PROYECTO FIN DE CARRERA

Presentado a

LA UNIVERSIDAD DE LOS ANDES
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA Y ELECTRÓNICA

Para obtener el título de

INGENIERO ELÉCTRICO

por

Andrés Felipe Zambrano Jacobo

Predicción de Irradiación Solar en Colombia a partir de Mediciones en Sitio y Satelitales Utilizando Redes Neuronales

---

Sustentado el 10 de diciembre de 2018 frente al jurado:

Composición del jurado

- *Asesor*: Luis Felipe Giraldo Trujillo, Profesor Asistente, Universidad de Los Andes.
- *Jurados* : José Fernando Jiménez Vargas, Profesor Asociado, Universidad de Los Andes.

Contenido

1 INTRODUCCIÓN

La contaminación generada por los combustibles fósiles y el constante desarrollo tecnológico en fuentes de generación de energía renovable ha permitido que cada día estas tecnologías, principalmente eólica y solar fotovoltaica, tengan una mayor participación en la canasta energética mundial. Sin embargo, en el contexto colombiano, a pesar de que la llegada de fuentes renovables genere un impacto positivo en el medio ambiente y permita diversificar la canasta energética para evitar ser altamente dependientes del recurso hídrico y las volatilidades del precio del combustible, la intermitencia de estas fuentes supone un reto a superar para garantizar la correcta prestación del servicio eléctrico.

Al analizar la electricidad como servicio, se debe tener en cuenta la necesidad de igualar la oferta con la demanda en todo instante de tiempo. El desequilibrio entre oferta y demanda del servicio genera variaciones en la frecuencia del sistema, afectando negativamente la seguridad, confiabilidad y estabilidad de la red. Por este motivo, es de vital importancia pronosticar la disponibilidad del recurso renovable para establecer correctamente el despacho energético de acuerdo con las restricciones físicas que supone la intermitencia del recurso [1].

De acuerdo con la revisión bibliográfica realizada en [2], los métodos más utilizados para solucionar el problema descrito corresponden a ARIMAX (Autoregressive Integrated Moving Average with Exogenous Variable) y ANN (Artificial Neural Network). Estos trabajos analizados en [2], enfocados en la estimación de corto plazo, tienen, en su mayoría, un horizonte de predicción de una hora, y utilizan mediciones en sitio como descriptores principales. Sin embargo, la utilización de mediciones en sitio no es suficiente debido a que, como lo indica ese trabajo, la alta variabilidad de las energías renovables, principalmente solar, se debe a factores atmosféricos como los niveles de aerosoles y las alteraciones en la nubosidad, los cuales son medidos satelitalmente. Adicionalmente, de acuerdo con la investigación realizada, no se han hecho trabajos enfocados a la predicción de irradiación solar en Colombia, con horizonte de tiempo inferior a 2 días, utilizando estos métodos. Por este motivo, este proyecto se enfocará inicialmente en utilizar redes neuronales para predecir la irradiación solar sobre la superficie terrestre de Colombia definiendo horizontes temporales desde 1 hasta 48 horas y reconociendo aquellos descriptores que sean más importantes para elaborar la predicción.

Por otra parte, existen escenarios en los cuales no se puede disponer de históricos en sitio. En estos casos, para poder utilizar la predicción por los métodos descritos sería necesario realizar mediciones por extensos períodos de tiempo lo cual retrasaría la construcción y puesta en marcha del parque solar fotovoltaico. Debido a esto, el segundo enfoque principal de esta investigación es proponer una metodología de predicción en nuevas estaciones sin necesidad de tener mediciones con más de un día de anticipación en el punto a predecir. De esta manera no se necesitará ningún tipo de históricos, en el lugar de interés, para entrenar el modelo predictivo. Finalmente, para cumplir con estas tareas, se utilizarán como descriptores las mediciones satelitales provenientes de la base de datos MERRA-2 [3], con resolución temporal y espacial de una hora y 0.5° x 0.625° respectivamente. Además, se usarán mediciones en sitio, facilitadas por el IDEAM, para validar el modelo.

2 OBJETIVOS

2.1 Objetivo General

Predecir la irradiación solar en diversos puntos de Colombia, con un horizonte temporal entre 1 y 48 horas, a partir de mediciones en sitio y satelitales, utilizando Redes Neuronales.

2.2 Objetivos Específicos

- Determinar los descriptores relevantes para predecir irradiación solar terrestre.
- Establecer un modelo de red neuronal que permita predecir la irradiación solar con un rendimiento similar al encontrado en la literatura.
- Definir una métrica para predecir irradiación solar en estaciones que no se tienen en cuenta en el entrenamiento.

2.3 Alcance y productos finales

Se realiza una investigación de tipo exploratoria y descriptiva en la cual se adaptarán métodos ya aplicados y comprobados en diferentes ámbitos para realizar una primera aproximación a la predicción de irradiación solar en Colombia de muy corto plazo, es decir, con un horizonte de predicción inferior a 48 horas. Por este motivo, el producto final de este trabajo corresponde a pruebas que demuestren la efectividad de las redes neuronales para realizar predicciones en el horizonte mencionado y una metodología que permita mejorar la predicción en estaciones de las cuáles no se tengan mediciones previas para entrenar la red neuronal.

Este trabajo será el paso inicial para la posible creación de un aplicativo dinámico que muestre la predicción de irradiación solar con una antelación y resolución temporal y espacial superior, el cual requerirá de mayor cantidad de puntos de medición en sitio y el acceso, en tiempo real, a la información necesaria para la predicción.

3 DESCRIPCIÓN Y JUSTIFICACIÓN DE LA INVESTIGACIÓN

Teniendo en cuenta que la predicción de irradiación solar es un mecanismo para superar la intermitencia de la energía solar fotovoltaica y facilitar su inclusión en la canasta energética colombiana, este trabajo considerará los dos posibles escenarios a los que se puede enfrentar cualquier parque solar fotovoltaico que desee predecir a irradiación recibida para estimar su posible generación.

En el caso de parques solares con mediciones previas, como el ubicado en el municipio de El Paso, Cesar, construido por el grupo ENEL [9], se puede realizar una predicción basándose en históricos propios que tengan en cuenta las dinámicas meteorológicas características del lugar. Para este escenario, este trabajo mostrará la utilización de las redes neuronales como

mecanismo de predicción, especificando el rendimiento que se puede alcanzar para horizontes temporales de 1 a 48 horas y determinando aquellos descriptores que son más importantes al elaborar esta predicción. De esta manera, en caso de no disponer de todos los descriptores utilizados en esta investigación, quien desee entrenar su propio modelo predictor conocerá aquellas variables indispensables para hacerlo.

Adicionalmente, para aquellas entidades que deseen construir nuevos parques solares, se les presenta la posibilidad de elaborar una predicción de la irradiación solar a partir de históricos en lugares distintos, eliminando la necesidad de tomar mediciones durante más de 5 años para tener un conjunto de datos que permitan entrenar el modelo predictivo y facilitando cualquier estudio de factibilidad para este tipo de proyectos. Para cumplir con este objetivo, esta investigación definirá métricas que determinen cuáles son aquellos lugares específicos que se deben tener en cuenta para elaborar un modelo predictor, basado en redes neuronales. De esta manera, se podrán obtener resultados similares al caso previo sin necesidad de disponer de históricos en el sitio de predicción y mejorando el rendimiento de predicciones realizadas a partir de todos los datos disponibles.

4 NORMATIVIDAD

Para realizar este estudio, es necesario conocer las restricciones en términos normativos del manejo de información, específicamente de los datos proveídos por el IDEAM de acuerdo con la resolución 2367 de 2009 [12]. En primer lugar, esta resolución establece que toda la información estará disponible a la comunidad y será responsabilidad de los individuos particulares su utilización. En el caso específico de los datos utilizados en esta investigación, se aclaró que las mediciones no están completamente validadas, sin embargo, pueden ser usadas para propósitos enfocados en investigación sin intereses económicos de por medio y aceptando la completa responsabilidad por la posterior utilización de la investigación.

Adicionalmente, la resolución establece que los datos, a pesar de estar disponibles a terceros, requieren de 15 días hábiles para su entrega, sin garantizar un formato específico. Por este motivo, en trabajos futuros donde se desee realizar predicción en tiempo real es necesaria la vinculación directa con el IDEAM quienes serán los que deben facilitar los datos necesarios de manera periódica en la mayor brevedad posible. De esta manera, cualquier publicación que provenga de esta vinculación deberá ajustarse a las características específicas de esta entidad.

Con respecto a los datos satelitales obtenidos de MERRA-2, [12] establece que estos son libres y están a disposición de la comunidad científica con la citación respectiva, aunque, al igual que en el caso del IDEAM, estos presentan un retardo para el público general. Por este motivo, si se desea diseñar un aplicativo en tiempo real es necesaria la vinculación directa con el departamento encargado de la NASA y ajustarse a la reglamentación particular para el uso de esta información.

