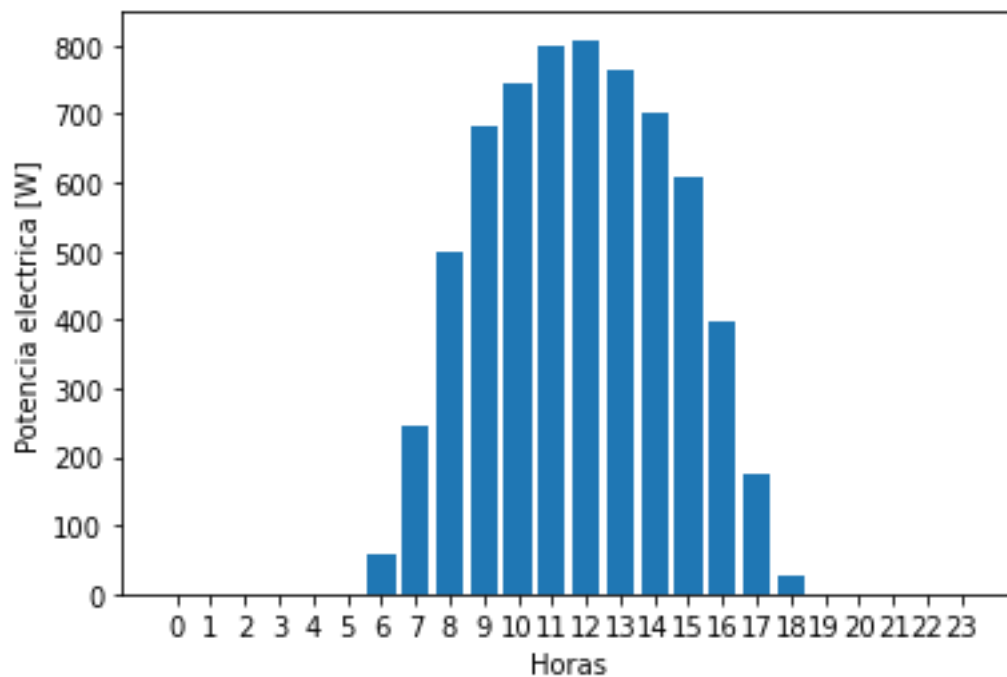


Ilustración 2. Distribución horaria de potencia eléctrica media de las dos plantas fotoeléctricas.

En la figura 2 se puede observar que el pico de generación de energía eléctrica ocurre al medio día y se producen más de 700 W instantánea, esta gráfica se realizó con la base de datos con la que se trabajó y que más adelante se describe.

Teniendo la suficiente información de estas plantas se pueden construir algoritmos capaces de pronosticar la potencia generada durante el día, creando así predicciones que son información valiosa para la toma de decisiones de los autogeneradores.

Este documento muestra la aplicación de diferentes técnicas de aprendizaje automático y profundo en un dataset que contiene información de 34 días registrados de dos plantas de energía diferentes, entre datos de generación y datos de medición obtenidos con sensores de diferentes variables climatológicas [5].

Cada una de las plantas cuenta con dos juegos de datos, uno de ellos tiene información de la generación de energía de la planta y el otro registra mediciones de sensores. Tener modelos que puedan estimar de una manera “precisa” la cantidad de potencia que se puede generar es de gran utilidad para tomar decisiones en torno a la calidad del servicio que se pueda prestar o la misma adquisición o venta de energía.

Se mostrarán las métricas de algunos algoritmos tradicionales y avanzados del aprendizaje automático para poder realizar un análisis comparativo entre ellos y determinar cuáles serían las mejores opciones de implementación.

## II. PLANTEAMIENTO DEL PROBLEMA

Las FERNC particularmente la solar y eólica son de fácil acceso, es decir que están disponibles en cualquier lugar del planeta, generan cero emisiones y son simples en su estructura. Esto ha provocado que gradualmente se acelere la instalación de generadores fotovoltaicos alrededor del mundo, sin embargo, la potencia generada en estos sistemas cambia por naturaleza en función de las variables climáticas del lugar de instalación y esto se convierte en un problema para la operación segura en las redes de potencia.

Debido a que las variables climatológicas son influyentes y altamente correlacionadas con la generación de potencia eléctrica, y además, cambian rápidamente durante el tiempo, se hace necesario diseñar herramientas que permitan describir de alguna manera el comportamiento de estos sistemas de generación tomando en cuenta todas las variables involucradas y el registro disponible en el tiempo.

### III. OBJETIVOS

#### *A. Objetivo general*

Predecir el comportamiento de la generación de potencia eléctrica por horas en un sistema de paneles solares a partir de modelos de aprendizaje de máquina y aprendizaje profundo.

#### *B. Objetivos específicos*

- Realizar búsqueda y preprocesamiento de dataset sobre generación de potencia a partir de energía solar.
- Implementar diferentes algoritmos de aprendizaje automático sobre el dataset utilizado.
- Analizar las métricas de los diferentes modelos para la predicción del comportamiento de generación de potencia.
- Comparar los modelos de aprendizaje clásico y profundo implementados.
- Elegir el modelo que mejor predice el comportamiento de la curva de generación de potencia eléctrica.

### IV. MARCO TEÓRICO

En la actualidad la demanda de energía eléctrica en Colombia busca ser suplida por medio de tecnologías más eficientes que no tengan relación con las fuentes de energía convencionales, que hacen uso de recursos como el petróleo, carbón y el agua [6]. De este modo, surgen las tecnologías que emplean la energía solar y que se presentan como soluciones amigables debido a su facilidad en instalación, y adicionalmente se cuenta con un recurso inagotable que tiende a reducir sus costos al pasar de los años.

Actualmente la capacidad de generación de energía eléctrica en Colombia es de 17326 MW, para suplir una demanda máxima anual de 10700 MW, donde el 82% es de fuentes hidráulicas, el 16% de combustibles fósiles y el 1.14% de Fuentes de Energía Renovables No Convencionales (FERNC), las cuales incluyen la biomasa, la eólica y la solar, por lo que solo se cuenta con aproximadamente 173 MW en el Sistema Interconectado Nacional (SIN)[7], lo cual presenta una proporción muy baja de generación de energía eléctrica haciendo uso del recurso solar.

Colombia cuenta con un promedio de radiación solar de 4.5 kW h/m<sup>2</sup> en todo el territorio [8], correspondiente a una cantidad suficiente para considerarse una potencial generación de energía eléctrica. La figura 4, presenta los promedios mensuales de irradiación en diferentes ciudades del país en el año 2015.

