UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA MECÁNICA

EVALUACIÓN DE MÉTODOS SUPERVISADOS DE APRENDIZAJE AUTOMÁTICO EN PRONÓSTICOS SOLARES INTRA-HORARIOS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL MECÁNICO

JOSÉ IGNACIO MOYA HERNÁNDEZ

PROFESORA GUÍA:
MÓNICA ZAMORA ZAPATA

MIEMBROS DE LA COMISIÓN:
VIVIANA MERUANE NARANJO
WILLIAMS CALDERÓN MUÑOZ

SANTIAGO DE CHILE
2023

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL MECÁNICO
POR: **JOSÉ IGNACIO MOYA HERNÁNDEZ**
FECHA: 2023
PROF. GUÍA: MÓNICA ZAMORA ZAPATA

EVALUACIÓN DE MÉTODOS SUPERVISADOS DE APRENDIZAJE AUTOMÁTICO EN PRONÓSTICOS SOLARES INTRA-HORARIOS

En el último tiempo, el planeta se ha visto sometido ante la crisis climática, un fenómeno que atenta contra su prosperidad producto de las elevadas emisiones de gases de efecto invernadero. En base a lo anterior se requiere una acción rápida, sobre todo por parte de las industrias con mayor cuota de responsabilidad, como lo es el sector energético. Afortunadamente, a nivel nacional, ha habido un gran incremento en la incorporación de fuentes de energías limpias, entre ellas, la de mayor contribución a la matriz energética, la energía solar fotovoltaica (PV). Si bien esto es algo positivo, trae consigo nuevos desafíos a futuro para los agentes involucrados en la logística del sistema energético nacional, principalmente dada la dependencia a las condiciones meteorológicas que presenta la energía PV. Esto ha generado un alza en el desarrollo e investigación de modelos de pronóstico de irradiancia solar incidente, que permitan reducir la incertidumbre producida por las fluctuaciones en el recurso solar de manera confiable, contribuyendo a la operación de un sistema interconectado más eficiente.

En función de lo anterior, el presente trabajo de título se enfoca en desarrollar y evaluar modelos de pronóstico de irradiancia solar de corto plazo, desarrollados por métodos de aprendizaje automático para la localidad de Tocopilla, región de Antofagasta, en el norte de Chile. Para lograr lo anterior se proponen cuatro métodos de aprendizaje automático a desarrollar: k-Nearest Neighbors (kNN), Gradient Boosting (GB) y dos arquitecturas de redes neuronales, una red neuronal Feed Forward (FFNN) y otra red recurrente, Long-Short Term Memory (LSTM), los cuales son evaluados para predicciones con distintos horizontes de tiempo intra-horarios. Especialmente, se evalúan sus desempeños individuales en base a su distribución de error para dos componentes de irradiancia, la global (GHI) y la directa (DNI), además de comparar en relación a un método de referencia, en este caso kNN, los otros tres métodos restantes.

Los resultados muestran que los cuatros modelos evaluados presentan desempeños satisfactorios para la predicción de irradiancia GHI y DNI para horizontes de tiempo menores a 30 minutos. Con horizontes de tiempo superiores, la red LSTM es la que logra sobrellevar esta complicación, presentando la menor distribución de error, entregando así un pronóstico confiable y preciso. En este aspecto, este método presenta la mayor capacidad de pronóstico para todos los escenarios evaluados, con un errores de pronóstico MAE que van desde un 0.04 % a un 1.98 % en el mejor y peor escenario respectivamente, lo que llega a ser hasta un 27 % menor que el método de referencia usado para GHI e incluso presenta una diferencia aún mayor en el caso de DNI, reafirmando así la utilidad de su capacidad de recursión y memoria selectiva inherente de la red LSTM respecto a sus rivales.

A mis padres, familia y amigos
por el apoyo incondicional.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

En el mundo moderno, uno de los temas que ha ido tomando cada vez más relevancia dentro de los últimos años, es la eminente crisis climática a la que se ve enfrentada la especie humana. Esto ha provocado que se deba tomar acción en diversos sectores industriales para reducir la generación de gases de efecto invernadero. Entre dichos sectores industriales, uno de los más relevantes, es el sector energético, pues es una de las industrias que más poder tiene de inclinar la balanza y producir un cambio real en esta materia. En el último tiempo, el desarrollo en lo que respecta a la incorporación de fuentes de energía renovables no convencionales (ERNC) a la matriz energética nacional ha tenido un incremento sin precedentes. A septiembre de 2022 el 40.69 % de la capacidad instalada a nivel país proviene de tecnologías asociadas a las ERNC, donde en particular, la energía solar fotovoltaica (PV) aporta con más de la mitad de esa contribución, representando un 23.54 % del total, lo cual equivale a 7.76 GW [1].

En base a lo anterior, es evidente que la energía solar PV representa un componente fundamental en la transición a una matriz energética más limpia. Sin embargo, trae consigo un desafío importante, específicamente ligado a las materias de logística y distribución de la energía. Esto resultado de un defecto que presenta este tipo de tecnología, y es que tiene una gran susceptibilidad a las condiciones climáticas dado el efecto directo que tienen estas sobre la irradiancia solar, provocando fluctuaciones en el recurso solar, lo cual se traduce directamente también en fluctuaciones para la generación de potencia eléctrica. Por esta razón es importante tener en cuenta que de no abordar de manera eficiente dichas fluctuaciones, la estabilidad del sistema eléctrico se verá comprometida, tanto en la disponibilidad energética como en el aspecto económico, ya que para compensar el déficit que generan dichas fluctuaciones muchas veces se requiere la intervención inmediata de otras fuentes de energía más convencionales, afectando así el costo marginal de producción.

Es por esta razón que es clave la implementación de modelos matemáticos que permitan disminuir la incertidumbre. La previsión de irradiancia solar es la base del pronóstico de potencia eléctrica y una predicción precisa de la potencia generada tiene la capacidad de mejorar el nivel de estabilidad y seguridad de la red eléctrica como también guiar a una mejora en la economía nacional. Además, la predicción de potencia es importante para los agentes involucrados en la generación, pues contribuye positivamente en la toma de decisiones, ya sea al momento de diseñar diferentes tipos de plantas o estaciones de generación, determinar reservas para almacenamiento de energía y proporcionar información para la negociación del mercado eléctrico, principalmente en el corto plazo [2].

En el último tiempo, el desarrollo y la investigación de nuevos modelos de pronóstico de irradiancia solar ha estado en alza, pues se busca mejorar las estimaciones realizadas para el sector energético. Estos modelos se pueden dividir en dos clases principales: modelos físicos y modelos estadísticos. Los modelos físicos se centran en ecuaciones capaces de describir el estado físico y la dinámica de la capa atmosférica. Estos modelos físicos consisten en ecuaciones no lineales bastantes complejas, que deben ser abordadas utilizando métodos numéricos para obtener soluciones aproximadas, lo cual implica un gran costo computacional. Por otro lado, los modelos estadísticos incluyen modelos en base a series de tiempo, datos satelitales, como también modelos basados en imágenes del cielo. Estos son de menor complejidad que los modelos físicos, lo que conlleva a un menor tiempo de cómputo para realizar el pronóstico [2]. Otra gran ventaja de este tipo de modelos, en particular de los modelos basados en series de tiempo, es la disponibilidad de datos para ser usados como datos de entrada, pues variables meteorológicas, tales como temperatura ambiente, humedad relativa, velocidad del viento, dirección del viento, entre otras, son de fácil acceso en la mayoría de los escenarios. En efecto, un pronóstico de irradiancia solar convencional se realiza usando como parámetros data histórica de irradiancia solar en conjunto a las variables ambientales mencionadas.

Es este tipo de modelos estadísticos, aplicados como tecnologías de aprendizaje automático (Machine Learning), una de las metodologías para abordar la fabricación de pronóstico de irradiancia solar, como por ejemplo el trabajo realizado por Pedro et al. [3], donde se busca evaluar técnicas de aprendizaje automático tales como k-Nearest-Neighbors (kNN) y Gradient Boosting (GB), para contrastarlas con modelos más clásicos de pronóstico como es el método de persistencia. Lo que destaca de este trabajo es el incluir nuevas fuentes de información como son imágenes del cielo en conjunto a la data histórica de irradiancia y parámetros ambientales, en búsqueda de una mayor precisión de la nubosidad y por ende de la irradiancia. En particular, en este trabajo los resultados evaluados muestran un error porcentual del orden del 3 % al 8 % para irradiancia global y de un 5 % a un 15 % para irradiancia directa, lo cual tiene incidencia directa en el set de datos utilizados, pero se sostiene como una referencia base a este tipo de métodos. Otra aproximación interesante que ha tomado el desarrollo de pronóstico solar es el uso de técnicas de aprendizaje profundo (Deep Learning) pues han demostrado gran desempeño, haciendo uso de arquitecturas variadas de redes neuronales (NN) multicapa [2], como también redes recurrentes (RNN), donde son de gran interés las redes Long-Short Term Memory (LSTM) [4].

En razón de lo antes expuesto, en este trabajo de título se propone analizar de manera efectiva los resultados obtenidos por distintos algoritmos supervisados de aprendizaje automático (ML) y redes neuronales profundas (DNN) junto con sus rendimientos, para la estimación de irradiancia solar a corto plazo en la zona norte de Chile, en base a parámetros ambientales y datos históricos de irradiancia solar.

1.1. Objetivos y alcances

1.1.1. Objetivo general

Evaluar modelos de pronóstico de irradiancia solar, de corto plazo, desarrollados mediante técnicas de aprendizaje automático, en base a datos obtenidos para la provincia de Tocopilla,

en la región de Antofagasta en el norte de Chile.

1.1.2. Objetivos específicos

- Evaluar los rendimientos de los modelos aplicados para el conjunto de datos seleccionado bajo distintos horizontes de pronóstico intra-horarios.
- Estudiar el comportamiento de los pronósticos de irradiancia solar global y de irradiancia normal directa.
- Comparar la capacidad de pronóstico de cada modelo en base a un modelo de referencia.
- Estudiar resultados obtenidos por las métricas de error y analizar causalidad.

1.1.3. Alcances

En base al objetivo general, el alcance del presente trabajo de título se concentra en modelar pronósticos solares a corto plazo de manera efectiva con cuatro métodos supervisados de aprendizaje automático, para operación del recurso solar.

Es importante mencionar que el trabajo de título posee un carácter de investigación respecto al modelamiento del sistema, esto implica que, si bien se buscará obtener resultados óptimos, no está dentro de los alcances mejorar métricas previamente establecidas por terceros.

Finalmente, este trabajo no apunta a evaluar un sistema de pronóstico operacional en el que las variables ambientales son dadas por un modelo de pronóstico del tiempo, sin embargo, la metodología podría adaptarse para ello.

Capítulo 2

Antecedentes

2.1. Irradiancia solar

La irradiancia o irradianza solar I se define como la cantidad de potencia incidente a causa de radiación electromagnética por unidad de área $[W/m^2]$, lo que de manera simplificada se puede expresar como:

$$I = \frac{P_{inc}}{A_s} \tag{2.1}$$

Donde P_{inc} es la potencia incidente y A_s el área de la superficie impactada. Dado que la potencia eléctrica generada por una celda fotovoltaica depende directamente del nivel de irradiancia recibida, esta magnitud es de gran importancia en el estudio de la generación de energía solar. Es importante notar que solo alrededor del 40 % de la energía solar interceptada en la parte superior de la atmósfera de la Tierra pasa a la superficie. La proporción es diferente acorde a las condiciones climáticas, dado que en un día soleado esta proporción será mayor en comparación a un día nublado [2].