5 MARCO TEÓRICO

5.1 Clear Sky Model (CSM)

El CSM es un modelo que representa las dinámicas astronómicas del sol. Como se describe en [2], este modelo asume una situación ideal sin nubes para estimar la irradiación que llega a determinado punto en la tierra, teniendo en cuenta la latitud, longitud y elevación del lugar donde se desea predecir la irradiación. Con estos parámetros, el modelo se encarga de calcular la posición del sol con respecto al lugar establecido previamente y finalmente predice la irradiación global, directa y difusa en sitio. La implementación de este modelo se realiza mediante la librería descrita en [13].

5.2 Artificial Neural Networks (ANN)

Las ANN, como lo describe [4], intentan imitar el funcionamiento de un cerebro humano en la medida en que utilizan información experimental de determinados parámetros, denominados descriptores, y almacenan este conocimiento a partir de los pesos de cada conexión entre los elementos de la red, simulando el proceso sináptico entre neuronas del cerebro humano. La figura 1 muestra la analogía entre el modelo matemático de una neurona artificial y su contraparte biológica. En ella se observa un conjunto de entradas (Dendritas) provenientes de otras neuronas, o del exterior, asignándole un peso específico a cada una (Sinapsis), las cuales acceden a la neurona (Cuerpo celular) para computar una función de activación y determinar la salida (Axón) hacia la siguiente neurona o hacia el exterior. Esta similitud con el funcionamiento del cerebro humano le ha permitido convertirse en un método fundamental para el procesamiento de imágenes, reconocimiento de patrones y elaboración de pronósticos. Por ende, como lo manifiesta [2], se ha convertido en uno de los métodos predominantes para predecir irradiación solar y generación solar fotovoltaica de corto plazo.

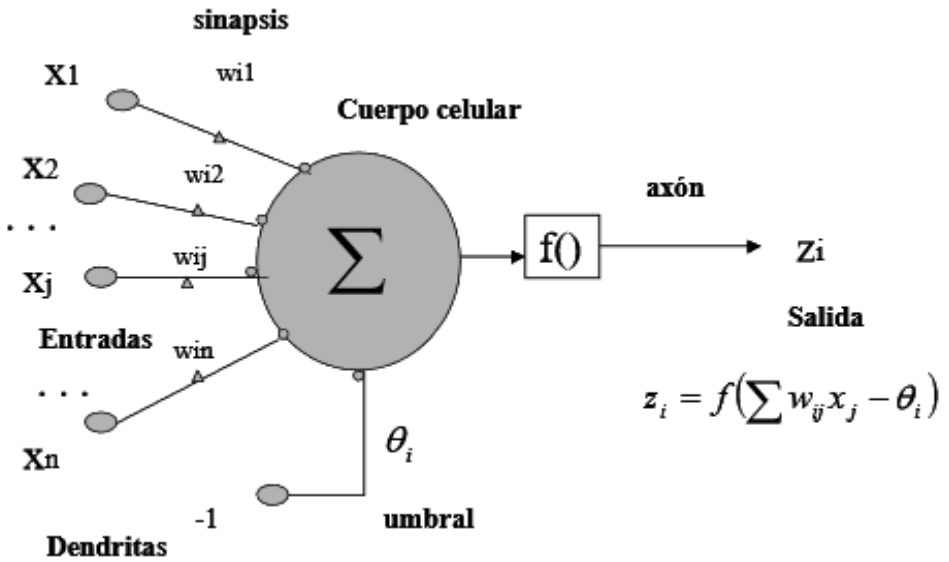

Figura 1. Neurona Artificial. Tomado de [4].

5.3 Indicadores de Rendimiento

Existen diversos indicadores de rendimiento que pueden ser utilizados para validar la efectividad en la predicción realizada por cualquier algoritmo de *Machine Learning*, como los descritos en [5] que se han utilizado específicamente en el problema de predicción de irradiación solar. Entre ellos se encuentra el *Mean Absolute Error* (MAE), que se encarga de obtener el promedio de los errores de cada predicción sobre el conjunto completo de validación. De esta manera, permite tener una mayor sensibilidad sobre el error promedio que tiene el algoritmo en término de las unidades que se desean predecir. El MAE se define de la siguiente manera:

$$MAE = \frac{1}{N}\sum \left| y - y_{pred} \right|$$

Un segundo indicador que permite medir el error de un conjunto de predicciones corresponde al *Mean Bias Error* (MBE) el cual calcula la diferencia entre las medias de los valores reales y la predicción realizada. De esta manera se puede conocer si el algoritmo tiende a cometer determinado error. El MBE se define a continuación:

$$MBE = \frac{1}{N}\sum y - y_{pred}$$

Otro indicador utilizado para analizar la efectividad de las predicciones es el *Root Mean Square Error* (RMSE) y su versión normalizada (NRMSE), los cuales, al tener en cuenta el cuadrado de las diferencias entre la predicción y la medida real, permiten darles una mayor relevancia a los errores provenientes de mediciones más altas. Estos indicadores se definen de la siguiente manera:

$$RMSE = \sqrt{\frac{1}{N}\sum \left(y - y_{pred}\right)^2}$$

$$NRMSE = \frac{\sqrt{\frac{1}{N}\sum \left(y - y_{pred}\right)^2}}{y_{mean}}$$

El último indicador que se utiliza en este trabajo corresponde al *Determination Coefficient* (R^2), el cual establece una relación entre el error de la predicción y la desviación de las medidas reales con respecto a la media, la cual puede utilizarse como medida de la calidad del modelo para replicar los resultados. Su descripción matemática es la siguiente:

$$R^2 = 1 - \sum \frac{\left(y - y_{pred}\right)^2}{(y - y_{mean})^2}$$

5.4 Weighted Mahalanobis Distance (WMD)

Usualmente, para describir la distancia entre dos puntos x_i y x_j, se utiliza la definición de norma euclidiana teniendo en cuenta el cuadrado de la diferencia de cada componente **c** que caracterice a los puntos en cuestión:

$$d(x_i, x_j) = \sqrt{\sum_{c=1}^{N} \left(x_i^c - x_j^c\right)^2} = \sqrt{\left(x_i - x_j\right)^T \left(x_i - x_j\right)}$$

Existen escenarios en los cuales cada componente puede tener una importancia distinta para conocer la diferencia entre dos puntos. Por este motivo, como se explica en [6], se puede introducir una matriz de pesos *W* que permita establecer la importancia de cada componente o, en el escenario de *Machine Learning*, descriptor y las posibles correlaciones entre ellos:

$$d_{Mahal}(x_i, x_j) = \sqrt{\left(x_i - x_j\right)^T W \left(x_i - x_j\right)}$$

La figura 2 muestra un ejemplo gráfico de una posible transformación que permite otorgarle más importancia a determinado descriptor. En caso de tener el mismo peso, las curvas de nivel del cálculo de distancia son círculo, pero, al asignarle pesos distintos, éstas pueden transformarse en elipses.

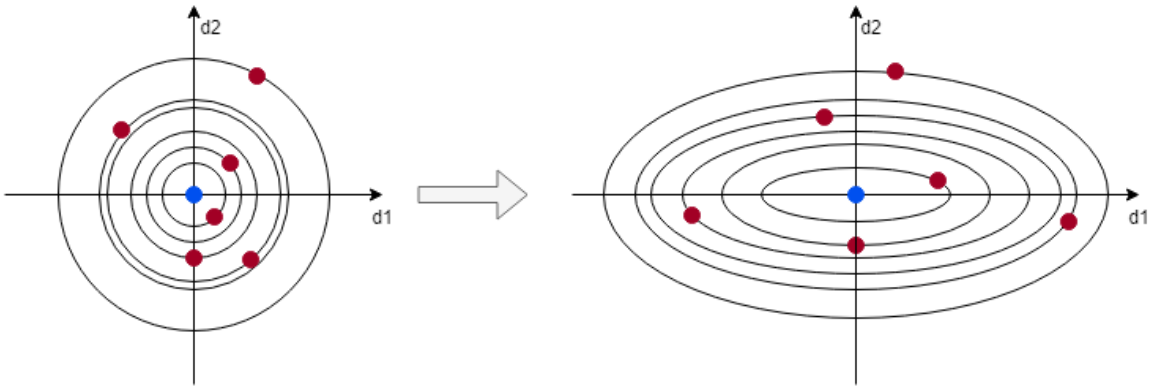

Figura 2. Transformación de WMD.

5.5 Sequential Backward-Forward Selection (SBS-SFS)

Los algoritmos SBS y SFS son utilizados principalmente para resolver problemas de predicción o clasificación que tienen elevada dimensionalidad. Como lo explica [7], la idea de estos algoritmos es adicionar (SFS) o eliminar (SBS) subconjuntos de datos hasta obtener el conjunto de entrenamiento que permita un mejor resultado.