Codigo	Estacion	Municipio	Departamento	Latitud	Longitud	Elevacion (m.s.n.m.)	Entidad	Valor promedio (Wh/m <sup>2</sup> por dia)												Promedio Anual	Años de Informacion	Fecha Inicio	Fecha Final
								ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC				
0027015070	Apto. Olaya Herrera	Medellin	Antioquia	6.22	-75.58	1490	IDEAM (conv.)	4382.6	4409.5	4296.7	4165.2	4050.5	4321.6	4668.1	4605.7	4595.1	4419.0	3958.3	4149.4	4335.1	10	ene-85	jun-97
0037050510	Apto. Santiago Perez	Arauca	Arauca	7.07	-70.73	128	IDEAM (conv.)	5156.3	4484.1	4617.1	4371.2	4292.6	4204.0	4299.1	4416.1	4895.9	4702.5	4786.5	5212.4	4619.8	4	ene-86	ene-92
0020204512	Las Flores	Barranquilla	Atlántico	11.04	-74.82	2	IDEAM (aut.)	6383.8	6369.2	6804.4	6072.2	5971.7	5968.9	6266.6	5594.7	5573.7	5286.9	5399.2	5722.7	5951.2	6	nov-09	dic-14
0014015020	Apto. Rafael Nuñez	Cartagena	Bolivar	10.43	-75.50	2	IDEAM (conv.)	5987.7	6412.9	6378.8	6012.8	4951.4	4988.7	5644.2	5213.1	5291.3	5014.5	4988.2	5748.8	5552.5	7	feb-90	dic-00
0024035130	LPTC	Turija	Boyaca	5.55	-73.35	2690	IDEAM (conv.)	5688.1	5184.2	4695.9	4678.5	4282.1	4092.7	4299.9	4515.6	4926.2	4625.1	4350.7	4550.3	4657.4	6	ene-95	dic-01
0020915020	E.M.A.S.	Manizales	Caldas	5.09	-75.51	2207	IDEAM (aut.)	3943.7	3649.3	3695.8	3857.8	3461.1	3546.0	4074.1	3983.2	4117.1	3720.8	3490.2	3667.0	3767.2	10	may-05	dic-14
0044035050	Mecagual Francia	Risaralda	Cauca	1.50	-75.66	257	IDEAM (aut.)	4063.3	3937.4	3546.6	3465.3	3084.8	3204.2	3127.7	3277.0	3270.0	3269.8	3951.1	3918.9	3617.5	10	jul-05	dic-14
0030521502	Apto. Yopal	Yopal	Casimere	5.32	-72.39	330	IDEAM (aut.)	5765.7	5239.7	4701.4	4569.8	4296.0	4240.9	4186.0	4433.6	4974.8	4885.2	5291.8	5521.0	4835.1	5	nov-09	dic-14
0020803060	Fidencor	Valledupar	Cesar	10.46	-73.25	184	IDEAM (aut.)	5420.1	5685.3	5317.3	5441.3	5253.8	5386.7	5517.8	5375.4	4996.8	4865.5	4950.3	5152.7	5280.4	10	sep-05	dic-14
	Monteria	Monteria	Córdoba	8.81	-75.85	17	FEDERPROZ	4345.6	4389.8	4371.3	4173.9	3973.2	4327.5	4770.2	4429.5	4262.2	3923.6	4018.8	4099.2	4247.1	4	oct-11	abr-14
0021205791	Apto. Eldorado	Bogotá	Cundinamarca	4.71	-74.15	2541	IDEAM (conv.)	4681.9	4312.7	4322.2	3716.7	3506.0	3658.9	3917.3	4168.2	3947.8	3961.0	4017.7	4241.4	4037.7	23	mar-81	dic-04
	Iniria	Puerto Iniridia	Guaina	4.02	-67.67	90	IDEAM (SUTRON)	4500.1	4327.1	3939.4	4140.2	3634.7	3628.7	3542.4	3891.2	4257.1	4117.2	4079.2	4202.1	4021.6	4	feb-97	sep-02
0021115020	Apto. Berito Salas	Neva	Hulla	2.93	-75.28	439	IDEAM (conv.)	4836.0	4700.4	4590.5	4628.9	4552.2	4550.1	4509.7	4656.6	4785.1	4782.3	4607.8	4618.0	4651.5	14	mar-90	ago-03
0015060510	Apto. Almirante Padilla	Riohacha	La Guajira	11.52	-72.92	4	IDEAM (conv.)	5202.8	5556.0	5761.0	5898.3	5618.0	5975.8	6237.6	6045.2	5832.8	5247.8	4977.5	4916.6	5605.8	17	sep-91	mar-14
0000150150	Univ. Tecnológica de Magdalena	Santa Marta	Magdalena	11.22	-74.19	7	IDEAM (aut.)	5539.4	5904.8	5855.5	5756.4	5698.0	5402.9	5370.9	5201.1	5325.3	4721.4	4787.3	5301.5	5405.4	7	ago-07	dic-14
0030903020	Apto. Vanguardia	Villavieja	Meta	4.15	-73.62	423	IDEAM (conv.)	4784.9	4514.5	4337.1	4565.8	4699.2	4650.1	4542.9	4993.2	5307.6	5286.1	4747.4	4580.0	4750.7	14	ene-90	dic-14
0020505010	Bojana	Pasto	Nariño	1.16	-77.28	2620	IDEAM (aut.)	3749.1	3499.2	3497.4	3698.9	3695.8	3715.9	3877.4	4006.3	4124.7	3988.8	3765.2	3475.0	3742.8	10	may-05	abr-03
0018015010	Apto. Camilo Daza	Cúcuta	Norte de Santander	7.92	-72.50	250	IDEAM (conv.)	4277.9	4116.4	4177.9	4104.3	4539.4	4487.7	4632.4	4147.4	4932.2	4737.2	4374.1	4167.3	4443.8	12	ago-89	nov-13
0026125090	Armenia	Armenia	Quindío	4.53	-75.69	1458	IDEAM (aut.)	3918.2	3837.4	3919.7	3857.4	3691.9	3866.7	4265.6	4175.5	4333.9	3893.8	3879.2	3567.8	3933.8	10	dic-05	nov-86
0026135040	Apto. Matecaña	Pareira	Risaralda	4.80	-75.73	1342	IDEAM (conv.)	4279.4	4406.3	4283.9	4099.7	3905.1	3940.5	4243.6	4362.0	4273.2	4338.6	4183.3	4315.4	4210.9	7	oct-90	oct-13
0017015010	Apto. Sesquicentenario	San Andrés	San Andrés y Providencia	12.58	-81.70	1	IDEAM (conv.)	4422.2	5166.1	5733.2	5967.8	5025.4	4705.8	4914.2	4868.8	4753.2	4430.9	3747.8	4045.4	4818.3	3	ene-01	dic-14
0025025070	Unsucre (Puerta Roja)	Sincolajo	Sucre	9.20	-75.39	221	IDEAM (aut.)	4843.9	4896.3	4733.4	4420.1	3860.0	4411.8	4600.9	4354.3	4233.7	3929.5	3733.4	4309.1	4368.0	10	may-05	dic-99
0021245040	Apto. Perales	Baquijé	Tolima	4.42	-75.13	928	IDEAM (conv.)	4615.6	4578.8	4621.2	4651.4	4627.9	4717.9	4896.1	4986.2	4846.8	4679.7	4404.6	4332.7	4663.2	9	nov-89	dic-14
0002605007	Univalle	Calí	Valle del Cauca	3.38	-76.53	992	IDEAM (aut.)	4385.4	4360.8	4373.0	4303.9	4138.2	4299.1	4628.5	4843.8	4631.4	4256.1	3998.5	3971.9	4332.6	9	nov-06	dic-14

Ilustración 3. Promedios Mensuales de irradiación global media recibida en superficie para principales.

Los beneficios de esta fuente de energía son demasiados, uno de ellos está relacionado con la economía. Si se llega a incrementar la generación de esta fuente de energía, se tendrá una menor demanda de esta en todo el SIN, provocando así una reducción en el precio de bolsa de energía eléctrica [ $\frac{COP}{kWh}$ ]; en la actualidad el precio fluctúa con relación a los recursos hidráulicos [10], lo cual permite tener seguridad en el consumidor, ya que es posible pronosticar la cantidad de generación, de manera que se pueda prever el ahorro monetario esperado. En la figura 4, se puede observar en color rojo el precio de bolsa de la energía eléctrica que varía de acuerdo con los embalses en las centrales hidroeléctricas del país, mostrados en color azul, los cuales tienen un comportamiento establecido, que permite identificar la variación de los precios.