La irradiancia solar puede ser estudiada en base a sus distintas componentes, donde destacan para propósitos del trabajo las siguientes:

- Irradiancia normal directa (DNI): Corresponde a la cantidad de radiación de onda solar recibida por unidad de área, que impacta sobre una superficie siempre perpendicular a los rayos incidentes.
- Irradiancia horizontal difusa (DHI): Corresponde a la cantidad de radiación de onda solar recibida por unidad de área, que no proviene directamente desde la dirección del sol, sino que ha sido desviada por partículas en la atmósfera, provocando que impacte a la superficie desde múltiples direcciones.
- Irradiancia global horizontal (GHI): Corresponde a la radiación de onda solar total por unidad de área medida en una superficie horizontal a la tierra. Esta es la componente más relevante para los cálculos de proyección de potencia eléctrica PV. Esta componente además incluye tanto a la DNI como la DHI, es decir, puede ser expresada en función de estas matemáticamente de la siguiente forma:

$$GHI = DHI + DNIcos(\theta_s) \tag{2.2}$$

Donde θ_s corresponde al ángulo zenital solar respecto a la superficie terrestre en el punto de interés.

Conocer las componentes de la irradiancia es de gran interés dentro de las tecnologías que usan el recurso solar, como ya se mencionó, la de mayor relevancia para pronóstico de potencia generada suele ser la GHI, pues considera un espectro más amplio, no obstante, para tecnologías de concentración solar (CSP) y paneles fotovoltaicos con función de seguimiento solar, las cuales se caracterizan por mantener una posición perpendicular al rayo incidente, la DNI es de gran importancia.

En la Figura 2.1 se presenta una muestra de las irradiancias medidas por la estación solarimétrica Crucero II, ubicada en la provincia de Tocopilla, entre los días 21 y 24 de enero de 2021. En particular la Figura muestra el comportamiento de la irradiancia frente a diferentes condiciones climáticas, tal que el primer y segundo día corresponde a días principalmente soleados, mientras que los últimos dos muestran la irradiancia cuando se tiene en días parcialmente nublados. Analizando el gráfico se puede extraer, que si bien es evidente la correlación de la irradiancia con respecto al paso horario durante la jornada del día, las fluctuaciones provocadas por el efecto de las condiciones meteorológicas, principalmente nubosidad, no pueden ser predichas de manera trivial, por lo que es necesario la aplicación de modelos de pronóstico con mayor grado de complejidad.

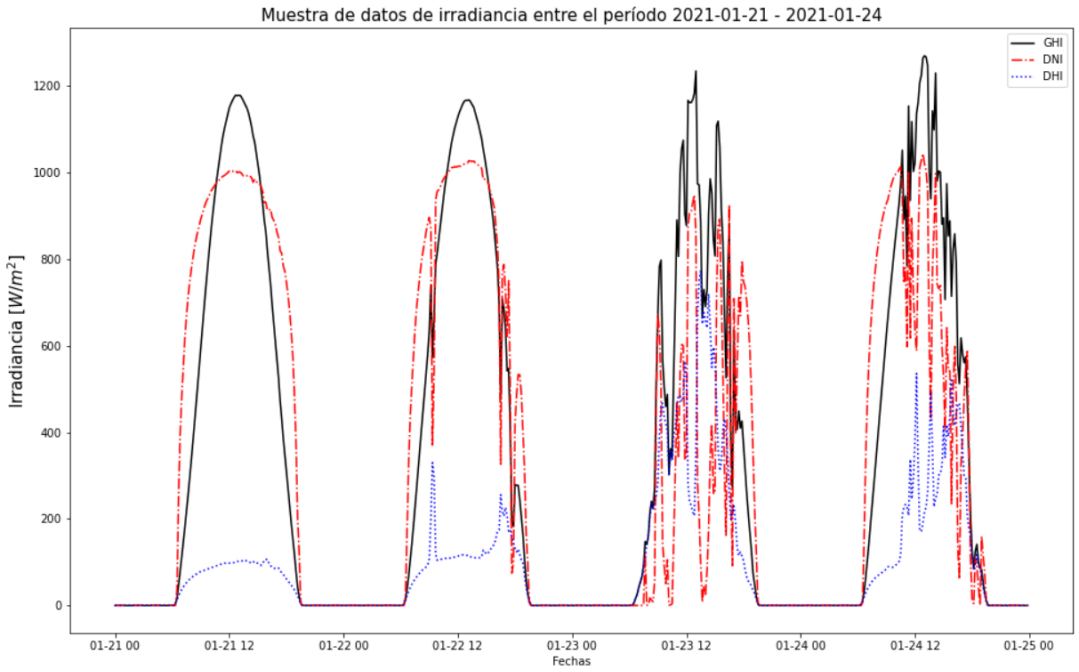

Figura 2.1: Comparación de las distintas componentes de las irradiancia solar registradas entre el 21 y 24 de enero de 2021.

2.2. Pronóstico solar

Un pronóstico solar fotovoltaico tiene como principal objetivo estimar la irradiancia solar incidente y posteriormente la potencia eléctrica generada, en un intervalo de tiempo, denominado horizonte de pronóstico, en base a los datos observados durante el período previo al intervalo deseado. Existen pronósticos a largo plazo, comprendidos en horizontes de un día en adelante, pronósticos de corto - mediano plazo con horizontes que van desde 1-6 horas y pronósticos de corto plazo o también llamados inmediatos, comprendidos en intervalos menores a 1 hora, usualmente entre 15 a 30 minutos. Es importante mencionar que mientras más largo sea el horizonte de pronóstico, menor será la precisión del algoritmo.

Los pronósticos solares son requeridos para abordar las necesidades de planificación, operación y equilibrio de las redes de transmisión eléctrica. Los pronósticos solares a corto plazo intra-horarios son particularmente útiles para labores operacionales en el sector, como la toma de decisiones en centrales eléctricas en tiempo real, el equilibrio de la red, despacho de unidades móviles y activación de otras fuentes de energía ante posibles fluctuaciones en el recurso solar que provoquen incumplimiento en la demanda energética, además de influir en otros aspectos tales como el comercio energético. Por otro lado los pronósticos con horizontes de tiempo más largos son de interés para empresas de servicios públicos, además de mejorar el rendimiento de la unidad encargada del equilibrio de la red, pues obtienen una mirada hacia el futuro, que si bien no es tan precisa como en un pronóstico a corto plazo, permite mantener un registro aproximado de la irradiancia o potencia eléctrica, tal que el balance no requiera una variación excesiva y por tanto que el tiempo de respuesta sea más corto [5].

En resumen, la predicción de irradiancia solar es una tecnología que permite la integración de un nivel cada vez mayor de energía solar en la matriz energética, pues mejora la calidad de la energía entregada a la red y reduce los costos secundarios asociados con la dependencia del clima. La combinación de estos dos factores ha sido la motivación impulsora para el desarrollo de un complejo campo de investigación que apunta a producir mejores capacidades de pronóstico solar y contribuir a la transición a una red con cada vez mayor contribución de ERNC [5].

2.3. Aprendizaje automático

El aprendizaje automático o machine learning (ML) es el campo de estudio encargado de desarrollar algoritmos que permiten a las computadoras extraer e identificar patrones de los datos. El aprendizaje automático posee una gran variedad de aplicaciones en la sociedad moderna, donde se encuentran principalmente los motores de búsqueda, análisis de mercado para los diferentes sectores de actividad, aplicaciones en diagnósticos médicos, detección de fraude en el uso de tarjetas de crédito, entre otras [6].

La gran mayoría de algoritmos de ML pueden ser clasificados en dos categorías, aprendizaje supervisado y aprendizaje no supervisado.

El objetivo del aprendizaje supervisado es generar una función a partir de los datos de entrenamiento, la cual se encarga de mapear desde los valores de los atributos que describen una instancia del valor, como puede ser los datos de entrada, hasta el valor de otro atributo,

conocido como atributo de salida o de destino, es decir, los resultados. La función creada busca ser capaz de predecir el valor de cualquier atributo de entrada, a partir de un conjunto de datos de ejemplo, denominados datos de entrenamiento, para lo cual debe generalizar patrones usando estos datos, y así, aplicarlos a los atributos de entrada que no han sido previsualizados, también denominados datos de prueba o testeo.

El aprendizaje supervisado opera mediante la búsqueda en una variada gama de funciones diferentes, con la finalidad de encontrar la que mejor representa la variación entre atributos de entrada y de salida. No obstante, en cualquier conjunto de datos de carácter más complejo, la cantidad de posibles combinaciones entre atributos de entrada y sus correspondientes caminos hacia el atributo de salida escala a un número muy elevado, tal que el algoritmo no puede probar todas las funciones posibles, por ende, un algoritmo de aprendizaje automático se diseña para preferir ciertos tipos de funciones mientras lleva a cabo su búsqueda. Dichas preferencias se conocen como sesgo de aprendizaje.

Por otro lado, el aprendizaje no supervisado se diferencia de su contraparte, en que no hay un atributo de salida definido. Esto implica que se trabaje únicamente con los atributos de entrada, los cuales muchas veces se trabajan como variables aleatorias, construyendo funciones de densidad para el conjunto de datos.

Una consecuencia de esto es que los algoritmos de aprendizaje no supervisado pueden ser aplicados sin tener que invertir tiempo en etiquetar las instancias del conjunto de datos con un atributo de salida. No obstante, esto implica que el proceso de aprendizaje, es decir, la inferencia de patrones, sea más compleja.

Cabe mencionar que también existen métodos que se encuentran en una categoría intermedia, y se les denomina métodos semisupervisados, los cuales requieren solo una parte del set de datos objetivo o de salida etiquetado.

2.4. Algoritmos de aprendizaje automático

En la presente sección se describen los cuatro algoritmos de aprendizaje automático para la generación de pronósticos de irradiancia solar de interés para este trabajo de título.

2.4.1. k-Nearest-Neighbors (kNN)

El algoritmo k-Nearest-Neighbors consiste en la clasificación de atributos de series de datos, en base a su similaridad con una serie de referencia. Es un modelo que destaca principalmente dado que no requiere de muchos parámetros para su entrenamiento, por lo tanto no presenta un costo computacional muy elevado y su desempeño es satisfactorio [7].

Considerando el caso referente a pronósticos, el modelo kNN predice los valores próximos correspondientes a una serie de tiempo. Sean atributos de entrada $\vec{x}$, para los cuales se desea estimar sus parámetros de salida. Sean $\overrightarrow{X}$ un subconjunto de los datos de entrada, destinados a entrenamiento, con sus respectivos datos de salida $\overrightarrow{Y}$. Cada uno de estos

subconjuntos contiene series de tiempo particulares. Se pueden definir de la siguiente manera.

$$\overrightarrow{X} = \{\overrightarrow{X}_1, \overrightarrow{X}_2, ..., \overrightarrow{X}_{N-1}, \overrightarrow{X}_N\} \tag{2.3}$$

$$\overrightarrow{Y} = \{\overrightarrow{Y}_1, \overrightarrow{Y}_2, ..., \overrightarrow{Y}_{N-1}, \overrightarrow{Y}_N\} \tag{2.4}$$

El modelo busca determinar las series más similares entre la serie de interés $\vec{x}$ y cada una de las series $\overrightarrow{X}_i$ de referencia. La cantidad de parámetros de referencia queda dado por el número de vecinos del modelo, es decir el modelo entregará las k series más similares, esto en base a la distancia euclidiana e_i, dada por la siguiente expresión.

$$e_i = \sqrt{\sum_j (x_j - X_{i,j})^2} \tag{2.5}$$

Donde x_j, $X_{i,j}$ están contenidos en $\vec{x}$ y $\overrightarrow{X}$ respectivamente. Una vez se encuentran determinadas las distancias, queda tomar el promedio de las k series de salida correspondientes a la menor distancia euclidiana (siendo k número entero a definir), dadas por $\overrightarrow{Y}_i$, lo cual se puede expresar como:

$$\vec{y} = \frac{\sum_{i=1}^{k} \overrightarrow{Y}_i}{k} \tag{2.6}$$

Siendo $\vec{y}$ los resultados del pronóstico (*output*) asociado a las series de tiempo de los atributos de entrada $\vec{x}$.

2.4.2. Gradient Boosting (GB)

El método de Gradient Boosting (GB) es un algoritmo que se usa tanto para aplicaciones de clasificación como de regresión. Consiste en proporcionar un modelo de predicción en base al ensamble de modelos de predicción más débiles, usualmente árboles de decisión [3]. En particular, para problemas de regresión, se centra en encontrar, dado un conjunto de datos, una función que minimice una función de pérdida asignada [8].