El SFS parte de diversos subconjuntos de los datos totales y realiza un entrenamiento particular para cada uno de ellos hasta encontrar el mejor subconjunto para realizar el

entrenamiento. Acto seguido, combina el mejor subconjunto encontrando con los demás y vuelve a entrenar y validar con cada uno de ellos hasta encontrar la mejor pareja de subconjuntos. De esta manera, continúa adicionando subconjuntos al mejor subconjunto encontrado en el paso anterior del algoritmo hasta tener en cuenta todos los datos disponibles. De manera análoga, el SBS parte del conjunto total de datos y comienza a eliminar uno por uno los subconjuntos que, de acuerdo con la validación, generan un impacto más negativo en la predicción o clasificación, hasta obtener únicamente uno de los subconjuntos, el cual considera que es el que permite un mejor resultado.

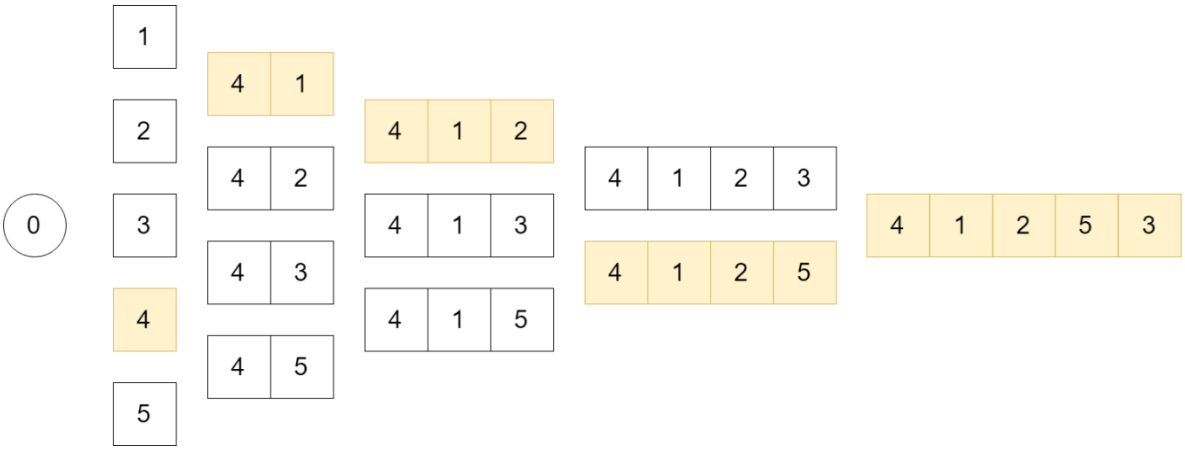

Figura 3. Ejemplo de SFS para obtener ranking de rentrenamiento

La utilización de estos algoritmos permite establecer un ranking de los subconjuntos que deben ser utilizados en el entrenamiento para obtener un mejor resultado que el conseguido a partir de un entrenamiento que tenga en cuenta la totalidad de los datos. La figura 3 muestra un ejemplo de la elaboración del ranking.

5.6 Learning to Rank

El problema de ranking intenta predecir un conjunto ordenado de elementos más cercanos, similares o deseados. A diferencia de otros algoritmos de *Machine Learning*, en este problema se tiene un vector solución en lugar de un número o clase, por lo cual es necesario utilizar una función de costo que tenga en cuenta el acierto o error en las posiciones de ese vector solución, en lugar de usar el cuadrado de la diferencia entre los valores predichos y los reales. Una de las funciones de costo más utilizadas corresponde a la *Normalized Discounted Cumulative Gain* (NDCG) la cual le otorga una relevancia a cada elemento del ranking, con respecto al ranking deseado, y acumula la suma de estas relevancias otorgando un mayor peso a los elementos que están en las primeras posiciones del ranking a partir de una función decreciente [10]:

$$NDCG(rank) = \sum_{i=1}^{k} \frac{rel_i}{\log(r+1)}$$

Las funciones de ese tipo poseen intervalos con derivadas iguales a cero y a infinito, por lo cual puede presentar problemas de convergencia a mínimos locales que no sean suficientemente buenos al resolver el problema de optimización. Por este motivo, en [10] se propone una función similar a la función de entropía usada en problemas de clasificación, que utilice como referencia las ubicaciones relativas entre los elementos del ranking predicho con respecto a sus posiciones ideales establecidas por el ranking original. De esta manera, el costo que aportan la ubicación relativa entre los elementos *i* y *j* donde el elemento *i* ocupa una mejor posición en el ranking deseado que j, está dado por:

$$C_{i,j}^{rank} = s_i - s_j + \ln(1 + e^{s_j - s_i})$$

La sumatoria de estas funciones de costo dada por cada pareja de elementos del ranking genera una tendencia a enviar hacia adelante en el ranking predicho a aquellos elementos que están mejor ubicados en el ranking original.

5.7 Principal Components Analysis (PCA)

Como lo explica [8], la técnica PCA corresponde a una proyección lineal de un conjunto de datos a un nuevo espacio en el cual las variables no estén correlacionadas. La componente principal corresponde a aquella dirección en que se puede capturar la mayor varianza posible de los datos. La segunda componente principal es la segunda dirección con mayor varianza y así sucesivamente hasta tener nuevamente un número de dimensiones igual al conjunto original de los datos, pero, a diferencia del conjunto original, maximizando la varianza entre ellos. Adicionalmente, PCA permite conocer el porcentaje de varianza total que representa cada componente principal por medio de los valores propios de la matriz de transformación. De esta manera se puede aplicar una reducción de dimensiones teniendo en cuenta un umbral de varianza deseado.

5.8 Kernelización de Métodos Lineales

De acuerdo con [6], el conjunto de métodos lineales es limitado para resolver determinados problemas de *Machine Learning*. Por este motivo, es necesario utilizar *Kernels* que permitan ampliar la familia de modelos para encontrar una mejor solución. Un *Kernel*, descrito como:

$$k(x_i, x_j) = \phi(x_i)^T \phi(x_j) = x_i^{H^T} x_j^H$$

Corresponde a una función *k* aplicada a dos vectores, de tal manera que el producto punto de los vectores resultantes al aplicar la transformación no lineal $\phi(x)$ a cada uno de ellos sea igual al producto punto entre los vectores originales operándolos en un espacio de Hilbert desconocidos. El principal beneficio de esta formulación consiste en poder reemplazar productos puntos con características lineales por *Kernels* que realicen el mismo producto punto en un espacio distinto y permitan incorporar comportamientos no lineales en el espacio

original. Entre los *Kernels* más utilizados se encuentra el *Kernel* de Base Radial (RBF), Polinomial (POL) y Sigmoidal (SIG) descritos a continuación:

$$k_{rbf}(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right)$$

$$k_{pol}(x_i, x_j) = \left(\gamma\left(x_i^T x_j\right) + r\right)^d$$

$$k_{sig}(x_i, x_j) = \tanh\left(\gamma\left(x_i^T x_j\right) + r\right)$$

Adicionalmente, al poder representar los *Kernels* previos como productos puntos, se pueden diseñar nuevos *Kernels* utilizando composiciones y combinaciones lineales entre los descritos previamente.

6 DATOS DE ENTRENAMIENTO Y VALIDACIÓN

6.1 Mediciones en Sitio

Se solicitaron mediciones en sitio de irradiación solar, temperatura y humedad relativa con temporalidad horaria de 10 estaciones del IDEAM buscando la mayor variabilidad entre estaciones. De esta manera se escogieron estaciones de diferentes puntos cardinales y elevación dentro del territorio colombiano. Estas estaciones se muestran en la tabla 1.

A partir de ahora, todas las estaciones serán nombradas dependiendo del departamento en que se encuentran para facilitar la comprensión de las posibles relaciones geográficas existentes entre ellas. Para conocer su ubicación exacta el lector puede remitirse nuevamente a la tabla 1.

Estación	Municipio	Departamento	Latitud (°)	Longitud (°)	Elevación (M.S.N.M)
Botana	Pasto	Nariño	1,16	-77,28	2820
Capurganá	Acandí	Chocó	8,51	-77,32	20
Carmen De Bolivar	Carmen de Bolívar	Bolívar	9,63	-75,1	190
El Diamante	Paz de Ariporo	Casanare	5,82	-71,42	160
Granja Paici	Uribia	La Guajira	11,59	-72,32	15
Ica Villavicencio	Villavicencio	Meta	4,14	-73,63	444
Las Brisas	Ansermanuevo	Valle del Cauca	4,78	-76,14	1982
Maceo	Maceo	Antioquía	6,57	-74,79	1112
San Vicente Del Caguán	San Vicente del Caguán	Caquetá	2,16	-74,75	275
Univ. F/Co De Paula Santander	Cúcuta	Norte de Santander	7,9	-72,49	311

Tabla 1. Estaciones Meteorológicas IDEAM.