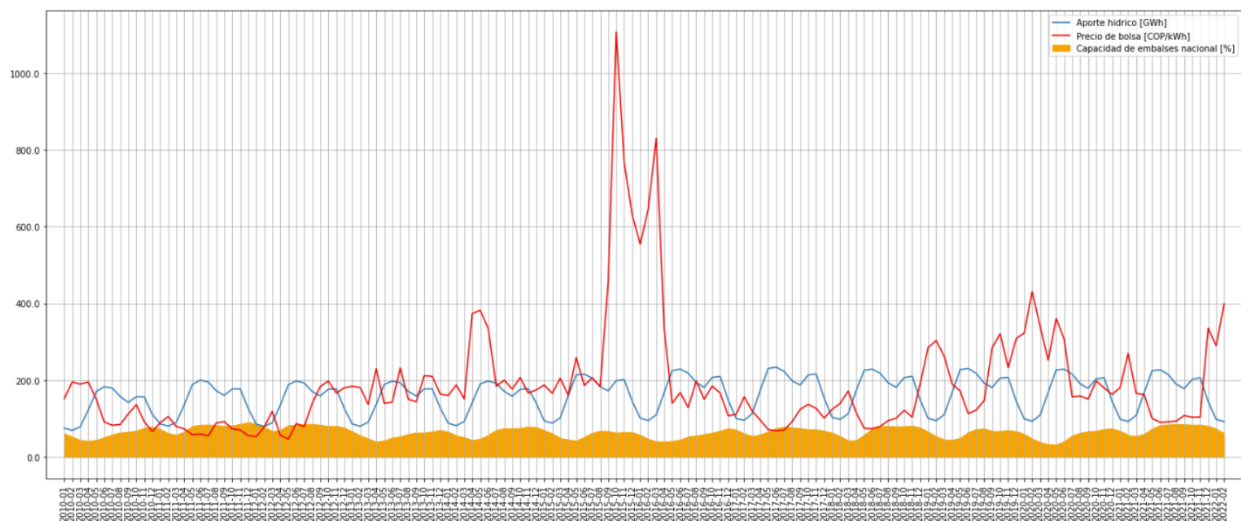


Ilustración 4. Aportes hídricos, precio de bolsa y capacidad de embalses en Colombia 2010-2022. Fuente: Elaboración propia y Datos de XM [13]

Cabe resaltar que el principal beneficio de la energía solar es que es gratuita, además cualquier persona natural o jurídica puede generar energía eléctrica con sistemas FV (Fotovoltaicos) aumentando así la independencia energética y propiciando formas de pronóstico de la generación o de retorno monetario, actividades de vital importancia para brindar seguridad a las personas que hagan uso de esta tecnología debido a que, si se produce un racionamiento energético o caída del circuito del operador de red, estos aun podrán sostenerse por medio de la autogeneración con ayuda de baterías.

Para realizar estos pronósticos se hace uso de variables climatológicas como la radiación solar, las cuales pueden cambiar continuamente, como cuando las nubes oscurecen el sol, en este caso



la generación tiende a disminuir. Esta fluctuación trae una complejidad en la operación cuando los sistemas FV sean de una capacidad financiera y potencial altos (sistemas de potencia eléctrica), como lo es el deslastre de este sistema de generación y la disminución de la frecuencia de la red a la cual esté conectada, causando la caída de todo el sistema de potencia [12].

Debido a su importancia y al impacto que genera en la sociedad, el problema de la generación de energía solar ha sido ampliamente abordado desde el aprendizaje automático o machine learning utilizando técnicas de forecasting y aplicando los diferentes modelos clásicos y de aprendizaje profundo disponibles. Ejemplo de ello es la identificación de zonas potencialmente aptas para establecer plantas generadoras mediante la construcción de modelos predictivos basados en máquinas de vectores de soporte (SVR) y perceptrón multicapa (MLP) utilizando registros de medición de la radiación solar, diferentes características meteorológicas y variables de latitud y longitud para determinar ubicaciones [15], en estas implementaciones se han obtenido resultados bastante aceptables. Esta necesidad de crear modelos predictivos mediante data registrada de radiación solar surge del altísimo costo de los equipos de medición de esta, además del costo derivado del mantenimiento de estos equipos [16].

Otro ejemplo de aplicación de machine learning en predicción de energía solar se llevó a cabo en Turquía, donde mediante data registrada de diferentes provincias, con variables como la radiación solar, temperatura ambiente entre otras se implementaron modelos utilizando técnicas como SVR, método de los k vecinos más cercanos (KNN), redes neuronales y aprendizaje profundo. Mediante las métricas obtenidas pudieron determinar que los mejores resultados se lograron con redes neuronales y aprendizaje profundo, en general obtuvieron buenas predicciones [17].

También se han empleado en predicción de generación de potencia mediante paneles solares, modelos específicos de aprendizaje profundo como redes LSTM para tratamiento de series temporales o forecasting, así como redes neuronales y convolucionales, y también la creación de redes híbridas LSTM-convolucional que dan paso a modelos mucho más robustos y con excelentes resultados de predicción [18].

## V. METODOLOGÍA

Se realizó la búsqueda de bases de datos estructuradas relacionadas con la generación de potencia a partir de paneles solares. Se seleccionó la base de datos gratuita “Solar Power Generation Data” de Ani Kannal de la comunidad de científicos de datos y profesionales de Deep Learning, Kaggle [14]. La base de datos corresponde a información registrada de 34 días, del 15/05/2020 al 17/06/2020 en intervalos de 15 minutos, de dos plantas generadoras, donde cada planta tiene dos documentos, el primero en generación de las unidades y el segundo variables fenomenológicas tomadas por sensores del estado de las celdas fotovoltaicas.

A continuación se describen las variables de la base de datos:

### A. Variables del campo de generación

- DATE TIME: Fecha y hora de cada observación, fue registrada en intervalos de 15 minutos.
- PLANT ID: identificador de planta, es un dato común en todo el archivo.
- SOURCE KEY: Identificador del inversor.
- DC POWER: Cantidad de potencia DC generada en el intervalo de 15 minutos en KW.
- AC POWER: Cantidad de potencia AC generada en el intervalo de 15 minutos en KW.
- DAILY YIELD: Suma acumulada de potencia generada hasta ese momento.
- TOTAL YIELD: Rendimiento total del inversor hasta ese momento.

### B. Variables del campo de medición de sensores

- DATE TIME: Fecha y hora de cada observación, fue registrada en intervalos de 15 minutos.
- PLANT ID: identificador de planta, es un dato común en todo el archivo.
- SOURCE KEY: Identificador de panel de sensores, también es un dato común para el archivo.
- AMBIENT TEMPERATURE: Temperatura ambiente en la planta.

- MODULE TEMPERATURE: Temperatura del panel solar conectado al sensor.
- IRRADIATION: Cantidad de irradiación en el intervalo de 15 minutos.

Teniendo los datasets a disposición se realiza una exploración inicial de los datos; en estos no se encuentran datos nulos y no se detectan tantos valores atípicos que puedan afectar la posterior construcción de modelos predictivos, por lo que se toma la decisión de no realizar un tratamiento para estos. Se eliminan de los datasets los campos SOURCE KEY y PLANT ID ya que solo identifican paneles y plantas (solo 2) y no aportan información relevante para los modelos predictivos.

Para poder proseguir con el preprocesamiento de los datos, se unen los datasets de cada una de las plantas, es decir, generación de energía planta 1 con mediciones de sensores planta 1; dicha unión se realiza por el campo común DATE TIME, de igual manera se realiza para la planta 2. Finalmente se concatenan los datos de las 2 plantas para obtener un único dataset.

Debido a que la energía solar tiene una gran dependencia de las condiciones climáticas que pueden afectar en varias unidades el valor a predecir, resulta fundamental filtrar las características, para de esta forma obtener la previsión de generación de energía; debido a esto se aplican los métodos de filtrado para la selección de características como: [1]

- Correlación de Pearson (Algoritmo simple).
- Algoritmo de Relief (Algoritmo robusto de uso común).
- Algoritmo de análisis de información local (Método avanzado).

Teniendo el insumo final más las técnicas de filtrado anteriormente propuestas, se construye la matriz de correlación con las variables que en primera instancia servirán de entrada a los modelos que se crearán posteriormente, para este caso se excluyen del dataset las variables DATE TIME y AC POWER, ésta última es la variable a predecir, y se emplean las demás para la matriz como se puede observar en la figura 5.