El modelo requiere como valores de entrada un conjunto de datos de entrenamiento y una función de pérdida. Sea un conjunto de n datos, donde x_i representa los datos de entrada que permiten predecir los elementos de salida, también conocidos como *features* e y_i los datos de salida de entrenamiento observados o *targets*, es decir el set de datos se compone como:

$$\{(x_i, yi)\}_{i=1}^{n} \tag{2.7}$$

Y sea una función de pérdida diferenciable $L(y_i, F(x))$, en este caso y la más comúnmente usada para el método de GB:

$$L(y_i, F(x)) = \frac{1}{2}(y_i - F(x))^2 \tag{2.8}$$

Donde $F(x)$ corresponde al valor predicho para el valor observado y_i. Una forma sencilla de cuantificar el error de datos predichos en una muestra es en base a residuos. Se define el residuo r como la diferencia simple entre el valor observado y el valor predicho, tal como:

$$r = (Observado - Predicho) \tag{2.9}$$

Esto indica que en base a la función de pérdida seleccionada, se puede evaluar qué tan bien una línea recta se ajusta a los datos usando la suma de los residuos cuadrados. Cabe mencionar que el factor $\frac{1}{2}$ en la ecuación 2.8 se usa ya que durante el desarrollo del modelo se debe derivar la función de pérdida, de tal manera que al realizar este proceso, el resultado sea directamente el residuo y facilite los cálculos.

El primer paso es inicializar el modelo, esto se hace con un valor constante $F_0(x)$, dado por:

$$F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma) \tag{2.10}$$

La sumatoria permite que se vaya agregando una función de pérdida para cada valor observado, luego $F_0(x)$ busca el valor γ que minimice la sumatoria. Sea, por ejemplo $n = 3$, eso implica que el valor inicial $F_0(x)$ se pueda escribir de la siguiente manera:

$$F_0(x) = \frac{1}{2}(y_1 - \gamma)^2 + \frac{1}{2}(y_2 - \gamma)^2 + \frac{1}{2}(y_3 - \gamma)^2 \tag{2.11}$$

Dado que se requiere minimizar esta función, se deriva respecto a γ y se iguala a 0, como se aprecia a continuación:

$$\frac{\partial F_0(x)}{\partial \gamma} = -(y_1 - \gamma) - (y_2 - \gamma) - (y_3 - \gamma) = 0 \tag{2.12}$$

$$\gamma = \frac{y_1 + y_2 + y_3}{3} \tag{2.13}$$

Es decir, dada esta función de pérdida, el valor de γ que minimiza la suma, corresponde al promedio simple de los valores observados. Generalizando lo anterior, la constante que se necesita para inicializar el modelo está dado por:

$$F_0(x) = \gamma = \frac{\sum_{i=1}^{n} y_i}{n} \tag{2.14}$$

Este valor corresponde al primer árbol de decisión del modelo, conformado por solo una hoja o nodo, la cual predice todos los valores de la muestra como el promedio simple. Esto evidentemente es una aproximación demasiado gruesa, lo cual se corrige con el siguiente paso, el cual se encarga de generar un loop que crea el resto de árboles de decisión aditivos al modelo.

GB requiere fijar la cantidad M de árboles de decisión a generar, si bien este número puede variar de acuerdo a las distintas aplicaciones, es común considerar $M = 100$. Luego para cada $m \in M$ se debe calcular lo siguiente:

$$r_{i,m} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \tag{2.15}$$

Donde $r_{i,m}$ es el residuo de los datos en la posición i del árbol m, en otras palabras, se calcula el residuo para cada valor observado de la muestra.

El siguiente paso es ajustar un árbol de regresión para los valores de $r_{i,m}$ obtenidos, de tal manera que se creen regiones terminales u hojas $R_{j,m}$ para $j = 1, ..., J_m$, siendo j la cantidad de valores en cada hoja indexada al respectivo árbol m. De tal manera que lo que se predice en este paso serán los residuos en lugar de los valores observados *targets* y posteriormente se etiqueta cada una de las hojas como $R_{j,m}$.

Sigue calcular el $\gamma_{j,m}$ de cada uno de los valores correspondientes a cada hoja, siguiendo un procedimiento muy similar al paso inicial del modelo, de tal manera que para cada $j = 1, ..., J_m$ se calcula:

$$\gamma_{j,m} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{j,m}} L(y_i, F_{m-1}(x_i) + \gamma) \tag{2.16}$$

De manera análoga al desarrollo de la ecuación 2.10, dada la función de pérdida utilizada, el valor de salida para $\gamma_{j,m}$ siempre será el promedio de los residuos que estén contenidos dentro de la respectiva hoja j en el árbol m.

El último paso es actualizar el valor predicho en el paso anterior, $F_0(x)$, o de manera más

general, el valor $F_{m-1}(x)$, lo cual se realiza con la siguiente ecuación:

$$\begin{aligned} F_m(x) &= F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{j,m} I(x \in R_{j,m}) \\ I(x \in R_{j,m}) &= \begin{cases} 1 \ si \ x \in R_{j,m} \\ 0 \ si \ no \end{cases} \end{aligned} \tag{2.17}$$

La nueva predicción toma en cuenta la predicción realizada en el árbol previo $m-1$ y el presente árbol de decisión creado en el paso anterior. La suma de la ecuación 2.17 indica que se debe sumar los valores de salida $\gamma_{j,m}$ para todas las hojas, $R_{j,m}$, donde la muestra x pueda ser encontrada. Además se define ν como la tasa de aprendizaje o *learning rate*, el cual se encarga de ponderar el efecto que tiene cada árbol sobre la predicción final, un valor pequeño reduce el sesgo, mejorando la precisión del algoritmo en el largo plazo, sin embargo, se debe tener en cuenta que esto aumenta el tiempo de cómputo del modelo.

Finalmente, cuando $m = M$, $F_M(x)$ corresponde al output final del algoritmo GB.

2.5. Redes neuronales

El estudio del cerebro humano se remonta hacia cientos de años en el pasado y con el avance de nuevas tecnologías, era natural tratar de aprovechar el proceso de pensamiento que este posee. Las redes neuronales, así como otras técnicas de ML, presentan una habilidad notable para extraer significado de situaciones complicadas o datos imprecisos, se pueden utilizar para extraer patrones y detectar tendencias que son demasiado complejas para ser notadas a simple vista o en base a un análisis con métodos más convencionales [9].

Una red neuronal (NN) consiste en un modelo matemático el cual se construye en base la conexión de arreglos de capas de neuronas o nodos, dichos arreglos pueden tener distintas configuraciones acordes a una arquitectura definida. Cada neurona toma un conjunto de datos de entrada y los redirecciona hacia los valores de salida. La manera en la que interaccionan se centra en el uso de pesos numéricos asignados a cada neurona, estos pesos se van ajustando a medida que se entrena al modelo. En otras palabras, se puede entender a una neurona como una función de regresión lineal de múltiples entradas, con la consideración de que la salida de esta función de regresión lineal pasa a través de otra función, a la cual se le denomina función de activación, de carácter comúnmente no lineal.

Las funciones de activación gobiernan el comportamiento de las neuronas, pues dictan el comportamiento y combinación de los valores de entrada a la neurona, los pesos y sesgos asociados. El resultado que se obtiene posterior a la aplicación de la función de activación es lo que ingresa a la siguiente capa de neuronas [10]. Este proceso es el principio que aplica para las redes neuronales más tradicionales, donde la propagación de información es en una sola dirección. A este tipo de redes se les conoce como Feed Forward Neural Network (FFNN).

El desarrollo de técnicas de aprendizaje profundo ha hecho que se experimente con distintas configuraciones respecto a los arreglos de las redes neuronales en conjunto a variaciones en la dirección en la que estas propagan información. Para el presente trabajo de título son de interés las redes neuronales recurrentes (RNN), en particular su variación más común, que corresponde a las redes Long-Short Term Memory (LSTM), para ser comparada con los métodos estocásticos kNN y GB, además del arreglo más tradicional de las FFNN.

2.5.1. Feed Forward Neural Network (FFNN)

Este tipo de redes FFNN, como se mencionó anteriormente, representan el concepto más tradicional de una red neuronal. La red se compone de una primera capa de entrada (*input layer*), seguido de esto, se encuentra una o más capas ocultas (*hidden layers*) y finalmente la capa de salida con el valor resultante asociado, dicha capa se denomina como (*output layer*). Cada capa puede estar compuesta de múltiples neuronas, como se esquematiza en la Figura 2.2. Cada neurona i tiene asignada una serie de pesos contenidos en el vector W_i, los que se encargan de ponderar los valores de entrada que recibe de las neuronas de la capa anterior x_i. El valor de la neurona resultante y_i se puede representar con la siguiente ecuación vectorial:

$$y_i = f(W_i \cdot x_i + b) \tag{2.18}$$

Donde f corresponde a la función de activación encargada de desencadenar la salida de la neurona i. Esta función corresponde a una transformación no lineal la cual determina el comportamiento de la neurona para la resolución de su valor de salida, ejemplos de funciones de activación son la función ReLU, sigmoide, sinusoidal, hiperbólica entre otras. En cuanto a b, este factor se le conoce como el sesgo (*bias*) de cada capa. Su principal utilidad es que en caso de que haya valores de entrada nulos a la neurona, este factor permite que dichos nodos puedan ser activados igualmente.

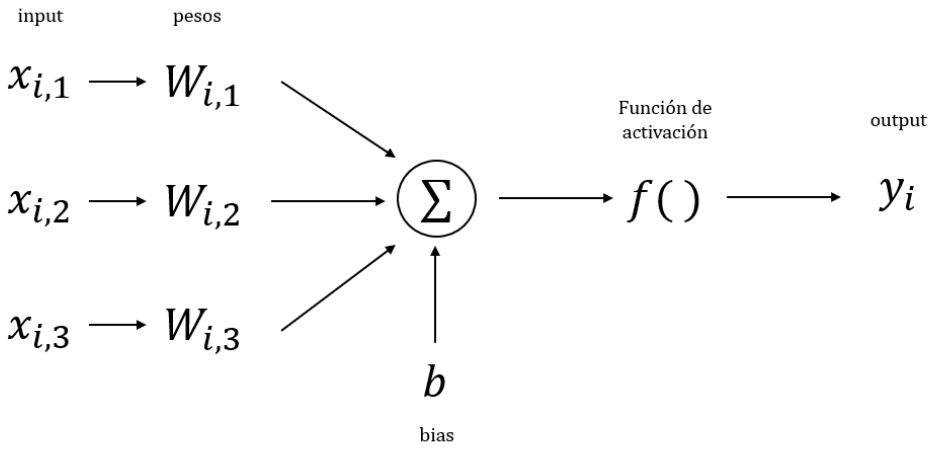

Figura 2.2: Esquema de una neurona artificial en una red multicapa.

2.5.2. Long-Short Term Memory (LSTM)

Las redes LSTM son una extensión de las redes neuronas recurrentes o RNN, las cuales tienen la cualidad de aprender de experiencias que han ocurrido en instantes de tiempo pasado como las RNN tradicionales, pero con una memoria más amplia y selectiva. Esto dado la incorporación de una celda de memoria y de puertas (*gates*) de olvido y de entrada. Cuando ambas se encuentran cerradas, el contenido de memoria de la celda permanecerá sin modificar entre un paso de tiempo y el siguiente. Esta estructura en base a puertas permite que la información pueda ser retenida por una gran cantidad de pasos de tiempo a la vez de ir deshaciéndose de la información que el modelo no considere relevante y así no sobrellenar la celda de memoria [10].

En otras palabras, cada celda LSTM tiene como atributos de entrada tanto información de la capa anterior x_t, al igual que en una FFNN, como también información correspondiente al paso de tiempo (*timestep*) seleccionado a revisar por el modelo h_{t-1}, este representa la memoria a corto plazo de la celda. Este proceso se repite con los atributos de salida de la celda, cuya información es propagada hacia las neuronas de la celda siguiente cumpliendo la misma labor.