6.2 Mediciones Satelitales

Se utilizó la base de datos MERRA-2 [3] para obtener las mediciones satelitales, con temporalidad horaria, de presión, columna de ozono, precipitación de hielo, agua y vapor, y viento en sentido este-oeste y norte-sur. La resolución de estas mediciones es de 0.5° para latitud y 0.625° para longitud, por lo cual es imposible tener la medición exacta en cada una de las estaciones que se utilizan para validar. Por este motivo, se estableció la cuadrícula delimitada por estas mediciones satelitales alrededor del territorio colombiano y se determinaron los cuatro extremos del cuadrado que rodeaban a cada una de las estaciones para promediar los valores de cada variable y obtener una aproximación de las mismas en los 10 puntos coordenados del estudio.

6.3 Organización de los datos

Para cada una de las estaciones, se creó una matriz con todas las mediciones disponibles para cada una. Los datos a usar como descriptores corresponden a latitud, longitud, elevación de la estación, hora y fecha en la que se desea hacer la predicción, las mediciones disponibles *h* horas atrás, donde *h* corresponde al horizonte de predicción, para tener en cuenta las últimas mediciones disponibles.

Además, se agregaron las mediciones 24, 48, 72, 96 o 120 horas atrás para tener en cuenta el comportamiento estacional de la irradiación solar, como se propone en [14], el cual tiende a repetirse cada 24 horas. A pesar de que en la mayoría de los casos se usó como referencia del fenómeno estacional la medición del día previo, en algunos casos en los que esa posición de la serie de tiempo estuviera vacía se realizaba un algoritmo iterativo hacia atrás hasta encontrar la medición previa más cercana durante la hora de predicción. En los casos específicos con *h* de 24 y 48 horas se utilizaron las mediciones de 2 o 3 días atrás al momento que se desea predecir debido a que no tiene sentido utilizar las mismas medidas para describir el último comportamiento conocido y el factor estacional, ni usar mediciones que se hayan tomado en un horizonte inferior al de la predicción.

Finalmente, se añade un descriptor adicional que corresponda al CSM para el momento que se desea predecir. También se crea un vector que contenga el valor de la irradiación real de tal forma que sea el conjunto solución que se utilizará para el entrenamiento o validación dependiendo de la prueba a realizar. Los descriptores y la irradiación de cada estación se almacenan en una estructura de dimensión 10 para utilizarla en cada una de las diferentes pruebas de predicción.

7 PREDICCIÓN EN ESTACIONES CONOCIDAS

7.1 Entrenamiento Red Neuronal

Tras observar que todas las estaciones tenían entre 20000 y 80000 datos, no se decidió usar ningún algoritmo para balancear el número de muestras por estación. Para entrenar la red neuronal, se tuvieron en cuenta el 90% de los datos disponibles por estación, reservando los

datos restantes para validar. Adicionalmente se aplica estandarización utilizando media y desviación estándar para que todos los descriptores se encuentren en el mismo orden de magnitud. Tras realizar el preprocesamiento, se realizaron diversas pruebas, para el caso de 1 hora de adelanto, con diferentes topologías, funciones de activación y algoritmos de solución para los pesos de la red neuronal, teniendo en cuenta una cantidad máxima de capas intermedias igual a 3, para evitar costos computacionales excesivos en el entrenamiento. Con el mejor modelo encontrado se entrenó y validó una red neuronal para escenarios de 1, 2, 3, 4, 5, 6, 12, 24 y 48 horas de anticipación utilizando los indicadores de rendimiento descritos previamente.

7.2 Resultados de Predicción

De las pruebas realizadas, la red con mejores indicadores de rendimiento tiene 3 capas intermedias con 20, 20 y 5 neuronas respectivamente. Esta red neuronal utiliza función de activación ReLU (Unidad Lineal Rectificada) y usa como algoritmo de solución del problema de optimización el gradiente descendiente estocástico optimizado ADAM propuesto en [11]. Con esta topología, se realizó la predicción con un horizonte desde 1 hasta 48 horas obteniendo los resultados mostrados en la tabla 2. En ella se puede observar que al aumentar las horas de antelación en la que se realiza la predicción, el error se incrementa mientras la calidad del modelo disminuye. Esto se debe a que existe una mayor incertidumbre al haber más tiempo para que el clima varíe con respecto a lo esperado.

HORAS	MAE (W/m^2)	MBE (W/ m^2)	RMSE (W/ m^2)	NRMSE (%)	R^2 (%)
1	44,63	1,18	89,16	47,94	89,21
2	55,94	-1,07	107,35	57,72	84,36
3	58,9	0,94	112,79	60,64	82,74
4	61,35	2,84	115,84	62,28	81,79
5	61,5	-0,24	117,22	63,02	81,36
6	61,58	1,52	117,26	63,04	81,34
12	63,41	4,42	119,07	64,02	80,76
24	63,13	-2,33	119,39	64,19	81,66
48	63,83	4,72	120,88	65,02	80,17

Tabla 2. Entrenamiento con 28 descriptores.

Al comparar estos resultados con los compilados en [2] se puede observar que el RMSE es muy similar para un horizonte de predicción inferior a 4 horas manteniéndose en una escala de 88-117 W/m^2, mientras que para escenarios superiores se obtiene un RMSE bastante inferior al de estos estudios previos. Es importante reconocer que estos resultados no son directamente comparables debido a que los indicadores de rendimiento dependen altamente de las características geográficas y atmosféricas del lugar donde se realiza la predicción, por lo cual los resultados obtenidos únicamente se pueden contrastar con otros estudios que también se hayan realizado en Colombia, sobre los cuales no se encontró información. A pesar de esto, la comparación con estudios previos en otros lugares del planeta permite observar que el orden de magnitud de los indicadores es similar y por ende los resultados

pueden considerarse suficientemente buenos para este primer acercamiento que permitirá comparaciones con estudios futuros que se realicen en el país.

Adicionalmente, se entrenó utilizando el total de descriptores satelitales antes de promediar para obtener una medición por variable para cada estación en sitio, es decir, teniendo en cuenta los vértices del cuadrado que rodea a cada estación al establecer la cuadrícula de mediciones satelitales. La tabla 3 muestra que adicionar estos descriptores no tiene un efecto positivo en la predicción, a pesar de incrementar el costo computacional del entrenamiento, por lo cual no debería realizarse este aumento de descriptores para el entrenamiento.

HORAS	MAE (W/ m^2)	MBE (W/ m^2)	RMSE (W/ m^2)	NRMSE (%)	R^2 (%)
1	45,33	-0,14	89,19	47,95	89,2
2	54,93	3,67	105,94	56,96	84,77
3	59,44	2,17	112,69	60,59	82,77
4	61,01	3,37	116,39	62,58	81,62
5	62,09	0,62	117,72	63,29	81,2
6	62,28	2,76	117,39	63,11	81,3
12	63,09	1,37	118,55	63,74	80,93
24	62,45	6,26	119,5	64,25	80,62
48	64,05	6,04	122,1	65,68	79,76

Tabla 3. Entrenamiento con 70 descriptores.

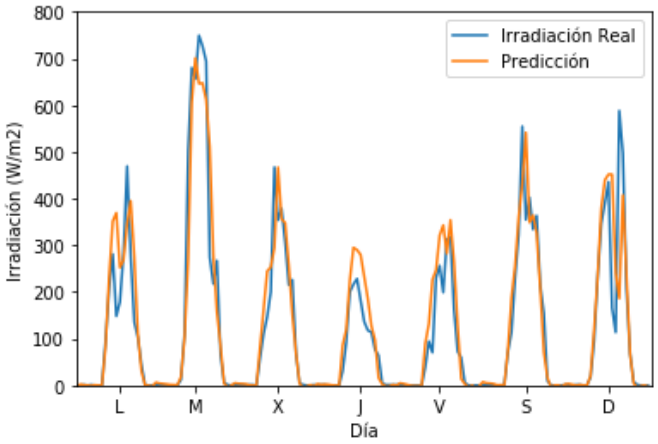

Figura 4. Predicción con horizonte temporal de 1h.

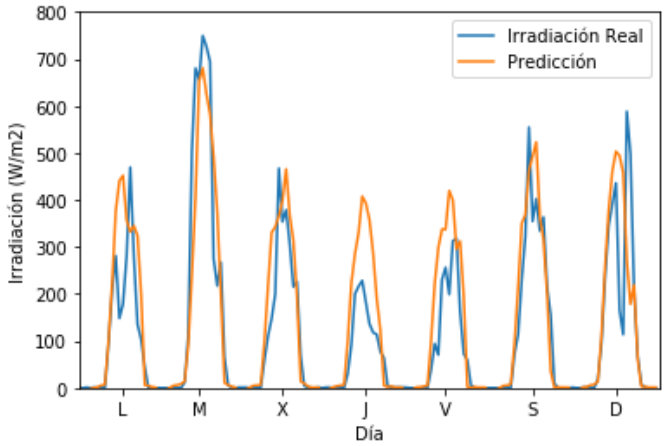

Figura 5. Predicción con horizonte temporal de 2h.