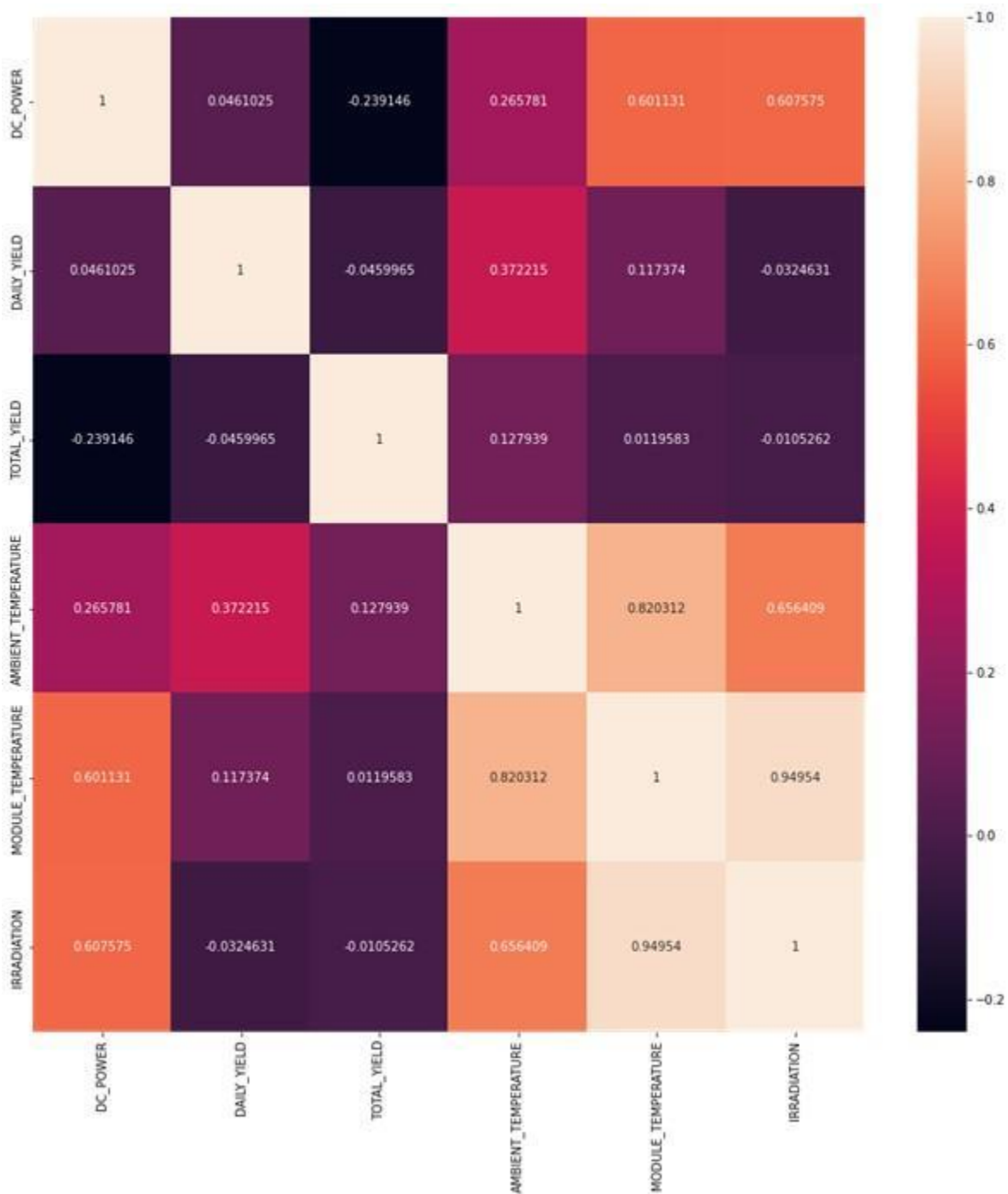


Ilustración 5. Matriz de correlación de características para predicción de potencia.

Con base en la matriz de correlación se toma la decisión de no emplear la variable MODULE TEMPERATURE debido a su valor de correlación de 0.95 con la variable IRRADIATION, por lo que no se requieren las dos y se elige trabajar solo con la segunda. También se decide prescindir de la variable DC POWER teniendo en cuenta que esta y AC POWER indican

---

potencia generada y ambas podrían ser consideradas como variable objetivo, por tanto, se debe descartar una, adicionalmente la variable más importante es la AC debido a que es a la que se conecta con el SIN (Sistema Interconectado Nacional).

En este punto se tiene listo el dataset para comenzar a realizar los experimentos de predicción de la potencia AC en el tiempo empleando los diferentes modelos de regresión disponibles para tal fin y poder comparar las salidas y métricas de dichos modelos.

seguidamente se realiza la implementación de los principales modelos clásicos disponibles y algunos de deep learning. Para el modelo KNN se fijaron los parámetros en distancia Manhattan y número de vecinos  $K=3$ . En el modelo Random Forest se configuró el parámetro de número de estimadores en 8. Para Gradient Boosting se eligieron parámetros de loss = 'squared error' y número de estimadores = 500.

El modelo SVR solo se entrenó con el dataset de la planta 1 y los parámetros configurados fueron kernel RBF, valor  $C = 200$  y  $\gamma = 0,0001$ . También se realizaron pruebas con el modelo de regresión lineal. Con el Perceptrón multicapa se emplearon 2 capas ocultas y se entrenaron modelos con diferente número de neuronas en las capas para encontrar la mejor configuración.

Finalmente se aplicaron redes de deep learning LSTM y convolucional. Para el modelo LSTM se utilizaron 4 capas con 64 celdas en cada una y para el modelo convolucional se emplearon 3 capas Conv1D con 64 filtros y  $\text{kernel}=3$ .

Para cada uno de los modelos se realizaron experimentos utilizando el dataset que contenía los datos de planta 1 y planta 2 y también ejecuciones empleando únicamente el dataset de planta 1 para comparar los resultados.

## VI RESULTADOS

### A. Modelo KNN

El primer modelo empleado fue KNN, se utilizó el método GridSearchCV en la búsqueda de los mejores hiper parámetros posibles entre las medidas de distancia Euclidiana, Manhattan y Chebyshev y para diferentes valores de K, también empleando metodología de cross validación KFold. Se obtuvo que los mejores hiper parámetros para la creación de un modelo KNN serían distancia Manhattan y número de vecinos  $K = 3$ . Luego de crear el modelo con estos parámetros y realizar la predicción, se obtuvo una varianza explicada de prueba de aproximadamente 0.89, lo que supone una buena métrica para una primera iteración. En la gráfica 6 se observa la salida del modelo o predicción contra el tiempo en comparación con la variable de prueba:



Ilustración 6. Modelo KNN para generación potencia eléctrica.

### B. Modelo de Regresión Lineal

En este modelo más básico aún, también se empleó metodología KFold, y luego de realizar la predicción con las variables de prueba se obtuvo una varianza explicada de 0.74. En la gráfica 7 se puede apreciar la potencia generada en contraste con la predicción para este modelo:



Ilustración 7. Modelo de Regresión Lineal para generación potencia eléctrica.

### C. Modelo Random Forest

Para este modelo no fue posible emplear el método GridSearchCV en la búsqueda de los mejores hiper parámetros debido al costo computacional, sin embargo, mediante el ajuste del mismo, realizando entrenamiento con diferente número de estimadores fue fácil identificar que luego de 8 la variación de la métrica empleada era mínima y utilizar más árboles sería innecesario por lo que se fija este parámetro en 8. Con este se hace la cross validación y luego de realizar la respectiva predicción se obtiene una varianza explicada de 0.94 y el comportamiento descrito en la figura 8:



Ilustración 8. Modelo de Random Forest para generación potencia eléctrica.