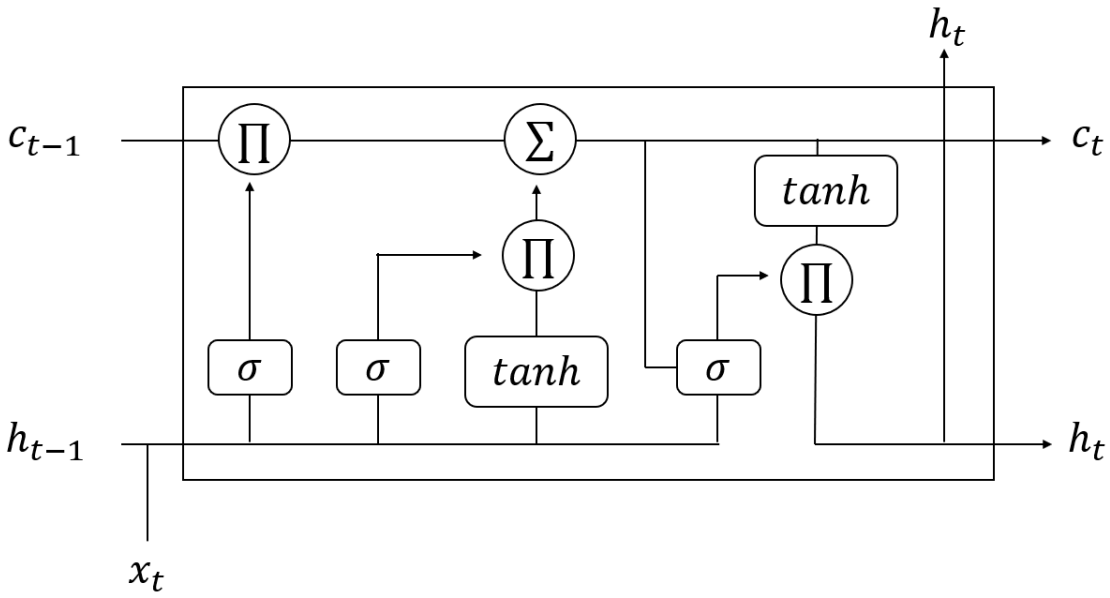

Figura 2.3: Diagrama de bloque de una celda LSTM.

La Figura 2.3 muestra una esquematización de una celda LSTM. El flujo de información sigue el siguiente orden, donde se tiene la puerta de entrada o *input gate*, la cual recibe los atributos de entrada a la celda x_t y h_{t-1}, los pondera y a través de una función sigmoide σ determina que porcentaje de esta información será añadida a la celda de memoria c_t, la cual ya contiene lo que proviene del paso de tiempo previo c_{t-1}. A esta celda de memoria también se le conoce como la memoria a largo plazo de la celda [11]. La operación que realiza la puerta de entrada se puede expresar como:

$$i_t = \sigma(W_i \cdot x_t + U_i \cdot h_{t-1} + b_i) \tag{2.19}$$

Donde W_i y U_i son los pesos pertenecientes a la puerta de entrada asociados a la información de x_t y h_{t-1} respectivamente, mientras que b_i es el *bias* de la puerta. De igual manera se comporta la puerta de olvido o *forget gate*, la cual pondera y determina el porcentaje de información que será eliminada de la celda de memoria.

$$f_t = \sigma(W_f \cdot x_t + U_f \cdot h_{t-1} + b_f) \tag{2.20}$$

Con lo anterior ya calculado, se actualiza la celda de memoria, haciendo uso de una función tangente hiperbólica de la siguiente manera:

$$c_t = f_t \cdot c_{t-1} + i_t \cdot tanh(W_c \cdot x_t + U_c \cdot h_{t-1} + b_c) \tag{2.21}$$

Por último, se tiene la puerta de salida o *output gate*, la cual pondera y determina el porcentaje de información de la celda de memoria c_t que se almacenará en la memoria de corto plazo h_t para la siguiente celda LSTM, esto mediante el uso de otra función tangente hiperbólica, como se aprecia a continuación:

$$o_t = \sigma(W_o \cdot x_t + U_o \cdot h_{t-1} + b_o) \tag{2.22}$$

$$h_t = o_t \cdot tanh(c_t) \tag{2.23}$$

Es importante notar que todas las puertas del bloque LSTM, es decir las puertas de entrada, olvido y salida tienen funciones de activación sigmoide con el propósito de mantener una restricción porcentual de la información, pues el rango de dicha función se comprende en el espacio de [0, 1], donde el valor 0 representa a la puerta completamente cerrada y por el contrario, el valor 1, permite el paso de toda la información. Por otro lado la función de activación de entrada y salida de la celda de memoria usada en la explicación corresponde a una función tangente hiperbólica, dado que es la que regularmente se usa, pero no representa una regla general [10].

Capítulo 3

Metodología

En la presente sección se describen los recursos y la metodología a seguir para el correcto desarrollo de este trabajo de título, desde la obtención inicial de los datos hasta el proceso de aplicación de las métricas seleccionadas para evaluar los modelos de aprendizaje automático

3.1. Obtención de datos

3.1.1. Estación solarimétrica

La investigación se lleva a cabo utilizando datos medidos por la Estación de Medición Solar Crucero II, la cual corresponde a un proyecto realizado en 2012, instalado por el Centro Nacional de Medio Ambiente (CENMA) por encargo de Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ). Crucero II registra irradiancia global horizontal (GHI), irradiancia normal directa (DNI), irradiancia horizontal difusa (DHI), junto a otras variables meteorológicas y ambientales. La ubicación de la estación de medición solar está en la comuna de María Elena, Región de Antofagasta (22.28° S, 69.57° W), al interior de la subestación Crucero.

Figura 3.1: Imagen satelital de la ubicación del sistema de medición solar al interior de la subestación crucero, indicado por el marcador de color azul.

Los instrumentos con los que cuenta Crucero II para la medición de variables se listan en la Tabla 3.1 a continuación.

Tabla 3.1: Listado de equipos principales pertenecientes a la estación de medición solar Crucero II.

Equipo	Marca	Modelo
Anemómetro a 6 metros de altura	Theodor Friedrichs	4035.0000
Anemómetro a 12 metros de altura	Theodor Friedrichs	4035.0000
Veleta	Theodor Friedrichs	4123.0000
Termómetro e higrómetro	Theodor Friedrichs	3031.0050
Piranómetro GHI	Kipp and Zonen	CMP21
Piranómetro DHI	Kipp and Zonen	CMP21
Pirheliometro DNI	Kipp and Zonen	CHP1
Capturador de datos 1	Theodor Friedrichs	Combilog 1022
Capturador de datos 2	Campbell SCI	CR1000
Paneles solares	Kyocera	KD135SX-1PU
Seguidor solar	Kipp and Zonen	Solys 2

La estación de medición requiere de baterías para su funcionamiento. Las baterías proporcionan la energía requerida para el trabajo del capturador de datos (*datalogger*) y equipos asociados. El proceso de escaneo de canales de medición se realiza cada 1 segundo, sin embargo, el intervalo de registro de memoria del *datalogger* está configurado en ventanas de 10 minutos, es decir, cada 10 minutos se guardan los promedios de las mediciones escaneadas cada 1 segundo.

Los datos registrados por Crucero II comienzan a partir del 17 de agosto de 2012, lo que implica que dadas las ventanas de registro de 10 minutos, se tiene una serie de tiempo de 482,616 elementos, donde cada elemento corresponde a un vector con las mediciones de los atributos en ese instante de tiempo en particular. En la Tabla 3.2 se pueden apreciar los atributos que entrega el capturador de datos para el servicio de operación. Estos atributos son los que permiten el entrenamiento de los algoritmos de aprendizaje automático para ser capaces de predecir la irradiancia global horizontal y la irradiancia normal directa en diferentes horizontes de tiempo a futuro.

Tabla 3.2: Descripción de variables entregadas por el sistema de medición de datos de la estación solarimétrica Crucero II.

Atributos	Descripción
Fecha y hora	Fecha y hora correspondiente al registro de cada uno de los valores de la medición.
Irradiancia Global Horizontal	Irradiancia GHI en el plano horizontal registrada por la estación solarimétrica en W/m^2.
Irradiancia Normal Directa	Irradiancia DNI registrada por la estación solarimétrica en W/m^2.
Irradiancia Horizontal Difusa	Irradiancia DHI registrada por la estación solarimétrica en W/m^2.
Temperatura Ambiental	Temperatura ambiente registrada por la estación solarimétrica en $°C$.
Humedad Relativa	Humedad relativa registrada por la estación solarimétrica en formato porcentual.
Velocidad del Viento a 6 metros	Velocidad del viento medida a 6 metros de altura registrada por la estación solarimétrica en m/s.
Velocidad del Viento a 12 metros	Velocidad del viento medida a 12 metros de altura registrada por la estación solarimétrica en m/s.
Dirección del Viento	Dirección del viento registrada por la estación solarimétrica en grados sexagesimales.
Voltaje Batería	Voltaje de la bateria requerida en V para la operación del datalogger (recopilador de datos) y otros equipos.

3.2. Pre-procesamiento de los datos

Previo a la configuración de las arquitecturas, es importante pre-procesar los datos, lo cual, dadas las condiciones del presente trabajo de título, consiste en realizar la exploración de las variables contenidas en el set de datos y su posterior depuración.

3.2.1. Exploración de los datos

La exploración de los datos es el primer paso para el análisis de datos y selección de parámetros a utilizar en el modelamiento de arquitecturas de aprendizaje automático, pues permite sondear y visualizar los datos para una previa identificación de patrones, atributos relevantes, entre otros. En esta etapa se consideraron para el análisis datos correspondientes a todo el 2021, pero por claridad en la visualización, los gráficos muestran datos de los primeros cuatro meses del 2021, en particular, desde el 14 de enero hasta el 30 de abril, ya que los patrones observados se mantienen. Se realizan una serie de gráficos para comparar como se comportan las componentes de irradiancia GHI y DNI, respecto a las variables ambientales de interés en el set de datos. El primer caso de análisis está en la Figura 3.2, donde se presenta

un gráfico que muestra la correlación entre la temperatura ambiental y las irradiancias. El patrón que siguen tanto la GHI y DNI es claro, al aumentar junto al aumento de temperatura, hasta llegar a un máximo para luego descender paulatinamente dentro del ciclo de 1 día, dada la posición relativa del sol en el cielo.

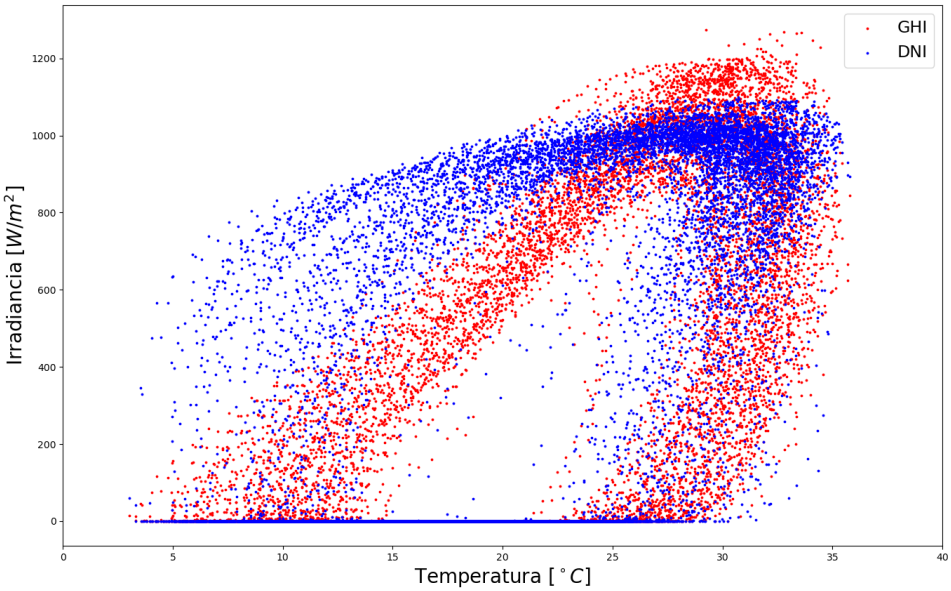

Figura 3.2: Gráfico de irradiancia GHI y DNI respecto a la temperatura ambiente entre los meses de enero y abril.