Las figuras 4 a 7 muestran la comparación entre la irradiación real y la irradiación predicha durante una semana con escenarios de predicción de 1, 2, 6 y 48 horas respectivamente. Para el escenario de 1 hora, se observa que el algoritmo puede predecir las dinámicas atmosféricas y las variaciones de la nubosidad, al igual que en el escenario de 2 horas, con una disminución en el rendimiento de la precisión para el segundo caso. Sin embargo, para los horizontes de 6 horas o superiores, el algoritmo sólo es capaz de predecir la distribución general de la irradiación sin considerar variaciones abruptas causadas por las nubes. Esto se debe principalmente a que, para un horizonte superior a 2 horas, no se conoce la

información necesaria para determinar la nubosidad en el momento que se desea conocer la irradiación solar.

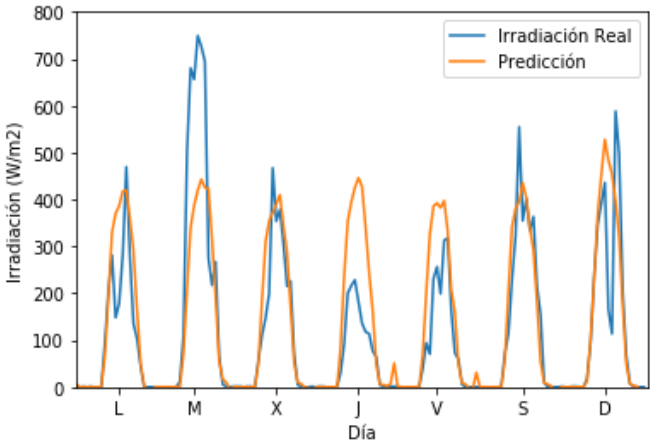

Figura 6. Predicción con horizonte temporal de 6h.

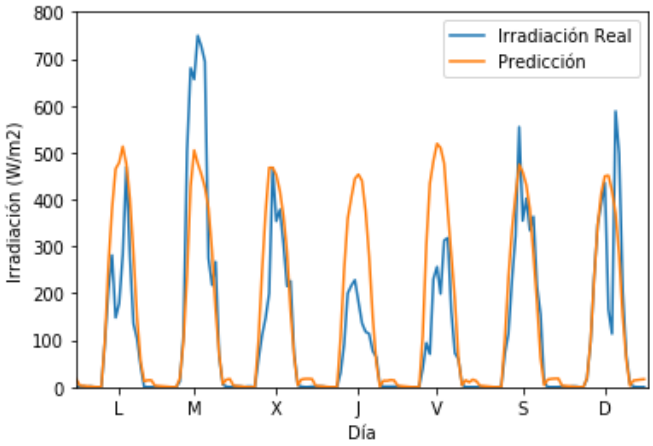

Figura 7. Predicción con horizonte temporal de 48h.

7.3 Feature Importance (FI)

Es posible que algunos de los descriptores sean muy poco relevantes para predecir la irradiación solar y por ende su adición en el entrenamiento contamine la predicción. Además, es importante reconocer las características fundamentales para predecir la irradiación, de tal manera que en futuros estudios se conozcan aquellas mediciones vitales para este proceso. Por este motivo se calcula el porcentaje de importancia de cada descriptor.

La medición del FI se realiza con un promedio ponderado de los pesos de cada descriptor con respecto a la primera capa de la red neuronal. No se tuvieron en cuenta capas posteriores debido a que, tras la primera capa, todas las neuronas están interconectadas con las de la siguiente, por lo cual se asume que la importancia de cada descriptor es constante después de haber pasado por la primera capa de la red.

$$FI_i = \frac{\sum_{j=1}^{N_{neuronas}} |w_{ij}|}{\sum_{i=1}^{N_{descriptores}} \sum_{j=1}^{N_{neuronas}} |w_{ij}|}$$

La tabla 4 resume el FI de los descriptores utilizados en el entrenamiento, obtenido como el promedio del FI calculado para todos los horizontes de medición, donde destaca la importancia vital de la hora del día, latitud, elevación, temperatura y el modelo CS, justificando la inclusión de este último en el entrenamiento. Adicionalmente, se observa la relevancia de las características estacionales para la predicción, especialmente en las variables de presión, irradiación y temperatura. También se puede apreciar que las mediciones en sitio tienen un FI mayor a las satelitales, aunque estas últimas, especialmente la presión atmosférica, velocidad de viento y precipitación de vapor, son necesarias para una

mejor predicción, debido a que están íntimamente relacionadas con la formación y movimiento de nubes.

Descriptores	FI (%)	Descriptores	FI (%)
Hora	13,52	Viento Norte (t-k)	2,46
Latitud	7,5	Prec. Vapor (t-k)	2,44
Clear Sky (t)	6,36	Viento Este (t-24n)	2,18
Temperatura (t-k)	6,07	Humedad (t-24n)	2,17
Elevación	6	Ozono (t-24n)	1,92
Irradiación (t-k)	5,9	Ozono (t-k)	1,88
Presión (t-k)	5,36	Prec. Vapor (t-24n)	1,88
Presión (t-24n)	4,88	Año	1,86
Longitud	4,31	Prec. Agua (t-k)	1,49
Irradiación (t-24n)	3,77	Viento Norte (t-24n)	1,47
Viento Este (t-k)	3,38	Prec. Agua (t-24n)	1,32
Temperatura (t-24n)	3,28	Prec. Hielo (t-k)	1,16
Humedad (t-k)	2,94	Prec. Hielo (t-24n)	1,1
Mes	2,65	Día	0,72

Tabla 4. Feature Importance

Por último, se realizaron predicciones utilizando los 10 descriptores más importantes, es decir, aquellos que tuvieran un FI superior a 3,5%, obteniendo los resultados mostrados en la tabla 5. En ella se puede observar que existe un ligero decremento en el rendimiento de las predicciones, aunque éstas siguen mostrando resultados similares a los obtenidos a partir del total de descriptores. Esto demuestra que determinando aquellos descriptores más importantes y utilizando únicamente estos para entrenar el algoritmo predictor se puede obtener un resultado suficientemente bueno sin necesidad de disponer de un número elevado de variables climatológicas y atmosféricas.

HORAS	MAE (W/ m^2)	MBE (W/ m^2)	RMSE (W/ m^2)	NRMSE (%)	R^2 (%)
1	45,75	2,6	91,25	49,06	88,7
2	59,06	-3,56	109,94	59,11	83,6
3	61,49	-3,26	115	61,83	82,05
4	64,53	-1,93	120,27	64,66	80,37
5	65,35	-1,42	121,17	65,14	80,07
6	65,27	-2,94	120,95	65,03	80,15
12	64,32	1,38	120,55	64,81	80,28
24	64,83	2,64	119,44	64,21	80,64
48	64,61	0,32	120,29	64,7	80,36

Tabla 5. Entrenamiento con 10 descriptores.

8 PREDICCIÓN EN ESTACIONES DESCONOCIDAS

Las predicciones realizadas hasta este punto requieren conocer información histórica en el sitio de predicción para entrenar el modelo predictivo. Sin embargo, requeriría demasiado tiempo tener esta información, por lo cual es importante tener una metodología que permita predecir la irradiación con un rendimiento similar sin necesidad de tener estos históricos, hasta que se disponga de ellos para utilizar la metodología previa.

8.1 Entrenamiento con Todos los Datos Disponibles

La primera aproximación a este problema, sugiere un entrenamiento a partir de todos los demás datos disponibles. Para esto, se aplicó un algoritmo de validación cruzada sobre todas las estaciones de tal forma que se entrenara utilizando 9 de las 10 estaciones disponibles y se validara utilizando la décima. Se escogió un horizonte de 1 hora debido a que, de acuerdo con las pruebas mostradas en la sección anterior, este escenario permite una mejor descripción de las dinámicas de las nubes y, por ende, facilita la observación y análisis del rendimiento de la predicción.

Los resultados obtenidos se resumen en la tabla 6. Como se esperaba, para la mayoría de estaciones se tiene una predicción más alejada de la realidad que la obtenida al utilizar los históricos en el lugar en que se deseaba predecir. Teniendo en cuenta únicamente el R^2, sólo dos estaciones superan el 89,21% promedio obtenido en el escenario de 1h usando históricos. También se puede observar que las estaciones cuyas predicciones tienen un rendimiento inferior corresponden a las ubicadas en Nariño y Meta, los cuales son puntos extremos en el mapa formado por las estaciones seleccionadas y tienen muy pocas estaciones a su alrededor para permitir elaborar una mejor predicción. Adicionalmente, la tabla 6, específicamente el elevado MBE, permite concluir que el principal error de las predicciones proviene de una elevada desviación de medias originada por utilizar ubicaciones diferentes en el entrenamiento.