### D. Modelo Gradient Boosting

En este método el costo computacional es aún mayor, por lo que tampoco se empleó GridSearchCV, aunque de la misma forma que en el modelo Random Forest, mediante el

ajuste en el entrenamiento utilizando el parámetro  $loss='squared\ error'$  se observó que luego de 500 estimadores la varianza explicada cambiaba mínimamente por lo que se optó por fijar  $n$  estimators en este valor. Igual que en los modelos previos se utilizó KFold como validación, la varianza explicada luego de la prueba fue de 0.92 aproximadamente. La potencia AC generada VS la predicción de esta para este modelo se puede observar en la figura 9:

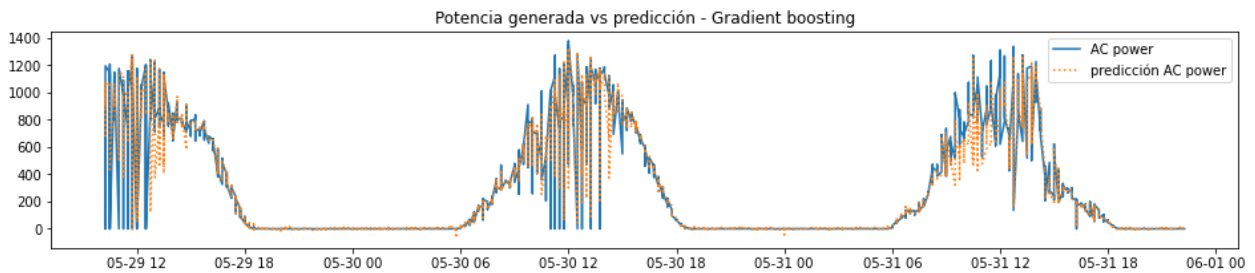


Ilustración 9. Modelo de Gradient Boosting para generación potencia eléctrica.

A modo de comparación en la tabla 1 se recopilan los 4 modelos de regresión utilizados hasta el momento en donde se compara la métrica de validación con la de test de cada uno, la métrica principal utilizada es la varianza explicada.

Tabla 1. Resultado de métricas iniciales de modelos clásicos.

Modelo	Score validación	Score test
Random Forest	0.95	0.94
Gradient Boosting	0.93	0.93
KNN	0.88	0.89
Regresión Lineal	0.76	0.74

De la anterior tabla se puede afirmar que para cada uno de los modelos el score de validación y el de prueba son muy parecidos, por lo que en teoría ninguno de ellos tendría problemas de sobreajuste evidentes.



También en la tabla 2 se registraron otras métricas de regresión diferentes a la varianza explicada para comparar cada modelo como lo son el mean squared error,  $R^2$  score entre otras:

Tabla 2. Resultado de métricas totales de modelos clásicos.

Modelo	Exp_var	RMSE	Max_error	$R^2$	MedAE
Random Forest	0.94	89.25	1303.38	0.94	0.23
Gradient Boosting	0.93	102.87	1093.64	0.93	5.3
KNN	0.89	126.87	1249.30	0.89	2.13
Regresión Lineal	0.74	191.41	1165.06	0.74	50.10

Al ver que las gráficas de la variable objetivo presentan varias caídas de potencia a 0 en horas del día de operación normal se pudo constatar que estos valores provienen del dataset correspondiente a la planta 2, y al graficar la salida solo para la planta 1 se observa que casi no se encuentran estos valores de potencia 0. Dado este caso y teniendo en cuenta que el dataset de planta 1 aún ofrece una información considerable se decide realizar los mismos experimentos con los mismos parámetros fijados anteriormente para comparar los resultados.

En estos experimentos se adiciona el modelo SVR que no se pudo ejecutar con los datasets de las 2 plantas unidos debido al costo computacional. Se utilizó kernel RBF, parámetro de regularización C de 200 y valor gamma de 0,0001. Los resultados obtenidos con la implementación de este modelo no fueron tan buenos como los anteriores y adicionalmente el entrenamiento de este sigue siendo costoso. Las figuras 10, 11, 12, 13 y 14 muestran la potencia AC generada VS la predicción para los modelos KNN, Random Forest, Gradient Boosting, SVR y Regresión Lineal respectivamente:

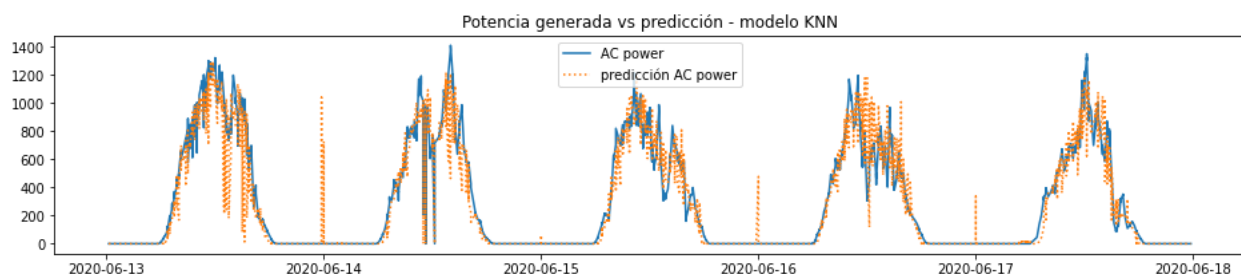


Ilustración 10. Modelo KNN para generación planta 1.

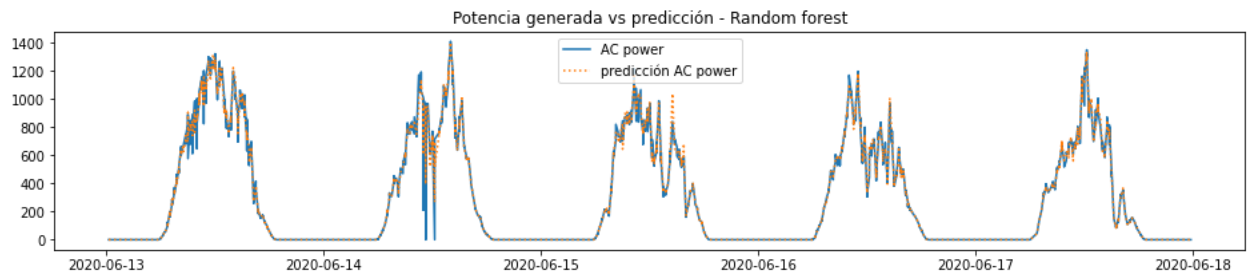


Ilustración 11. Modelo Random Forest para generación planta 1.

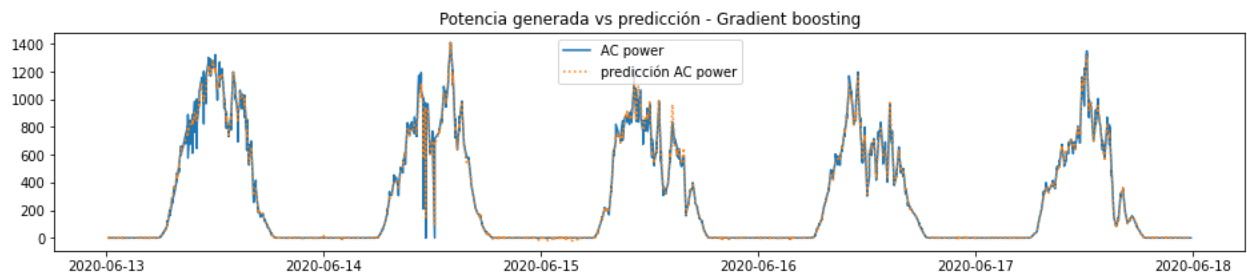


Ilustración 12. Modelo Gradient Boosting para generación planta 1.

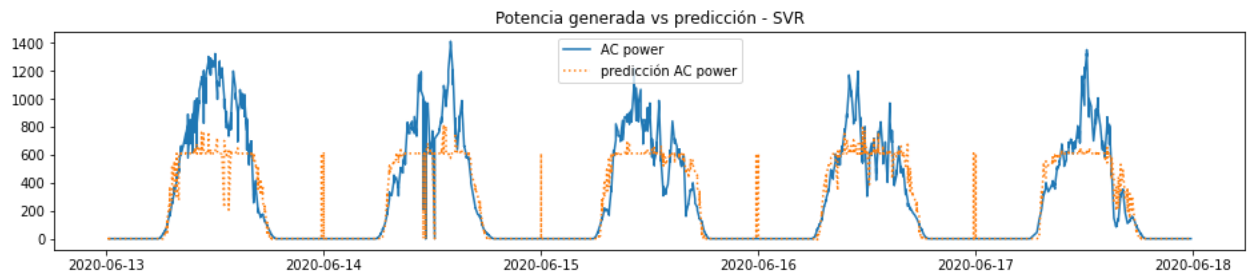


Ilustración 13. Modelo SVR para generación planta 1.