De manera similar, en la Figura 3.3 se analiza la correlación entre las irradiancias y la humedad relativa, donde se puede apreciar que en gran medida, los cúmulos mayores de puntos se encuentran para irradiancia alta y humedad baja, lo que además es coherente con el tipo de clima para la localidad estudiada.

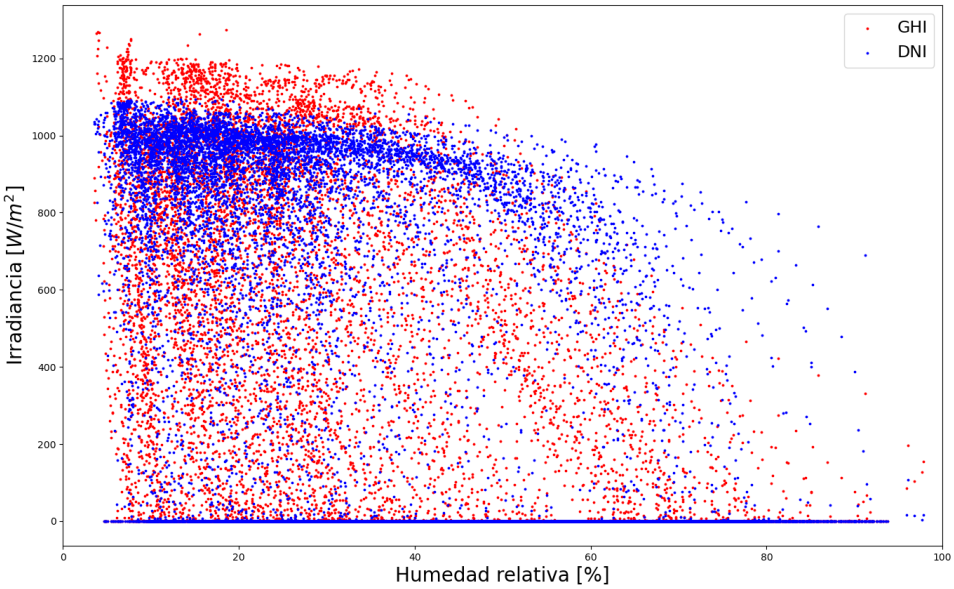

Figura 3.3: Gráfico de irradiancia GHI y DNI respecto a la humedad relativa entre los meses de enero y abril.

Lo anterior tiene sentido físico, pues la humedad relativa es un indicador directo de nubosidad, de tal manera que se confirma que este atributo tiene gran relevancia para el modelamiento de los pronósticos.

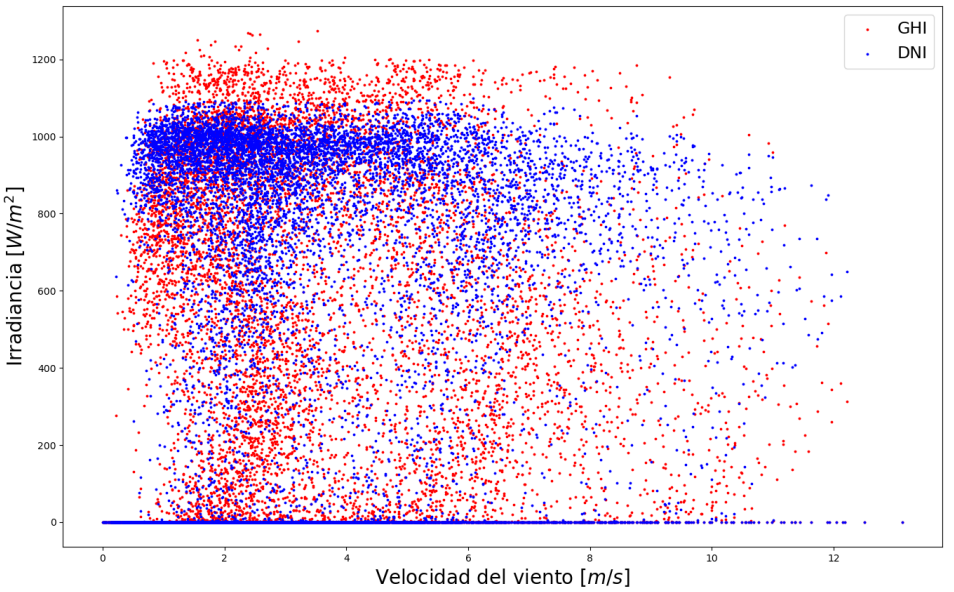

Figura 3.4: Gráfico de irradiancia GHI y DNI respecto a la velocidad del viento medida a 6 metros de altura entre los meses de enero y abril.

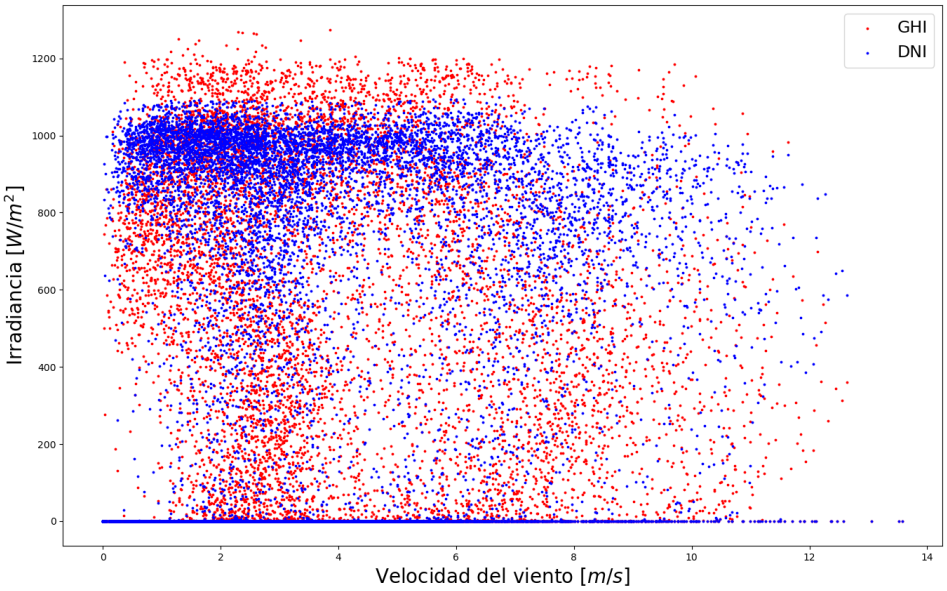

Figura 3.5: Gráfico de irradiancia GHI y DNI respecto a la velocidad del viento medida a 12 metros de altura entre los meses de enero y abril.

Finalmente, se tiene la correlación que hay entre las irradiancias y la velocidad del viento medida a distintas alturas, a 6 metros y 12 metros en la Figura 3.4 y Figura 3.5 respectivamente. En ambos gráficos no se logra apreciar una diferencia significativa, lo que implica que para el modelamiento de pronóstico estos dos atributos, al menos en cuanto a este análisis previo, son en gran parte equivalentes, lo que permite prescindir de una de estas mediciones. Respecto a los valores, la acumulación es preferencial para velocidades inferiores a 4 m/s, ya sea para irradiancias altas o bajas, lo cual, de manera preliminar, puede ser factor de que una menor velocidad del viento implica mayor estabilidad en el movimiento de nubosidad, ya sea en un día despejado o en uno nublado, sin embargo no se puede descartar que sea a causa simplemente de que la mayor cantidad de datos pertenezca a este espectro de la velocidad del viento.

Adicionalmente, es importante notar que tanto en la comparación respecto a la humedad relativa como a la velocidad del viento, se destaca un cúmulo lineal horizontal cuando la irradiancia es cero, lo cual viene dado ya que el set de datos incluye valores correspondiente al período nocturno, donde la humedad como también la velocidad del viento pueden variar en su espectro y no tendrá relevancia alguna con el valor de la irradiancia.

3.2.2. Depuración de los datos

La depuración de los datos permite la detección y posterior corrección o eliminación de datos incorrectos, contaminados o faltantes, como también ayuda a la eliminación de atributos redundantes. Esta etapa selectiva del set de datos posibilita una mayor agilidad en el modelamiento de pronóstico, disminuyendo el costo computacional. El primer paso es seleccionar el período de datos que se usa para la modelación y evaluación de los pronósticos.

Como se mencionó previamente, el set de datos inicial tiene mediciones desde agosto de 2012 en adelante, lo cual es una gran cantidad de datos de distintos años, que dadas las características de los patrones de irradiancia, no son del todo necesarios, pues si se consideran los datos obtenidos para un solo año natural, esos datos ya toman en cuenta el factor de variabilidad que pueden causar las estaciones del año. Por lo tanto, el período de tiempo analizado en este trabajo de título cubre desde el 14 de enero de 2021 hasta el 3 de enero de 2022, lo cual corresponde a un total de 354 días.

Lo siguiente en esta etapa de limpieza de los datos, para lo cual se aplican tres métodos simples de selección de parámetros:

- **Razón de valor faltante:** Esta razón busca descartar columnas de datos que presenten una gran cantidad de valores faltantes (NaN), se calcula como:

$$Razón\ de\ valor\ faltante = \frac{N°\ valores\ faltantes}{Total\ de\ datos} \tag{3.1}$$

  Se considera como límite de corte el 10 % de valores faltantes para la eliminación de una columna. Durante el período analizado ninguna columna fue eliminada bajo este criterio.

  Cabe mencionar que para la aplicación de las técnicas de aprendizaje automático de interés, es necesario que no exista la presencia de valores faltantes, por esta razón, si bien ninguna columna presenta una proporción de valores faltantes mayor al 10 %, el encontrar un valor faltante en cualquiera de las columnas de atributos implica la eliminación de la fila completa correspondiente a ese instante de tiempo. El resultado de lo anterior desemboca en la eliminación de solamente una fila del set de datos, que dado su gran tamaño, no tiene mayor implicancia.

- **Filtro de alta correlación:** Este filtro tiene como función eliminar columnas que sean redundantes entre sí. Bajo este criterio, respaldando lo visto en las Figuras 3.4 y 3.5, la medición para velocidad del viento a 6 metros de altura tiene una gran correlación con la medición realizada a 12 metros de altura, por lo cual se elimina este último atributo y se deja al set de datos solo con la medición realizada a 6 metros de altura. De manera cuantitativa, esto se mide con el coeficiente de correlación de Pearson. Sea una muestra de tamaño N $\{x_i, yi\}_{i=1}^{N}$, el coeficiente de correlación de Pearson $r_{x,y}$ está dado por la ecuación 3.2.

$$r_{x,y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{3.2}$$

  Donde $\bar{x}, \bar{y}$ representan la media muestral para cada conjunto x e y respectivamente. El coeficiente de Pearson toma valores entre $[-1, 1]$, tal que valores cercanos a 1 indican una correlación positiva y cercanos a -1 una correlación negativa, mientras que valores cercanos a 0 indican una baja correlación (aunque todavía pueden existir correlaciones

no lineales en las variables). Para la aplicación del filtro, se considera con una alta correlación a las variables que presenten un $r_{x,y}$ mayor a 0.9 y menor a -0.9.

- **Filtro de baja varianza:** Este filtro se emplea para las variables que presenten baja varianza, pues si la variable evoluciona muy poco con el paso del tiempo, no es mucha la información que se puede obtener de ella. Esto afecta directamente al atributo de voltaje de la batería del capturador de datos. Esta variable al no ser de carácter ambiental y dependiente netamente del equipo instalado, se consideraba susceptible a ser eliminada por ser poco relevante preliminarmente, lo cual se confirma con la aplicación de este filtro, pues para valores normalizados, una varianza menor a 0.01 comúnmente es el criterio a utilizar para la prescindir de la variable.