Estación	MAE (W/ m^2)	MBE (W/ m^2)	RMSE (W/ m^2)	NRMSE (%)	R^2 (%)
Nariño	62,07	30,52	109,65	69,94	77,77
Chocó	48,71	16,13	93,31	52,86	88,03
Bolivar	43,45	9,14	83,89	40,93	90,77
Casanare	60,83	-8,61	97,39	47,52	87,91
La Guajira	47,28	27,54	81,54	34,99	93,34
Meta	41,01	17,06	86,34	84,89	78,37
Valle del Cauca	60,61	-29,8	114,22	60,72	82,66
Antioquía	55,95	13,45	109,75	54,3	85,31
Caquetá	93,82	-30,11	111,15	67,65	81,39
Norte de Santander	61,1	-15,86	114,44	56,49	85,76

Tabla 6. Validación Cruzada utilizando todos los datos disponibles.

Teniendo en cuenta que algunas estaciones están particularmente alejadas de otras, es posible que al utilizar todos los datos disponibles para realizar los entrenamientos se obtengan resultados inferiores que seleccionando específicamente un conjunto de estaciones. Por ejemplo, las características geográficas y atmosféricas de Nariño son muy distintas a las de Meta, por lo cual se podría obtener un mejor resultado al no utilizar esta estación para establecer el algoritmo predictor. Bajo esta intuición se proponen las metodologías descritas en las siguientes secciones.

8.2 Sequence Forward Selection

Como se mencionó previamente, la idea general del SFS es establecer un ranking del orden específico de estaciones que deben ser utilizadas en el entrenamiento para obtener el mejor resultado posible. A pesar de que el algoritmo SFS es una aproximación al ranking ideal, debido a que no prueba todas las posibles combinaciones entre estaciones, es una aproximación que permite comprobar la hipótesis planteada sin recurrir a los excesivos costos computacionales que requeriría entrenar con todas las combinaciones posibles hasta obtener el mejor resultado.

Para cada estación se determinó el orden en que deben adicionarse los subconjuntos de entrenamiento, correspondientes a los datos de otras estaciones, para obtener el mejor resultado posible en cada iteración del algoritmo. Como indicador a minimizar se utilizó el RMSE, por lo cual en cada iteración se añadía al conjunto total de entrenamiento el subconjunto cuya adición generó un menor RMSE en la predicción.

Desde este punto, se mostrarán los resultados obtenidos específicamente para la estación de Botana en Nariño, debido a que, de acuerdo con los indicadores de la tabla 6, es la estación donde se observa un menor rendimiento de la predicción. La tabla 7 muestra el orden de adición de estaciones de acuerdo con el algoritmo SFS para predecir en la estación mencionada.

Estación	MAE (W/ m^2)	MBE (W/ m^2)	RMSE (W/ m^2)	NRMSE (%)	R^2 (%)
Caquetá	77,59	-7,24	117,17	74,75	74,61
Bolívar	67,78	-42,60	114,59	73,10	75,72
Chocó	80,95	-66,90	141,52	90,27	62,96
Casanare	61,75	-7,63	100,30	63,98	81,39
Antioquía	68,25	29,27	107,25	68,42	78,73
Valle del Cauca	62,27	30,83	108,23	69,04	78,33
Norte de Santander	57,27	20,31	103,48	66,01	80,2
La Guajira	58,86	28,43	108,07	68,94	78,4
Meta	60,33	28,47	108,23	69,04	78,34

Tabla 7. Algoritmo SFS para estación de Nariño.

Se observa una mejoría del rendimiento del modelo predictor al utilizar únicamente 7 estaciones en lugar de entrenar con las 9 disponibles. De hecho, las dos estaciones que

deberían ignorarse para este entrenamiento están ubicadas en La Guajira y Meta, los cuales son dos de los departamentos más alejados geográficamente de Nariño. Sin embargo, la inclusión del departamento de Bolívar en el segundo lugar del ranking permite pensar que además de la distancia geográfica existen otras características importantes para determinar las estaciones con las cuales debe entrenarse el modelo. Por este motivo, se propone utilizar un algoritmo de optimización que permita definir una métrica para establecer las estaciones ideales para el entrenamiento.

8.3 Ranking usando WMD y PCA

Todas las estaciones poseen características propias o que pueden conocerse sin necesidad de tomar mediciones de irradiación en sitio, sino a partir de mediciones satelitales. Estas son latitud, longitud, elevación, temperatura, humedad, presión, columna de ozono, precipitación de hielo, agua y vapor, y viento en sentido Este-Oeste y Norte-Sur. Adicionalmente, también se puede conocer una aproximación de la irradiación característica en cada lugar por medio del modelo CS sin necesidad de haber tomado mediciones en sitio. Estas 13 variables permiten crear un vector que caracterice a cada una de las 10 estaciones del estudio.

En este punto se utiliza el concepto de WMD como medida de distancia entre los vectores mencionados, es decir, entre las estaciones meteorológicas disponibles. Al definir una matriz de pesos *W* diagonal y positiva definida, donde los pesos son las variables a decidir en el algoritmo de optimización, se puede encontrar una métrica que permita encontrar las estaciones con las que se debe entrenar, a partir de organizar de menor a mayor las distancias con respecto a la estación de referencia. De esta manera, se evita la necesidad de realizar el algoritmo SFS disminuyendo los elevados costos computacionales que se tendrían al superar las 10 estaciones.

La función de costo *J(W)* de este problema de optimización se define a partir de una modificación de la propuesta por [10] de tal manera que se intente minimizar las diferencias existentes con el ranking ideal obtenido mediante SFS. Además, esta función intenta hacer lo mismo con la WMD con respecto a la estación *k* en la que se desea predecir para distanciar lo máximo posible a aquellas estaciones *j* que de acuerdo con el ranking deben ser más lejanas que las estaciones *i*:

$$\min J(W) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} s_i - s_j + \ln(1 + e^{s_j - s_i}) + \lambda \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ik}^{mahal} - d_{jk}^{mahal} + \ln\left(1 + e^{d_{jk}^{mahal} - d_{ik}^{mahal}}\right)$$

Al resolver este problema de optimización, se observó que dependiendo de la inicialización el algoritmo convergía a diferentes mínimos locales. Esto se debe a que existen correlaciones entre las variables características de cada estación que no se están teniendo en cuenta en la solución del algoritmo de optimización. Por ejemplo, el modelo CS depende especialmente de latitud, longitud y elevación, por lo cual, al asignarle una alta importancia a este modelo,

probablemente disminuya el peso de las otras tres variables solamente porque su relevancia ya esté incluida en el modelo CS.

Para tener en cuenta estas correlaciones, se aplica PCA antes de resolver el problema de optimización, de manera que se tenga en cuenta las correlaciones existentes entre los descriptores de las estaciones. Además, se utilizan los valores propios de la matriz de transformación de PCA para inicializar la matriz de pesos *W* que se utiliza en el problema de optimización. De esta manera, se utiliza la varianza descrita por cada una de las componentes principales para asignarle más peso a las componentes más relevantes.

Estación	MAE (W/ m^2)	MBE (W/ m^2)	RMSE (W/ m^2)	NRMSE (%)	R^2 (%)
Valle del Cauca	124,14	115,01	164,92	105,20	49,70
Meta	151,42	146,54	192,88	123,04	31,20
Caquetá	84,56	64,52	138,54	88,38	64,50
Norte de Santander	57,84	24,08	103,03	65,72	80,37
Antioquía	59,83	20,40	102,87	65,62	80,43
Casanare	52,44	2,57	96,59	61,61	82,75
Bolívar	64,92	27,31	105,43	67,25	79,45
Chocó	69,91	43,64	116,02	74,01	75,11
La Guajira	65,38	26,33	109,68	69,96	77,76

Tabla 8. Ranking PCA-WMD para estación de Nariño.

La tabla 8 permite apreciar que, a pesar de obtener un ranking diferente al establecido por SFS, nuevamente se obtiene un mejor resultado al no utilizar determinadas estaciones para el entrenamiento. En este caso, el ranking vuelve a sugerir que La Guajira no debe utilizarse para este entrenamiento en específico. Sin embargo, las estaciones de Bolívar y Chocó que deberían ser incluidas según el ranking elaborado por el algoritmo SFS, no se deberían tener en cuenta de acuerdo con la métrica dada por el algoritmo de optimización. La diferencia entre ambos rankings radica en que al intentar aprender la matriz de pesos *W* el algoritmo intenta satisfacer hasta determinado punto todos los rankings, pero es incapaz de aprenderlos completamente porque un mejor indicador para el ranking de una estación específica podría damnificar el de otra. Por este motivo se propone una versión Kernelizada de este algoritmo para tener en cuenta comportamientos no lineales y tener más espacios disponibles para proyectar el vector representativo de cada estación en el espacio específico en que la distancia permita establecer el ranking ideal de entrenamiento.