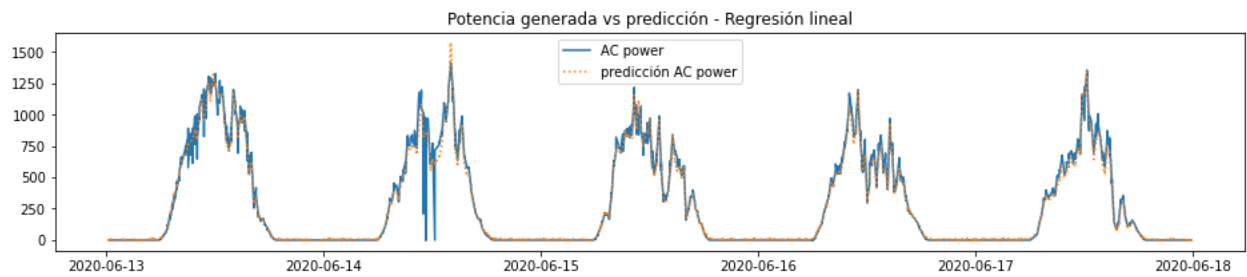


Ilustración 14. Modelo de Regresión Lineal para generación planta 1.

De igual manera se comparan las métricas iniciales de varianza explicada en validación y test para cada uno de los modelos y se recopilan métricas adicionales, esta información se visualiza en las tablas 3 y 4 respectivamente:

Tabla 3.Resultado de métricas iniciales para planta 1

Modelo	Score validación	Score test
Gradient Boosting	0.99	0.99
Random Forest	0.98	0.98
Regresión Lineal	0.98	0.98
KNN	0.88	0.89
SVR	0.74	0.77

Tabla 4.Resultado de métricas totales planta 1

Modelo	Exp_var	RMSE	Max_error	$R^2$	MedAE
Gradient Boosting	0.99	38.55	1024.02	0.99	2.38
Random Forest	0.98	40.27	973.26	0.98	0.40
Regresión Lineal	0.98	49.88	1074.43	0.98	9.71
KNN	0.89	130.18	1222.69	0.89	9.34
SVR	0.77	197.21	805.92	0.7	13.95

### E. Modelo Perceptrón Multicapa (MLP)

Luego de explorar los modelos clásicos pasamos a realizar experimentos con el perceptrón multicapa, se emplearon 2 capas ocultas. Se realizaron varios entrenamientos en los que se probó cambiando la cantidad de neuronas de cada capa y no se obtuvo una mejoría considerable en la métrica de error empleada, que en este caso fue RMSE. Se incorporaron capas de batchnormalization ya que en algunas ejecuciones el modelo no era capaz de entrenar. También se empleó técnica de regularización L2 para reducir el sobreajuste. El mejor resultado se obtuvo con 6 neuronas en la primera capa y 4 en la segunda con un RMSE de 135.36, se utilizó función de activación relu en las capas. En la figura 15 se observa la salida del modelo VS la potencia generada:

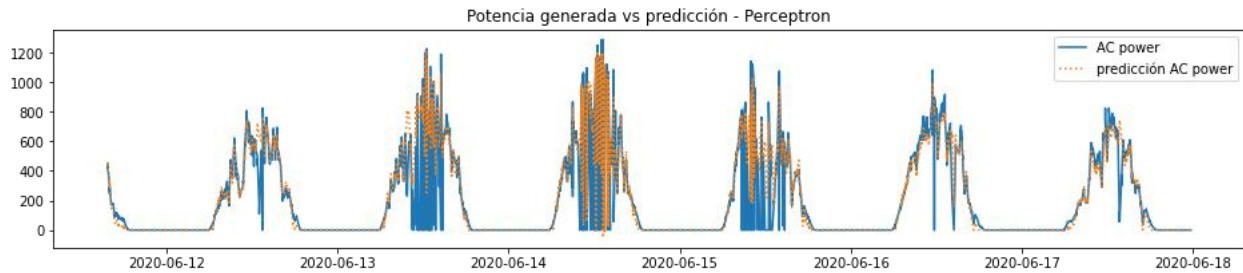


Ilustración 15. Modelo MLP para generación de potencia eléctrica.

En la tabla 5 se recopilaron algunos de los resultados de error obtenidos en los experimentos realizados para diferentes valores de neuronas en capa 1 y capa 2:

Tabla 5. Experimentos Perceptrón.

# Neuronas capa 1	# Neuronas capa 2	RMSE
6	4	135.36
7	5	139.78
4	6	143.67
6	5	148.64
5	3	149.07
6	3	155.59
5	4	155.85
2	3	157.57

De igual manera se realizó prueba utilizando únicamente el dataset de la planta 1 pero solo se empleó la configuración de 6 y 4 neuronas en la capa 1 y 2 respectivamente con la que se obtuvo el mejor resultado de los experimentos anteriores. En este caso el resultado fue mucho mejor, obteniendo un RMSE de 72,66. En la figura 16 se muestra la correspondiente predicción para este caso:

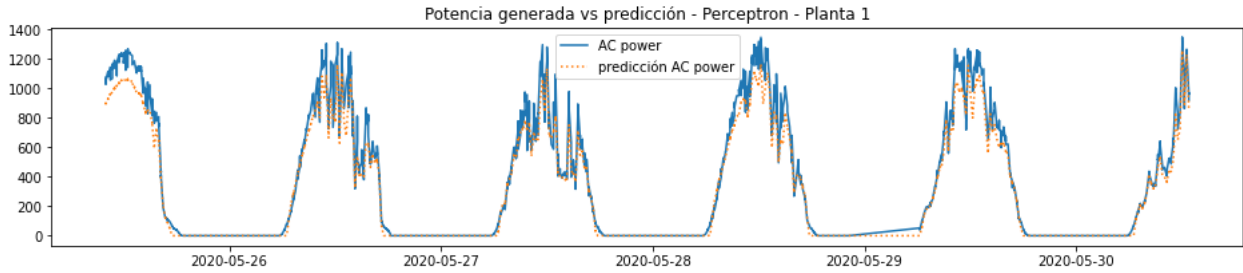


Ilustración 16. Modelo MLP para generación - Planta 1.

## F. LSTM

Para la implementación de este modelo se utilizaron 4 capas LSTM con 64 celdas cada una, con capas dropout intermedias para regularización. Se obtuvo en este caso un RMSE en test de 112,19. En la figura 17 se observa la salida generada por este modelo:

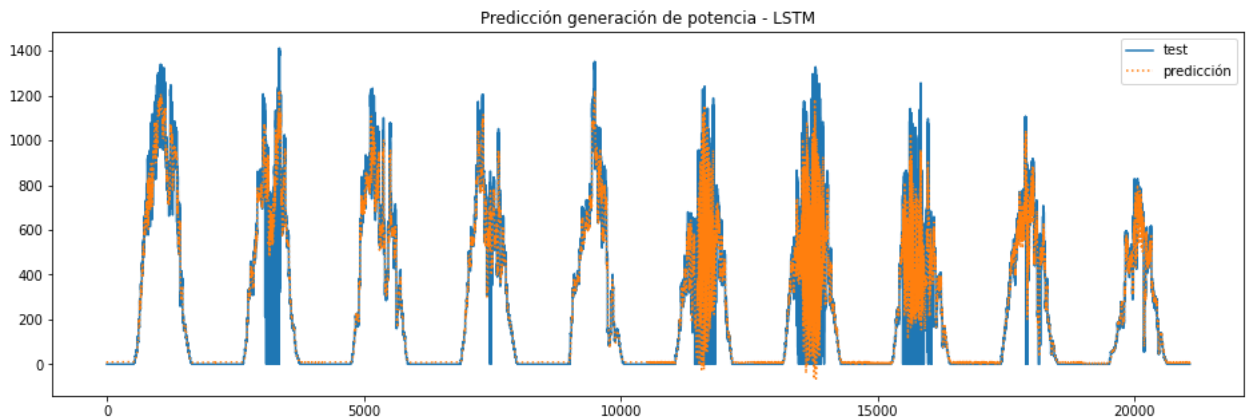


Ilustración 17. Modelo LSTM para generación de potencia.