Finalmente, el resultado de emplear los métodos de reducción de parámetros se traduce en el listado de atributos restantes para el modelamiento de los pronósticos, observado en la Tabla 3.3.

Tabla 3.3: Atributos resultantes posterior a la reducción de parámetros para ser usados en el modelamiento de pronósticos.

Atributos
Fecha y Hora
Irradiancia Global Horizontal
Irradiancia Normal Directa
Irradiancia Horizontal Difusa
Temperatura Ambiental
Humedad Relativa
Velocidad del Viento a 6 metros
Dirección del Viento

3.3. Creación de conjuntos de entrenamiento y testeo

Para llevar a cabo el proceso de entrenamiento de los algoritmos de interés se debe separar la información en dos subconjuntos que contienen series de tiempo particulares, un set de entrenamiento ($\overrightarrow{X}_{train}$, $\overrightarrow{Y}_{train}$), encargado de que el modelo ajuste los pesos y aprenda correctamente, y otro set de prueba o testeo ($\overrightarrow{X}_{test}$, $\overrightarrow{Y}_{test}$), cuya función es evaluar el desempeño del modelo al ponerlo a prueba con datos que no ha visto previamente, de tal manera de poder concluir si el algoritmo fue capaz de generalizar el problema o no. Dado que se está ante un problema de aprendizaje supervisado, los vectores $\overrightarrow{X}$ contienen los atributos que se usan para predecir y los vectores $\overrightarrow{Y}$ el valor objetivo que se apunta a obtener, en este caso, las irradiancias GHI y DNI. Se pueden definir de la siguiente manera:

$$\overrightarrow{X} = \{\overrightarrow{X}_1, \overrightarrow{X}_2, ..., \overrightarrow{X}_{N-1}, \overrightarrow{X}_N\} \tag{3.3}$$

$$\overrightarrow{Y} = \{\overrightarrow{Y}_1, \overrightarrow{Y}_2, ..., \overrightarrow{Y}_{N-1}, \overrightarrow{Y}_N\} \tag{3.4}$$

Donde los subconjuntos de entrenamiento y testeo están contenidos en estos vectores. El proceso de emparejamiento de atributos y valor objetivo requiere trabajo previo, pues en el set de datos inicial todos los atributos están alineados por su instante de tiempo en el que fueron registrados, pero el valor objetivo o *target* está pensado para ser un valor futuro, en particular, proyectados a 10, 30 y 60 minutos. El procedimiento para generar el vector de *target* $\overrightarrow{Y}$ se observa en la Figura 3.6.

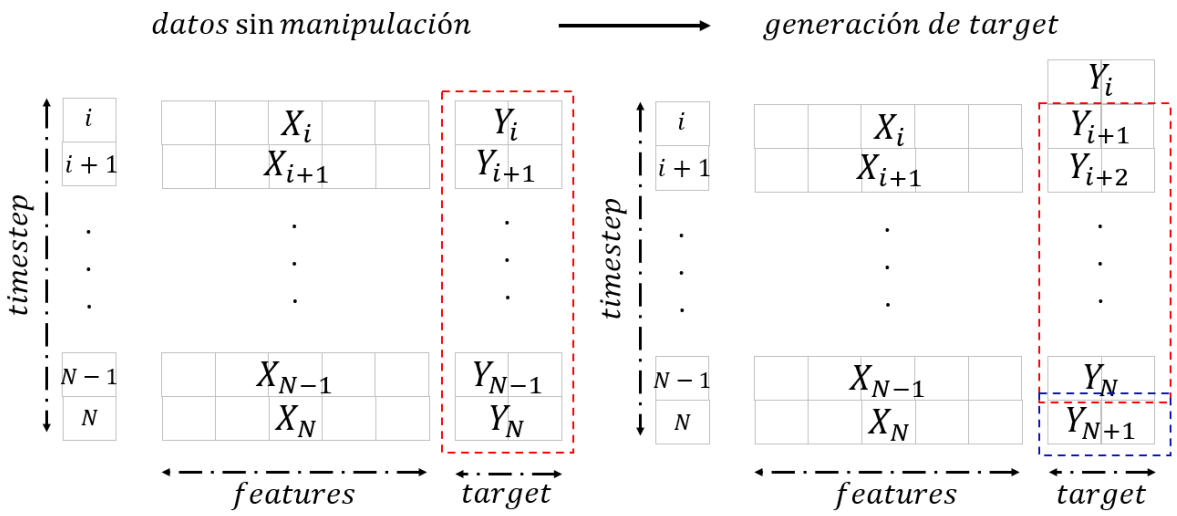

Figura 3.6: Esquema de procedimiento para selección de features y target correspondientes al emparejamiento de un paso de tiempo, alineando los atributos de entrada $\overrightarrow{X}_i$ y salida $\overrightarrow{Y}_{i+1}$.

Este procedimiento consiste en asignar al vector $\overrightarrow{Y}$ las irradiancias GHI y DNI, y desplazar el vector una cierta cantidad de *timesteps*, dependiendo el horizonte de pronóstico, es decir, para el caso de horizonte de 10 minutos, el valor objetivo debe ser desplazado un paso de tiempo, pues el registro de datos tiene una ventana de 10 minutos entre ellos. De manera análoga, para el pronóstico a 30 y 60 minutos, se debe desplazar el vector con los *targets* tres y seis pasos de tiempo, respectivamente.

Es importante mencionar que para el proceso de entrenamiento y testeo se usa un subconjunto del total de datos. En particular se toman en cuenta datos que van desde el 14 de enero de 2021 al 30 de noviembre de 2021, esto con el fin de abarcar prácticamente en su totalidad el ciclo anual pero dejando lugar a una parte de los datos de 2021 disponibles para probar los modelos con predicciones de datos desconocidos.

Por último, queda definir la proporción de datos de $(\overrightarrow{X}, \overrightarrow{Y})$ que serán asignados al subconjunto de entrenamiento y al subconjunto de testeo. Para este trabajo se utilizó una proporción $80\,\% - 20\,\%$ para los subconjuntos de entrenamiento y testeo respectivamente, pues es una

de las combinaciones típicas en la aplicación de algoritmos de aprendizaje automático. Se debe tener precaución de que se tenga en consideración una gran cantidad de datos para entrenar el modelo y una proporción suficiente para el subconjunto de testeo respecto al de entrenamiento (20 % o mayor) para evitar que en el proceso de entrenamiento del modelo este sufra sobreajuste (*overfitting*), lo que provoca que el algoritmo no sea capaz de generalizar la solución del problema y simplemente se aprende de memoria los valores objetivos del set de entrenamiento, tal que su desempeño en este proceso es en extremo bueno, pero al realizar el testeo con datos que no ha visto previamente, no es capaz de realizar predicciones coherentes.

3.4. Configuración de los modelos kNN y GB

La configuración de los modelos es un proceso que consiste principalmente en determinar los hiperparámetros, los cuales corresponden a parámetros que no se aprenden directamente en los modelos y deben ser especificados previamente. El procedimiento para encontrar los hiperparámetros óptimos comprende entregar un rango para cada hiperparámetro de interés y un método de búsqueda aleatoria determina la combinación que obtuvo un mejor desempeño dentro de un conjunto de validación, que está constituido por un subconjunto del set de entrenamiento.

3.4.1. Modelo kNN

Para el modelo de vecinos más cercanos, los hiperparámetros a determinar son los siguientes:

- **Número de vecinos:** Cantidad de valores que se verifican como similares al valor a predecir, en este caso se considera un vector con los valores enteros $\{1, 2, ..., 50\}$.

- **Función peso:** Función para ponderar el peso de cada vecino en el cálculo del valor a predecir. El módulo que se usa contiene dos opciones: $\{uniform, distance\}$. La función uniforme pondera de igual manera a todos los vecinos seleccionados mientras que la función basada en distancia pondera usando la inversa de la distancia (semejanza a cada atributo) respecto al valor de interés, es decir, de los vecinos seleccionados, los que estén más cercanos tendrán mayor influencia que los que se encuentren más alejados.

Cabe mencionar que existen otros hiperparámetros que pueden ser seleccionados en el modelo, como el algoritmo a usar para determinar cuales son los vecinos más cercanos o la métrica usada (siendo la distancia euclidiana la métrica por defecto), pero los anteriores son los más relevantes, de tal manera que el resto de los hiperparámetros se deja en sus valores predeterminados.

Posteriormente se evalúa cada una de las posibles combinaciones de los hiperparámetros mencionados. El resultado obtenido como óptimo es el que presenta menor error para cada configuración, el cual se puede observar en la Tabla 3.4.

Tabla 3.4: Hiperparámetros seleccionados para cada configuración de pronóstico del modelo kNN.

Horizonte de tiempo	$N°$ de vecinos GHI	$N°$ de vecinos DNI	Función peso
10	14	15	*uniform*
30	9	8	*uniform*
60	6	7	*uniform*

3.4.2. Modelo Gradient Boosting

De manera similar, el método de Gradient Boosting también posee hiperparámetros a configurar, los más relevantes son:

- **Función de pérdida:** Corresponde a la función de pérdida a minimizar en el modelo, dentro de las cuales se encuentran: {*squared error*, *absolute error*, *huber*, *quantile*}. En particular la función por defecto corresponde al *squared error* o como se describió en el capítulo de antecedentes sobre GB, la suma de los residuos al cuadrado mostrada en la ecuación 2.8.

- **Tasa de aprendizaje:** Pondera la contribución individual de cada árbol de decisión en la predicción. El valor por defecto es $\nu = 0.1$. Como se aclaró previamente, un valor pequeño de ν reduce el sesgo y mejora la precisión del modelo, pero aumenta el costo computacional requerido. Se usa el valor por defecto.

- **Número de estimadores:** Corresponde a la cantidad de árboles que se crearán en el modelo. El valor por defecto corresponde a $M = 100$, el cual fue el seleccionado para este hiperparámetro.

Para este modelo en particular, a diferencia del procedimiento realizado para kNN, se decide dejar como definitivos los valores predeterminados por razones de simplicidad. De tal manera que los hiperparámetros a utilizar se presentan en la Tabla 3.5.

Tabla 3.5: Hiperparámetros seleccionados para cada configuración de pronóstico del modelo GB.

Horizonte de tiempo	Función de pérdida	ν	M
10	*squared error*	0.1	100
30	*squared error*	0.1	100
60	*squared error*	0.1	100

3.5. Configuración de las arquitecturas de redes neuronales

La configuración de los modelos relacionados a redes neuronales, en este caso FFNN y LSTM, sigue un procedimiento parecido a la configuración de los modelos kNN y GB, pero se diferencia en que el foco principal está en el arreglo y cantidad de capas y neuronas, junto a las funciones de activación de cada capa.

3.5.1. Arquitectura FFNN

La estructuración de la arquitectura de esta red tuvo como filosofía el mantener una cantidad similar de parámetros entrenables que su contraparte LSTM, esto con la finalidad de que la comparación sea lo más justa posible y que un mejor resultado en alguno de los modelos no se deba netamente a este factor.

Adicionalmente al número de capas y neuronas, como se mencionó previamente, se debe seleccionar una función de activación correspondiente a cada capa. Entre las funciones de activación disponibles más tradicionales se encuentran la función tangente hiperbólica $f(x) = (exp(x) - exp(-x))/(exp(x) + exp(-x))$ y la función sigmoide $f(x) = 1/(1 + exp(-x))$. Sin embargo, en el último tiempo una de las funciones de activación más populares ha sido la unidad lineal rectificada (ReLU), la cual se expresa como $f(x) = max(0, x)$, es decir, el nodo se activa solo si el argumento está por sobre cierta cantidad. Si bien en decadas pasadas las funciones tangente hiperbólica y sigmoide fueron las más predominantes, la función ReLU ha demostrado, comúnmente, que tiene capacidad de un aprendizaje mucho más rápido en redes que presentan una cantidad elevada de capas [12]. Una representación gráfica de la función ReLU se muestra en la Figura 3.7.