8.4 Ranking utilizando Kernel

A partir de la definición de distancia WMD, esta puede ser descrita en términos de productos vectoriales sin realizar directamente la resta de la siguiente manera:

$$d_{mahal}(x_i, x_j) = \sqrt{x_i^T W x_i + 2x_i^T W x_j + x_j^T W x_j}$$

Conociendo la utilización de *Kernels* para realizar productos puntos entre vectores en un espacio de Hilbert diferente al original, además de saber que la matriz de pesos W es positiva definida y diagonal, la definición de distancia puede ser reescrita como:

$$d_{mahal}(x_i, x_j) = \sqrt{k\left(W^{1/2}x_i, W^{1/2}x_i\right) + 2k\left(W^{1/2}x_i, W^{1/2}x_j\right) + k\left(W^{1/2}x_j, W^{1/2}x_j\right)}$$

Definiendo la multiplicación de $W^{1/2}x_i$ como un nuevo vector x_i' la WMD kernelizada es:

$$d_{mahal}(x_i, x_j) = \sqrt{k(x_i', x_i') + 2k(x_i', x_j') + k(x_j', x_j')}$$

De esta manera, se puede utilizar el problema de optimización definido previamente sobre un conjunto de soluciones que están en un espacio no lineal modelado por la función *Kernel* utilizada para resolver los productos vectoriales. Tras probar los *Kernels* descritos en la sección 4.8, y diferentes combinaciones lineales y composiciones entre ellos, se obtuvo que el mejor resultado en términos de cercanía al ranking establecido por el algoritmo SFS, para todas las estaciones disponibles, corresponde a la composición de dos *Kernels* sigmoidales de la siguiente manera:

$$k_{sig}(x_i, x_j) = \tanh\left(\gamma \tanh\left(x_i^T x_j\right) + r\right)$$

La tabla 9 muestra nuevamente que la mejor predicción para la estación de Nariño se obtiene al no utilizar los datos de La Guajira para el entrenamiento, aunque, a diferencia de las pruebas anteriores, tiene en cuenta todas las otras estaciones. A pesar de ser el ranking más similar al establecido por el algoritmo SFS, los resultados de la predicción son ligeramente inferiores a las metodologías previas, lo cual puede ocurrir por la inicialización aleatoria de la red neuronal. Sin embargo, este rendimiento también se debe a que el ranking definido por el SFS no es el óptimo sino una aproximación del mismo que se genera a partir de las relaciones entre los subconjuntos de entrenamientos que se van concatenado. Por eso, una ligera variación en el ranking, generará una agrupación diferente de datos que puede desencadenar resultados inferiores a otros rankings que difiera más del deseado, pero agrupe subconjuntos de mejor manera.

Estación	MAE (W/ m^2)	MBE (W/ m^2)	RMSE (W/ m^2)	NRMSE (%)	R^2 (%)
Valle del Cauca	127,60	118,04	164,85	105,16	49,74
Caquetá	158,86	151,94	188,74	120,40	34,12
Antioquía	62,32	34,02	112,31	71,65	76,67
Meta	65,04	43,42	121,82	77,71	72,56
Chocó	68,24	44,44	115,30	73,55	75,42
Norte de Santander	69,44	41,98	108,65	69,31	78,17
Casanare	60,45	27,10	104,07	66,39	79,97
Bolívar	59,99	28,23	103,47	66,00	80,20
La Guajira	64,53	37,62	110,13	70,25	77,57

Tabla 9. Ranking Kernel-PCA-WMD para estación de Nariño.

La única forma de garantizar que una mayor similitud con el ranking ideal genere mejores resultados y se establezca de esta manera la métrica para decidir cuáles estaciones deben utilizarse para el entrenamiento es encontrando el mejor ranking posible. Además, se requiere tener una familia de espacios suficientemente grande como para que sea posible encontrar el espacio donde el ranking deseado y el ranking predicho sean iguales o al menos la concatenación de subconjuntos hasta el mejor resultado posible sea la misma. Como los costos computacionales y las restricciones de tiempo hacen imposible llevar este estudio hasta ese punto, se propone una metodología adicional que no requiere realizar entrenamientos sucesivos para establecer ese ranking.

8.5 Ranking a partir de FI y PCA

Como alternativa al problema de optimización planteado en la sección 7.2, se propone la utilización del FI calculado en el entrenamiento basado en históricos para establecer la métrica que indique las estaciones que deben ser utilizadas para entrenar el modelo predictivo. La tabla 10 muestra el resultado de utilizar únicamente la distancia euclidiana para determinar las estaciones más cercanas y establecer el ranking de estaciones que deben ser tenidas en cuenta para entrenar. Como denominador común vuelve a aparecer La Guajira como una de las estaciones que no debe usarse en el entrenamiento y los indicadores de rendimiento vuelven a ser similares a los obtenidos con otras métricas.

Estación	MAE (W/ m^2)	MBE (W/ m^2)	RMSE (W/ m^2)	NRMSE (%)	R^2 (%)
Valle del Cauca	131,50	123,03	171,75	109,56	45,45
Meta	117,90	110,60	162,93	103,93	50,91
Caquetá	110,42	102,06	165,25	105,42	49,50
Antioquía	66,28	42,77	110,67	70,59	77,35
Norte de Santander	74,09	-16,05	108,72	69,36	78,14
Bolívar	73,84	51,74	117,26	74,80	74,57
Casanare	60,34	12,46	100,74	64,26	81,23
Chocó	66,24	40,37	115,75	73,83	75,22
La Guajira	72,66	33,52	116,19	74,12	75,04

Tabla 10. Ranking con Distancia Euclidiana para estación de Nariño

Estación	MAE (W/ m^2)	MBE (W/ m^2)	RMSE (W/ m^2)	NRMSE (%)	R^2 (%)
Valle del Cauca	131,75	122,03	164,92	105,20	49,70
Meta	168,64	164,91	205,39	131,02	21,99
Antioquía	60,60	26,20	106,47	67,92	79,04
Caquetá	56,05	26,76	104,81	66,86	79,69
Norte de Santander	67,33	31,82	114,56	73,08	75,73
Casanare	62,29	33,62	110,26	70,34	77,52
Chocó	59,65	19,80	102,37	65,30	80,62
Bolívar	59,08	28,75	106,60	68,00	78,99
La Guajira	60,11	33,75	109,75	70,01	77,73

Tabla 11. Ranking con WMD usando FI para estación de Nariño.

Al añadir el FI como matriz *W* para establecer una métrica basada en la WMD, se obtienen los resultados mostrados en la tabla 11, los cuales son similares a los obtenidos únicamente a partir de la distancia euclidiana. Para tener en cuenta las posibles correlaciones entre los descriptores, se propone la utilización de PCA sobre los vectores característicos de cada estación tras haberle aplicado la matriz de pesos a cada uno.

Estación	MAE (W/ m^2)	MBE (W/ m^2)	RMSE (W/ m^2)	NRMSE (%)	R^2 (%)
Chocó	125,20	-108,17	203,00	129,50	23,79
Caquetá	90,96	-85,44	155,12	98,95	55,50
Casanare	113,29	-98,83	193,72	123,57	30,61
Bolívar	120,58	-81,33	190,21	121,34	33,09
Antioquía	63,41	-9,98	104,83	66,87	79,68
La Guajira	111,54	71,71	130,62	83,32	68,45
Meta	69,92	-38,78	113,53	72,42	76,17
Norte de Santander	94,50	-85,57	140,79	89,81	63,34
Valle del Cauca	61,68	23,40	106,46	67,91	79,04

Tabla 12. Ranking con WMD-PCA usando FI para estación de Nariño.

La tabla 12 muestra que al aplicar PCA no se obtiene un resultado mejor al de utilizar todos los datos, debido a que, al darle la misma importancia a todas las componentes principales, incluyendo aquellas que representan mínima varianza, no se tienen en cuenta aquellos descriptores realmente relevantes para diferenciar las estaciones entre sí. El principal ejemplo de esto es que esta es la única prueba en la que los datos de La Guajira no son unos de los menos relevantes. Para superar este inconveniente, se adiciona la matriz *L* de valores propios de la transformación PCA para darle más importancia a las componentes principales que representan más varianza. De esta manera se aplica nuevamente la WMD, pero en el espacio de PCA y usando esa matriz *L*.

Estación	MAE (W/ m^2)	MBE (W/ m^2)	RMSE (W/ m^2)	NRMSE (%)	R^2 (%)
Antioquía	132,46	104,38	154,49	98,55	55,86
Chocó	65,58	-24,83	114,74	73,19	75,66
Caquetá	133,09	112,47	154,46	98,53	55,88
Valle del Cauca	60,94	33,48	108,71	69,34	78,15
Meta	63,62	22,22	113,55	72,44	76,15
Bolívar	68,63	38,36	109,86	70,08	77,68
Casanare	63,36	34,66	111,34	71,02	77,08
Norte de Santander	65,03	22,70	104,12	66,42	79,95
La Guajira	65,83	40,63	112,88	72,00	76,44

Tabla 13. Ranking con WMD-PCA y matriz *L* usando FI para estación de Nariño.

La tabla 13 evidencia que este escenario requirió menos estaciones para alcanzar el mejor resultado posible, aunque los resultados en términos del R^2 son ligeramente inferiores. Esto puede ocurrir por la inicialización aleatoria de los pesos o la manera en que se crean los mini-batches de entrenamiento de la red neuronal.