Se emplea la misma arquitectura LSTM utilizando solo el dataset de la planta 1 y se obtiene un valor RMSE de 62,52. El gráfico de la predicción se observa en la figura 18:

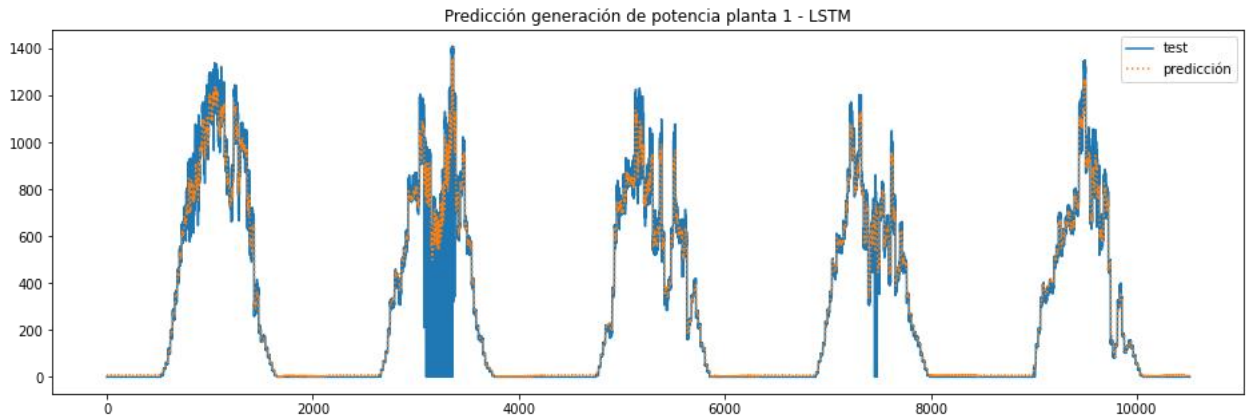


Ilustración 18. Modelo LSTM para generación - Planta 1.

### G. Red Convolucional

Como último experimento para esta serie de tiempo se empleó una arquitectura de capas convolucionales 1D. Se utilizaron 3 capas Conv1D con 64 filtros y kernel de 3 en cada una de ellas, adicionalmente capas intermedias dropout. Para esta configuración se obtuvo un RMSE de prueba de 83,06. La figura 19 muestra la predicción realizada por este modelo:

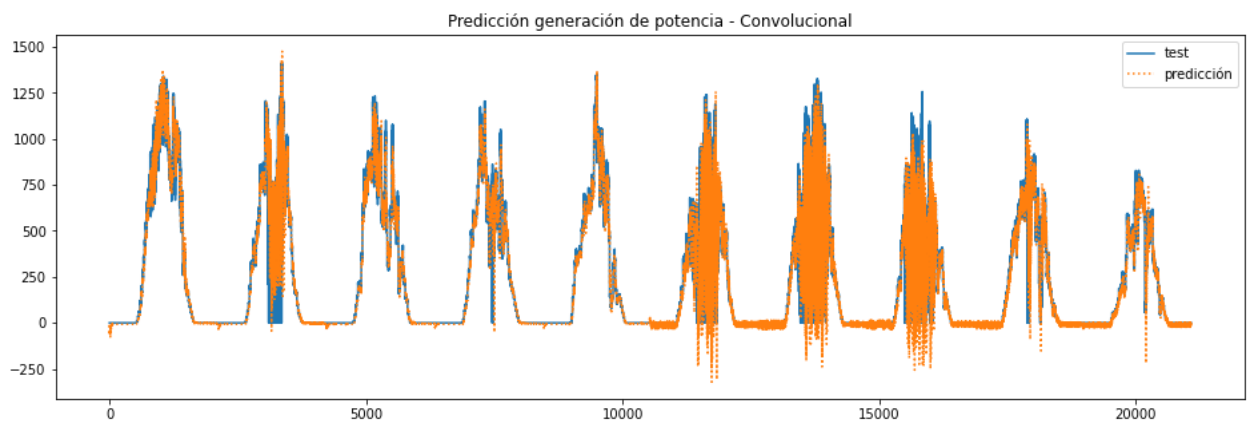


Ilustración 19. Modelo Convolucional para generación de potencia.

Para el mismo modelo entrenado solo con el dataset de la planta 1 se obtuvo un RMSE de 54,40 y en la figura 20 se puede ver la salida generada por el mismo:

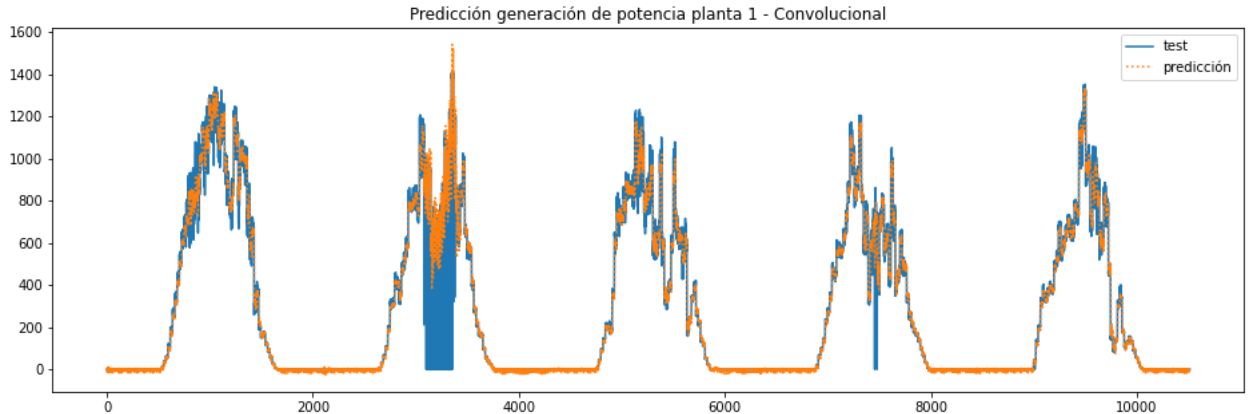


Ilustración 20. Modelo Convolucional para generación - Planta 1.

En la tabla 5 se presentan los resultados de predicción obtenidos en la implementación de los modelos, tanto clásicos como de deep learning empleando los datasets de las plantas 1 y 2, tomando como métrica de referencia el valor RMSE:

Tabla 6. RMSE de modelos clásicos y deep learning.

Modelo	RMSE
CONV1D	83.06
Random Forest	89.25
Gradient Boosting	102.9
LSTM	112.2
KNN	126.9
Perceptrón	135.4
Reg. Lineal	191.4

Se obtuvo el mejor resultado con el modelo de deep learning de capas convolucionales 1D, con un RMSE de 83,06, aunque con el modelo clásico Random Forest, se logró una métrica de error bastante cercana de 89,25, lo que supondría una buena opción de implementación pensando en la diferencia de costo computacional entre las 2 arquitecturas. También se observa que la implementación LSTM, cuya arquitectura es especial para problemas de series de tiempos, quedó por debajo de los modelos basados en árboles, claro está que esta métrica de error podría ser mejorada con la ejecución de más experimentos adicionando más capas LSTM y entrenando

hasta obtener el mejor RMSE, esto mismo se podría lograr con el modelo convolucional adicionando capas CONV1D.

En la tabla 6 se registran los resultados de los experimentos realizados con las mejores configuraciones obtenidas para los mismos modelos anteriormente mencionados, pero solo utilizando el dataset de la planta 1 con igual métrica de referencia RMSE:

Tabla 7. Mejores resultados planta 1.