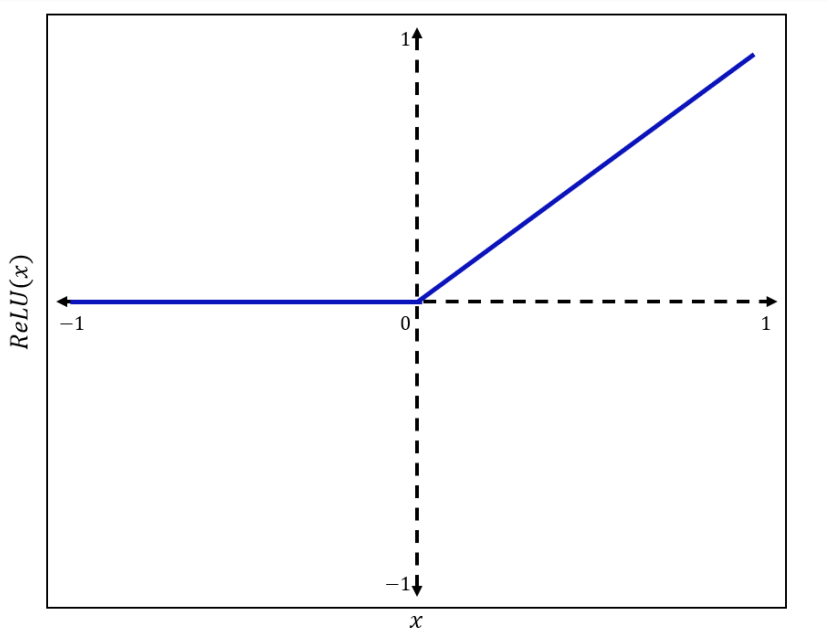

Figura 3.7: Muestra gráfica de la función de activación ReLU.

Dado lo anterior, la función ReLU fue seleccionada como función de activación de cada capa *Dense* (abreviación de *Densely Connected Neural Network*, que corresponde a lo que es una capa FFNN tradicional) exceptuando por la capa de salida, donde la función de activación en este caso corresponde a una función lineal, pues en un problema de regresión, se requiere que la salida sea un valor único.

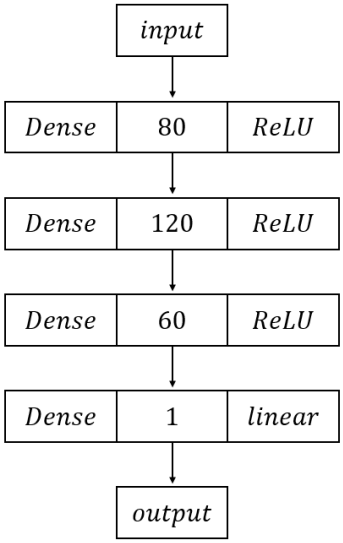

Figura 3.8: Esquema de la arquitectura utilizada para el modelo FFNN.

En la Figura 3.8 se observa la arquitectura resultante para el modelo FFNN, compuesta de tres capas *Dense* con 80, 120 y 60 neuronas cada una y finalmente la capa de salida que contiene solo una neurona, la cual da lugar al valor resultante de la predicción. La idea de usar tres capas es mantener el modelo dentro de la clasificación de machine learning y no pasar a ser una red de deep learning, donde el criterio tradicional es tener más de tres capas de neuronas. Cabe mencionar que el número de parámetros entrenables dada esta arquitectura corresponde a 17,841.

3.5.2. Arquitectura LSTM

Las redes LSTM, al ser un tipo de redes recurrentes, cuentan con una particularidad al momento de recibir los datos de entrada y es que dada su estructura, trabaja con un *input* tridimensional, compuesto por el tamaño de cada muestra, el número de *timesteps* que el modelo tendrá en cuenta al mirar hacia el pasado y la dimensión de los *features* de entrada. Esto implica que los datos deben ser preprocesados para reorganizar su dimensionamiento y ser ordenados en el formato deseado. Esta estructura se observa en la Figura 3.9.

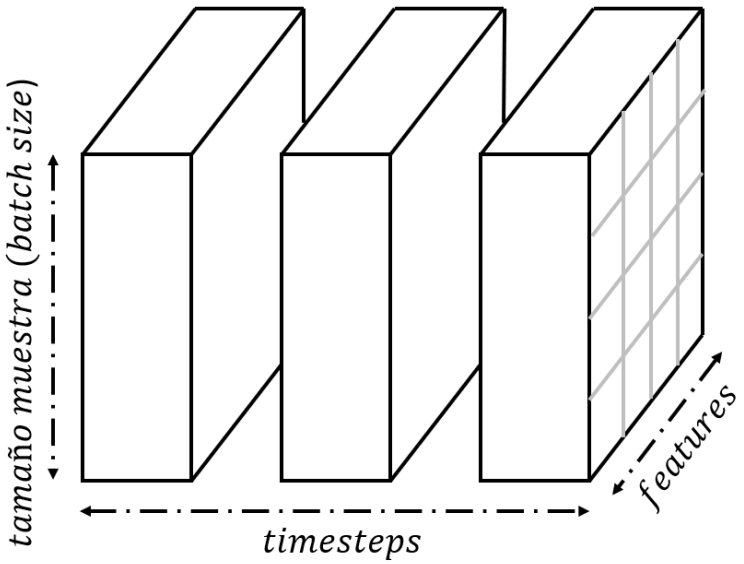

Figura 3.9: Esquema de predisposición de los datos como input para el modelo LSTM.

De los tres componentes del *input* tridimensional, el número de *timesteps* n anteriores que el modelo vuelve a incorporar como valor de entrada en cada etapa, es un factor a decidir. En este caso en concreto se usa $n = 3$, es decir, en cada celda LSTM, el modelo revisa adicionalmente la información de tres espacios de tiempo en el pasado, que dado por la distribución de los datos, corresponde a los 30 minutos previos para cada predicción.

Finalmente se procede a la construcción de la arquitectura al igual que en la red FFNN, esta se observa en la Figura 3.10. La red está compuesta de una capa LSTM con 50 neuronas y función de activación ReLU, seguida de dos capas *Dense* tradicionales, de 60 neuronas cada una y por último la capa de salida con su nodo único para entregar el valor resultante a la predicción. Es importante notar que se usa nuevamente la función de activación ReLU para las capas *Dense*, en base al argumento planteado previamente en la sección de la construcción de la arquitectura FFNN. Por otro lado, la cantidad de parámetros entrenables en este modelo escalan a 18,781, lo cual está en un orden bastante similar a su contraparte FFNN. Esto se realiza a propósito, pues como se mencionó previamente, permite que la comparación entre estos dos modelos sea más justa al momento de analizar los resultados.

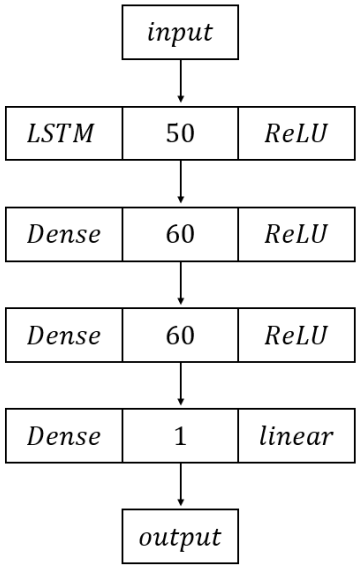

Figura 3.10: Esquema de la arquitectura utilizada para el modelo LSTM.

3.6. Entrenamiento y testeo

El proceso de entrenamiento y testeo es aplicado a los cuatro modelos de interés utilizando los subconjuntos $(\overrightarrow{X}_{train},\ \overrightarrow{Y}_{train})$ y $(\overrightarrow{X}_{test},\ \overrightarrow{Y}_{test})$ definidos previamente. Para los modelos de redes neuronales adicionalmente se define la función de pérdida a optimizar que se toma como criterio, la cual es el Mean Squared Error (MSE), y se minimiza en base al método *adam*. Finalmente se define el tamaño de la muestra (*batchsize*) a utilizar y la cantidad de épocas a entrenar. El *batchsize* es 100 y la cantidad de épocas seleccionadas corresponde a 250, número que no fue variado durante el proceso. El principio de funcionamiento del *batchsize*, es que si la muestra de entrenamiento tiene 1000 elementos y el *batchsize* utilizado es de 100, le tomará 10 iteraciones al algoritmo completar una época.

Es importante agregar que un menor número de tamaño de muestra reduce la memoria utilizada del equipo y normalmente entrena más rápido, sin embargo, de ser muy pequeña, puede verse afectada la precisión en la estimación.

3.7. Métricas de evaluación

Para cuantificar el desempeño de los modelos utilizados, se hizo uso de métricas estadísticas, las cuales permiten cuantificar el error correspondiente a la predicción, contenido en el vector $\widehat{Y}$, respecto a los valores observados reales en el vector $\overrightarrow{Y}$.

La primera métrica corresponde al Mean Absolute Error (MAE), la cual entrega la desviación promedio de los valores del pronóstico respecto a los observados.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\widehat{Y} - \overrightarrow{Y}\right| \tag{3.5}$$

De la misma manera que el MAE, el Root Mean Squared Error (RMSE) consiste en comparar los valores reales respecto a los obtenidos al realizar la predicción. Sin embargo, esta métrica tiene la particularidad de elevar al cuadrado el valor absoluto de la diferencia entre $\widehat{Y}$ e $\overrightarrow{Y}$, lo que provoca que los valores atípicos (*outliers*) ponderen más en el resultado del RMSE [13].

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\widehat{Y} - \overrightarrow{Y})^2} \tag{3.6}$$

El Mean Bias Error (MBE) sigue una fórmula similar al MAE, pero sin considerar el valor absoluto. Esta es la razón de la utilidad de esta métrica, pues representa la tendencia del modelo a subestimar o sobrestimar los valores observados en base al promedio simple del pronóstico.

$$MBE = \frac{1}{N}\sum_{i=1}^{N}(\widehat{Y} - \overrightarrow{Y}) \tag{3.7}$$

El Forecasting Skill (FS) muestra de manera porcentual la disminución o aumento de error de un modelo respecto a un modelo de referencia, esto permite tener una comparación directa entre la capacidad de un modelo y otro. En particular, dado que es el método de mayor simpleza, se usa como referencia el modelo kNN comparando la métrica RMSE.

$$FS = 1 - \frac{RMSE}{RMSE_{ref}} \tag{3.8}$$

Finalmente, la Skewness es una medida de la distorsión de simetría (asimetría) en un set de datos. Respecto a los pronósticos, cuantifica el sesgo que hay en los resultados, de manera similar al MBE, pero con otro criterio, pues tiene en consideración la mediana de los datos y su desviación estándar (SD), además del promedio simple.

$$Skewness = \frac{MBE - mediana(\widehat{Y} - \overrightarrow{Y})}{SD} \tag{3.9}$$

Capítulo 4

Resultados y discusión

En el presente capítulo se exponen y discuten los resultados entregados por los cuatro modelos propuestos: kNN, GB, FFNN y LSTM. Cada uno de ellos se evalúa para tres horizontes de tiempo intra-horarios a pronosticar: 10, 30 y 60 minutos con el objetivo de analizar el rendimiento de los modelos y su evolución a medida que se incrementa el horizonte de pronóstico. Los resultados expuestos por las métricas corresponden a la evaluación de cada uno de los modelos al ser entrenados y testeados con los mismos subconjuntos de entrenamiento y prueba, para de esta manera mantener igualdad de condiciones.

En conjunto a lo anterior, también se dispone de predicciones realizadas con datos de diciembre de 2021, los cuales fueron dejados fuera del proceso de entrenamiento con este propósito, además de sus respectivos gráficos de dispersión de los valores predichos comparados con los valores reales de irradiancia, para así tener una mayor claridad del comportamiento de cada uno de los modelos utilizados.

4.1. Análisis de modelos kNN y GB

En función de establecer una comparación más equitativa y sustentar una organización que permita visualizar de mejor manera los resultados, los modelos se analizan en dos subcategorías previo a la comparación global, la primera de estas subcategorías corresponde a los métodos estadísticos: kNN y GB, y la otra a las técnicas asociadas a redes neuronales: FFNN y LSTM. Adicionalmente se desglosa el análisis en estudio de rendimiento para GHI y DNI.