8.6 Resumen de las Metodologías

Para tener un panorama más claro de las diferentes metodologías descritas es necesario conocer los resultados que estas tuvieron para todas las estaciones, debido a esto la tabla 14 resume los indicadores de rendimiento obtenidos en el mejor escenario de cada una de estas metodologías.

Metodología	MAE (W/ m^2)	MBE (W/ m^2)	RMSE (W/ m^2)	NRMSE (%)	R^2 (%)	BEST
All data	58,41	-2,16	100,74	57,31	84,96	9
G	52,64	0,61	98,55	56,14	85,59	6,2
SFS	49,58	2,33	94,97	54,01	86,69	6,6
PCA-WMD	50,98	-2,28	97,00	55,07	86,25	6,8
Kernel	51,58	-1,59	97,85	55,60	85,90	7,1
Dist. Euclidiana	50,82	-1,37	96,29	54,79	86,32	7,2
WMD-FI	51,09	-0,43	96,83	55,03	86,19	5,4
PCA-WMD-FI	51,65	-0,70	97,45	55,51	85,91	7,3
L-PCA-WMD-FI	51,60	-3,36	97,72	55,64	85,84	6,6

Tabla 14. Resumen de Metodologías.

De acuerdo con la tabla 14, el mejor escenario de todas las metodologías suele alcanzar indicadores de rendimiento similares. Prueba de esto es que en porcentaje ninguna metodología supera a otra por más de 1% o de 3W/m^2. La última columna de esta tabla indica en promedio cuantas estaciones se necesitaron para alcanzar el mejor resultado de cada metodología. Esto permite observar claramente que la métrica basada únicamente en la WMD usando el FI calculado en la primera parte de este trabajo requiere menos estaciones o subconjuntos de entrenamiento que todas las otras metodologías, por lo cual puede ser más efectivo para descartar aquellas estaciones que no deben tenerse en cuenta en el entrenamiento.

9 CONCLUSIONES

En términos generales, se pudo observar la efectividad de las redes neuronales para aprender de mediciones históricas y predecir la irradiación solar en Colombia. Esta predicción debe ser elaborada utilizando parámetros satelitales y en sitio debido a que ambos demostraron tener una alta importancia en el entrenamiento de las redes neuronales. Entre los descriptores más importantes se encontraron mediciones tomadas no sólo en el instante previo a la predicción, sino variables tomadas 24 o 48 horas antes, demostrando la importancia del fenómeno estacional en estas series de tiempo, por lo cual no deben usarse únicamente los últimos valores de cada serie de tiempo. Adicionalmente, el modelo de cielo abierto CS demostró tener una relevancia elevada en el rendimiento de la predicción, por lo cual es un descriptor que debe tenerse en cuenta para modelos predictivos basados en redes neuronales, a pesar de ser otro tipo de modelo predictivo que no tiene en cuenta fenómenos vinculados con las nubes.

Con respecto a las variaciones de la nubosidad, la cual es la principal causante de la variación en la irradiación solar, se comprobó que la red neuronal es capaz de tenerlas en cuenta en la predicción para horizontes temporales inferiores a 3 horas. En escenarios con un horizonte mayor, a pesar de no predecir correctamente los altibajos de la irradiación causados por las nubes, el modelo es capaz de predecir la tendencia general del sol y por ende genera resultados equiparables, e incluso superiores, a los mostrados en estudios previos para este mismo horizonte. Sin embargo, es necesario reconocer que estudios en diferentes lugares no son directamente contrastables porque la facilidad de la predicción está directamente relacionada con las características climatológicas del lugar.

El estudio sobre estaciones que no se tuvieron en cuenta en el entrenamiento permitió comprobar que, para casos específicos, es recomendable utilizar un subconjunto de los datos en lugar de usar todos los disponibles. Esto ocurre principalmente porque existen determinados puntos coordenados que, por sus características geográficas y climatológicas, son más semejantes que otros al lugar de predicción.

Para determinar los subconjuntos de datos que deben ser utilizados en el entrenamiento, se puede utilizar un algoritmo de ordenamiento por distancia WMD basados en la importancia de cada descriptor para la red neuronal. Además, para un resultado superior, se puede utilizar un algoritmo de optimización que resuelva un problema de ranking. Sin embargo, esta segunda metodología debe enfocarse en que las estaciones seleccionadas por el algoritmo de optimización sean iguales a aquellas que deberían escogerse de acuerdo con el ranking ideal sin importar que el orden de estas estaciones hasta el umbral de selección sea distinto.

Con este nuevo enfoque, se propone la utilización de *Kernels* más complejos utilizando la idea de composición y combinación lineal tratados en este documento, y otros métodos de diseño que no se exploraron en este trabajo, de tal forma que se pueda encontrar un espacio en que las estaciones seleccionadas por la métrica sean iguales a las sugeridas por el algoritmo SFS. Adicionalmente, se puede utilizar un método más robusto que el algoritmo SFS para encontrar un subconjunto de entrenamiento que permita obtener mejores resultados, teniendo en cuenta que el incremento del costo computacional no sea demasiado elevado.

Por último, es necesario recordar que esta es una primera aproximación a la predicción de irradiación solar en Colombia a partir de redes neuronales. Por este motivo, los resultados expuestos deben ser contrastados con información adicional proveniente de más estaciones meteorológicas y otras metodologías que permitan obtener resultados superiores. De esta manera, se podrá elaborar una herramienta que, a partir de los descriptores necesarios, pueda elaborar predicciones en tiempo real que permita facilitar el problema de despacho energético en Colombia, abriendo la puerta a un incremento de la energía solar fotovoltaica disponible en la canasta energética del país.

10 AGRADECIMIENTOS

A Luis Felipe Giraldo por todo el tiempo invertido en consejos, explicaciones, revisiones bibliográficas y propuestas que permitieron la culminación de esta investigación con los resultados expuestos. A Gilberto Díaz y Alfredo Avendaño por sus ideas y conocimientos en Machine Learning que me facilitaron el desarrollo de los algoritmos de predicción. Finalmente, a todos mis amigos y familia por su apoyo moral durante todo el tiempo invertido en este trabajo.

11 REFERENCIAS

[1] S. Chow, E. Lee and D. Li, "Short-term prediction of photovoltaic energy generation by intelligent approach", Energy and Buildings, vol. 55, pp. 660-667, 2012.

[2] R. Inman, H. Pedro and C. Coimbra, "Solar forecasting methods for renewable energy integration", Progress in Energy and Combustion Science, vol. 39, no. 6, pp. 535-576, 2013.

[3] Global Modeling and Assimilation Office (GMAO), MERRA-2, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), 2015.

[4] M. C. Pantoja & H. E. Mutis, "Análisis comparativo de pronósticos realizados con redes neuronales, modelos Arima y procesos Garch para series de tiempo no estacionarias", Proyecto de Grado para optar al título de Maestria en Ingeniería Industrial, Universidad de los Andes, 2004.

[5] Sobri, S., Koohi-Kamali, S. and Rahim, N. (2018). Solar photovoltaic generation forecasting methods: A review. *Energy Conversion and Management*, 156, pp.459-497.

[6] Kulis, B. (2013). Metric Learning: A Survey. *Foundations and Trends® in Machine Learning*, 5(4), pp.287-364.

[7] Gheyas, I. and Smith, L. (2010). Feature subset selection in large dimensionality domains. *Pattern Recognition*, 43(1), pp.5-13.

[8] Wold, S., Esbensen, K. and Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3), pp.37-52.

[9] Enel Green Power Colombia, "Parque Solar Fotovoltaico El Paso", *Consejo Nacional de Operación*, 2018. [Online]. Available: https://www.cno.org.co/sites/default/files/archivosAdjuntos/enel_fotovoltaico_el_paso.pdf. [Accessed: 15- Nov- 2018].

[10] C. Burges, R. Ragno, and Q. Le. Learning to rank with nonsmooth cost functions. In Advances in Neural Information Processing Systems 18, pages 395–402. MIT Press, Cambridge, MA, 2006.

[11] Kingma, P. Diederik and J. Ba, “Adam: A method for stochastic optimization”, arXiv preprint arXiv:1412.6980, 2014

[12] "ESDS Policies | Earthdata", Earthdata.nasa.gov, 2018. [Online]. Available: https://earthdata.nasa.gov/earth-science-data-systems-program/policies. [Accessed: 23- Nov- 2018].

[13] W. F. Holmgren, C. W. Hansen and M. A. Mikofski, "pvlib python: a python package for modeling solar energy systems", Journal of Open Source Software, vol. 3, no. 29, p. 884, 2018.

[14] F. Hocaoğlu, Ö. Gerek and M. Kurban, "Hourly solar radiation forecasting using optimal coefficient 2-D linear filters and feed-forward neural networks", Solar Energy, vol. 82, no. 8, pp. 714-726, 2008.