Modelo	RMSE
Gradient Boosting	38.55
Random Forest	40.27
Reg. Lineal	49.88
CONV1D	54.40
LSTM	62.52
Perceptrón	72.66
KNN	130.18
SVR	197.21

Claramente se puede apreciar que solo con la data de la planta 1, las métricas de error para los modelos en general tienen una gran mejoría. En este caso los experimentos arrojaron mejores valores de RMSE para los modelos basados en árboles Gradient Boosting y Random Forest.

teniendo en cuenta los resultados obtenidos en los diferentes experimentos realizados, se podría afirmar que los modelos basados en árboles serían una buena opción para la implementación y despliegue de un modelo de predicción para generación de potencia eléctrica en días próximos para este problema específico, sin embargo, como se mencionó anteriormente, los modelos de deep learning LSTM y Convolucional podrían ser mejorados notoriamente con más tiempo de entrenamiento y experimentación para crear configuraciones que describen mucho mejor el problema de serie de tiempo abordado. También se podría considerar en una implementación de este problema específico la utilización únicamente del dataset de la planta 1, ya que, si bien representa la utilización de menos datos, aún sigue siendo una cantidad considerable para trabajar y supone la obtención de mejores resultados.



---

## VII. CONCLUSIONES

Luego de realizar la implementación de varios de los modelos de aprendizaje automático disponibles, clásicos y de aprendizaje profundo, aplicándolos al problema de generación de potencia de plantas solares, se pudo evidenciar que es posible desarrollar modelos que se ajustan muy bien al comportamiento esperado de este tipo de sistemas y pueden ser una herramienta muy útil para la toma de decisiones respecto al aprovechamiento de la energía, claro está que los datos empleados para entrenar los modelos pueden tener cierta cantidad de características altamente correlacionadas con la variable de salida, por lo que se hace necesario realizar una depuración empleando la matriz de correlación por ejemplo, y adicionalmente excluir variables que pueden ser consideradas como variable objetivo también, en este caso puntual, la potencia DC.

Aunque inicialmente se optó por emplear como métrica la varianza explicada, se pudo visualizar más claramente el desempeño de cada uno de los modelos implementados mediante la obtención de métricas adicionales como el RMSE. Gracias a estas, se pudo establecer que, con los modelos clásicos como Random Forest y KNN, aún se pueden lograr resultados de predicción bastante buenos para series de tiempo, incluso mejores que los modelos de deep learning, esto constituye una gran ventaja para situaciones en las que se requiere implementar modelos no tan simples y se cuenta con menos recursos.

También se pudo evidenciar, que las caídas de potencia presentes en la serie de tiempos abordada aumentan notablemente los valores de error de las métricas utilizadas, por lo que puede ser una buena opción realizar un tratamiento especial para estos datos específicos u optar por crear los modelos con solo una parte de la data, que de igual manera representa un porcentaje considerable para poder desarrollar modelos predictivos con resultados favorables.

A diferencia de los modelos clásicos, quizás la parte más compleja en la implementación de modelos de deep learning, para trabajar con series de tiempos, sean redes LSTM o convolucionales, es la preparación de la data antes del entrenamiento, se debe tener mucha

precaución cuando se está realizando la división del dataset en los diferentes conjuntos de entrenamiento, validación y test y estar constantemente validando las dimensiones para no tener resultados erróneos.

En general se lograron buenos resultados en todos los modelos implementados, exceptuando el SVR, en cuanto a la predicción del comportamiento de la curva de generación de la potencia. Sin embargo, los que mejor ajustaron su predicción al comportamiento requerido, fueron los modelos basados en árboles Random Forest y Gradient Boosting, aunque como ya se reiteró, modificando la arquitectura de los modelos de deep learning y realizando más experimentos podrían obtenerse los mejores resultados.

## REFERENCIAS

- [1] "Renewables 2014 Global Status Report" 2014. [En línea]. Disponible en: <https://www.ren21.net/wp-content/uploads/2019/05/GSR2014-Full-Report-English.pdf>.
- [2] "PARATEC" 2022. [En línea]. Disponible en: <https://paratec.xm.com.co/paratec/SitePages/generacion.aspx?q=capacidad>.
- [3] "RESOLUCIÓN No.030 DE 2018" 2018. [En línea]. Disponible en: <http://apolo.creg.gov.co/Publicac.nsf/1c09d18d2d5ffb5b05256eee00709c02/83b41035c2c4474f05258243005a1191>.
- [4] "Agencia Provincial de la Energía"2011.[En línea]. Disponible en: <http://alicantenergia.es/content/graficas-de-produccion-de-energia-fotovoltaica-a-tiempo-real/>.
- [5] "Solar Power Generation Data" 2022.[En línea]. Disponible en: <https://www.kaggle.com/anikannal/solar-power-generation-data>.
- [6] "XM - Compañía de Expertos en Mercados S.A. ESP - Demanda de Energía Nacional," 2015. [En línea]. Disponible en: <http://informesanuales.xm.com.co/2015/SitePages/operacion/3-1-Demanda-de-energia-nacional.aspx>.
- [7] "Documento de trabajo sobre el Sistema Interconectado Nacional, SIN" 2018.[En línea].Disponible en: <https://www.celsia.com/wp-content/uploads/2021/02/Documento-de-trabajo-sobre-el-Sistema-Interconectado-Nacional.pdf>.
- [8] C. A. Toledo Arias and A. Urbina Yeregui, "Evaluación de la energía solar fotovoltaica como solución a la dependencia energética de zonas rurales de Colombia," p. 1–49, Septiembre, 2013.
- [9] "Promedios mensuales de irradiación global media recibida en superficie para las principales ciudades del país ( $\frac{kWh}{m^2}$  por día)" 2018.[En línea].Disponible en: <http://atlas.ideam.gov.co/basefiles/AnexoPromediosmensualesdeIrradiacionGlobalMedia.pdf>.
- [10] "¿A qué se debe el alza en el precio de la energía en Colombia?"2022.[En línea].Disponible en: <https://www.valoraanalitik.com/2022/01/11/a-que-se-debe-alza-en-el-precio-energia-colombia/>.
- [11] "API-XM"2020. [En línea].Disponible en: [https://github.com/EquipoAnaliticaXM/API\\_XM](https://github.com/EquipoAnaliticaXM/API_XM).
- [12] Gómez, Marco Antonio "Pronóstico de la generación eléctrica de sistemas fotovoltaicos. Un inicio en cuba desde la universidad"2021. [En línea].Disponible en: [http://scielo.sld.cu/scielo.php?script=sci\\_arttext&pid=S221836202021000100253](http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S221836202021000100253).
- [13] Muñoz, A "Modelos de predicción a corto plazo de la generación eléctrica en instalaciones fotovoltaicas" 2014.[En línea].Disponible en: <https://dialnet.unirioja.es/descarga/tesis/44089.pdf>.

- 
- [14] “Kaggle” 2022.[En línea].Disponible en: <https://www.kaggle.com/anikannal/solar-power-generation-data>.
- [15] R. Martin, R. Aler, J. M. Valls and I. M. “GalvanMachine learning techniques for daily solar energy prediction and interpolation using numerical weather models”. 2015.[En línea]. <https://10.1002/cpe.3631>
- [16] Yu Fenga, Weiping Haoa, Haoru Lia, Ningbo Cuib, Daozhi Gongga,Lili Gaoa, “Machine learning models to quantify and map daily global solar radiation and photovoltaic power”.2019.[En línea]. <https://doi.org/10.1016/j.rser.2019.109393>
- [17] Ümit A ğbulut a, Ali Etem Gürel a,b,Yunus Biçen, “Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison”.2020.[En línea]. <https://doi.org/10.1016/j.rser.2020.110114>
- [18] Rui Chang a, Lei Bai a, Ching-Hsien Hsu, “Solar power generation prediction based on deep Learning”.2021.[En línea]. <https://doi.org/10.1016/j.seta.2021.101354>