4.1.1. Rendimiento de pronóstico GHI

De acuerdo a las iteraciones realizadas para el pronóstico de GHI, los mejores resultados obtenidos por los modelos kNN y GB se muestran en las Tablas 4.1 y 4.2 respectivamente. Notar que en los resultados para kNN, el FS es 0, pues este es el método usado como referencia para los otros tres.

Tabla 4.1: Métricas estadísticas obtenidas en la evaluación de desempeño de modelo kNN para pronóstico de GHI.

Horizonte	MAE	RMSE	MBE	Skewness	FS
10 minutos	0.0073	0.0219	-0.0012	-1.3230	0.00 %
30 minutos	0.0143	0.0329	-0.0009	-0.1418	0.00 %
60 minutos	0.0252	0.0520	-0.0011	0.3509	0.00 %

Tabla 4.2: Métricas estadísticas obtenidas en la evaluación de desempeño de modelo GB para pronóstico de GHI.

Horizonte	MAE	RMSE	MBE	Skewness	FS
10 minutos	0.0071	0.0216	-0.0009	-1.5528	1.38 %
30 minutos	0.0131	0.0298	-0.0035	-1.6153	9.53 %
60 minutos	0.0202	0.0370	-0.0040	-0.1875	28.83 %

Es importante notar que las métricas son aplicadas sobre valores normalizados, por ende, sus resultados también lo están. Lo primero que se puede observar en ambos modelos es el aumento paulatino en el error MAE y RMSE a medida que aumenta el horizonte de pronóstico, lo cual es un resultado esperado. Sin embargo, haciendo un análisis de la comparación directa de los dos modelos, acorde al FS, se puede apreciar que el error crece notablemente más rápido en el modelo kNN al aumentar la ventana de tiempo a predecir que en el modelo GB. En particular para el pronóstico con horizonte de 10 minutos, tanto kNN como GB se comportan de manera similar, con un desempeño ligeramente superior de un 1.98 % por parte de GB, lo cual se acrecienta de forma drástica al aumentar el horizonte de pronóstico a 30 y 60 minutos, tal que el FS muestra un desempeño 9.58 % superior en el primer caso y un 28.83 % en el segundo.

Esta diferencia en el desempeño tiene como factor principal que en el método kNN la medición de distancias se vuelve menos discriminatoria a medida que aumenta la dimensionalidad del problema, en este caso 8 *features*. A esto se le agrega otra dificultad inherente del mismo modelo, que es que requiere un proceso de escalamiento global sobre todos los atributos en función de que la ponderación de estos sea más justa. Sin embargo, dadas características propias de algunos de los atributos, hay ocasiones en donde no es posible, pues sus escalas individuales difieren mucho, como es el caso de la dirección del viento, medida en grados sexagesimales, en contraste con otras variables, tales como la temperatura, que se registra en grados celsius. Por esta razón, el problema planteado en este trabajo de título, es de carácter complejo para este modelo, y si bien su desempeño no es malo para el pronóstico a 10 minutos, al aumentar el horizonte de pronóstico, lo que le agrega dificultad a la predicción, el modelo kNN deja ver sus deficiencias.

Adicionalmente es importante tener en cuenta la susceptibilidad del modelo a los *outliers*, lo cual se aprecia al comparar los resultados individuales del MAE, donde la mejora en GB va desde 2.97 % en el caso más favorable para kNN a un 20.08 % en el menos favorable para este último, con los resultados individuales del RMSE, que muestra directamente el FS, donde

la mejora del GB aumenta hasta casi un 30 % en el escenario de pronóstico a 60 minutos. Esto es por lo mencionado previamente respecto a la consideración extra que tiene la métrica RMSE sobre los valores atípicos en comparación al MAE, y es que a pesar de que los árboles de decisión también son muy susceptibles a los valores atípicos, esto se aborda con el *boosting*, es decir, el ensamble de estos modelos de predicción débiles (árboles de decisión), corrigiendo el error y disminuyendo los *outliers* gradualmente con cada iteración, haciendo que el modelo GB rinda mejor en este aspecto.

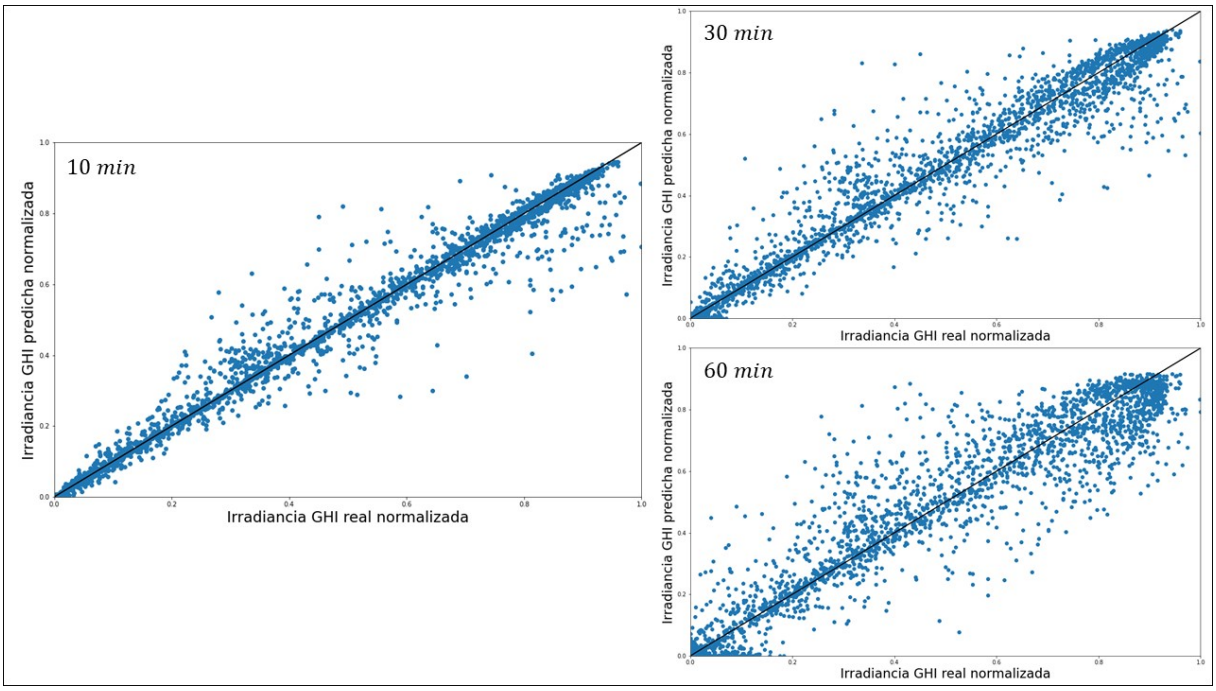

Figura 4.1: Dispersión de pronósticos GHI de modelo kNN respecto a los valores reales medidos en el mes de diciembre de 2021 para tres horizontes de pronóstico.

En la Figura 4.1 se observa la dispersión en el pronóstico, comparando los datos reales con los entregados por la predicción para los tres horizontes de tiempo, del modelo kNN. Esto permite complementar otro factor importante a analizar, como lo es el sesgo en los modelos, el cual está representado por las métricas MBE y *Skewness*. Se aprecia de manera preliminar al observar los valores entregados por estas métricas que ambos modelos poseen una tendencia a subestimar la irradiancia, pues son valores negativos, sin embargo, en una medida muy pequeña tanto para GB como para kNN. Notamos que en este último, la subestimación de GHI es muy estable, es decir, no presenta una gran variación, sin importar el horizonte de pronóstico al analizar la métrica MBE, cosa que no ocurre con el modelo GB, donde la subestimación incrementa junto con el aumento de horizonte de pronóstico, lo que observando la Figura 4.2, se nota que es producto de los valores atípicos del modelo, es decir, los valores atípicos que genera GB son en general por debajo de la irradiancia real.

Ambos gráficos de dispersión permiten entender de mejor manera el comportamiento del sesgo en los modelos, donde para kNN la estabilidad observada en las métricas para este aspecto viene dado por la manera de operar del modelo. Se observa que para irradiancias

bajas el modelo tiende a sobreestimar los resultados mientras que para irradiancias altas el modelo tiende a subestimarlos, lo cual destaca con mayor claridad en los gráficos de los pronósticos a 30 y 60 minutos. Esto provoca que el sesgo se compense, entregando los valores para MBE previamente discutidos. El sesgo diferenciado del modelo kNN se puede explicar por la cantidad de vecinos seleccionados: al buscar vecinos de un valor bajo, es más probable que incluya algunos valores mayores, aumentando así el promedio. Análogamente, lo mismo puede ocurrir para valores altos de irradiancia.

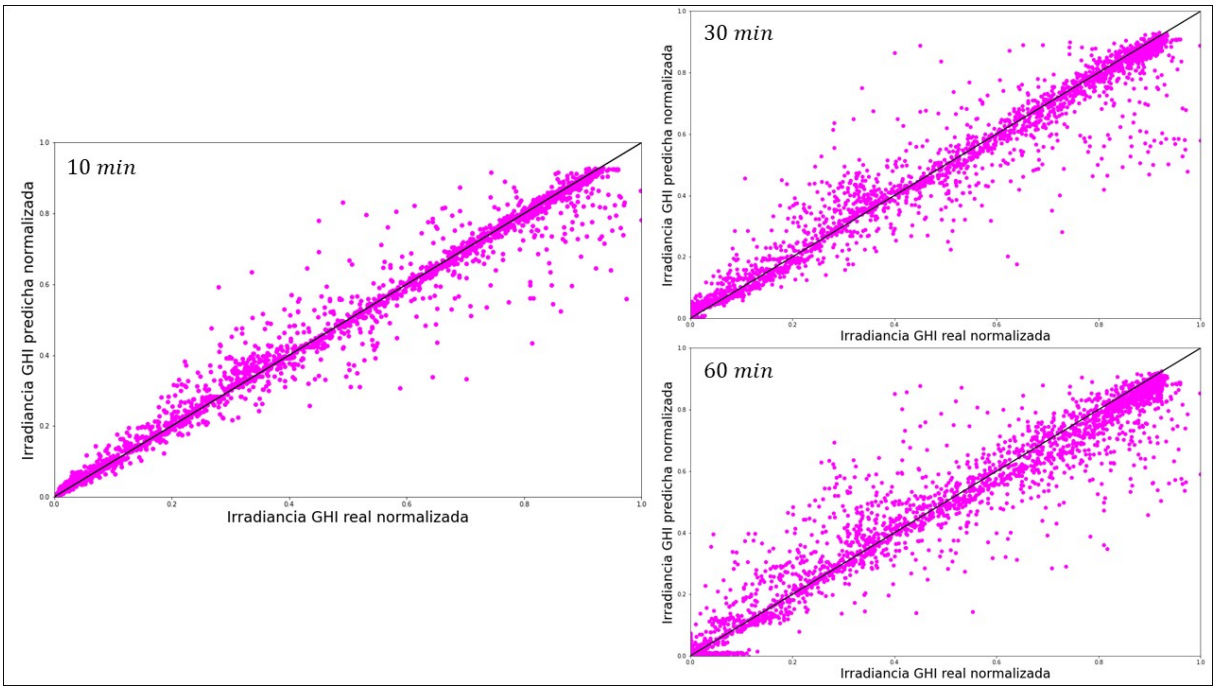

Figura 4.2: Dispersión de pronósticos GHI de modelo GB respecto a los valores reales medidos en el mes de diciembre de 2021 para tres horizontes de pronóstico.

Respecto a la dispersión en el modelo GB, esta es más pareja y también cuenta con la característica de que es más difícil reconocer si hay algún patrón recurrente en el sesgo. Por otro lado, lo que sí es evidente, es el mejor rendimiento que este modelo presenta para las predicciones con horizontes de 30 y 60 minutos de la Figura 4.2 en relación a sus correspondientes del modelo kNN en la Figura 4.1, pues se logra observar una concentración mucho mayor de cúmulos de puntos cercanos a la línea diagonal de los gráficos.

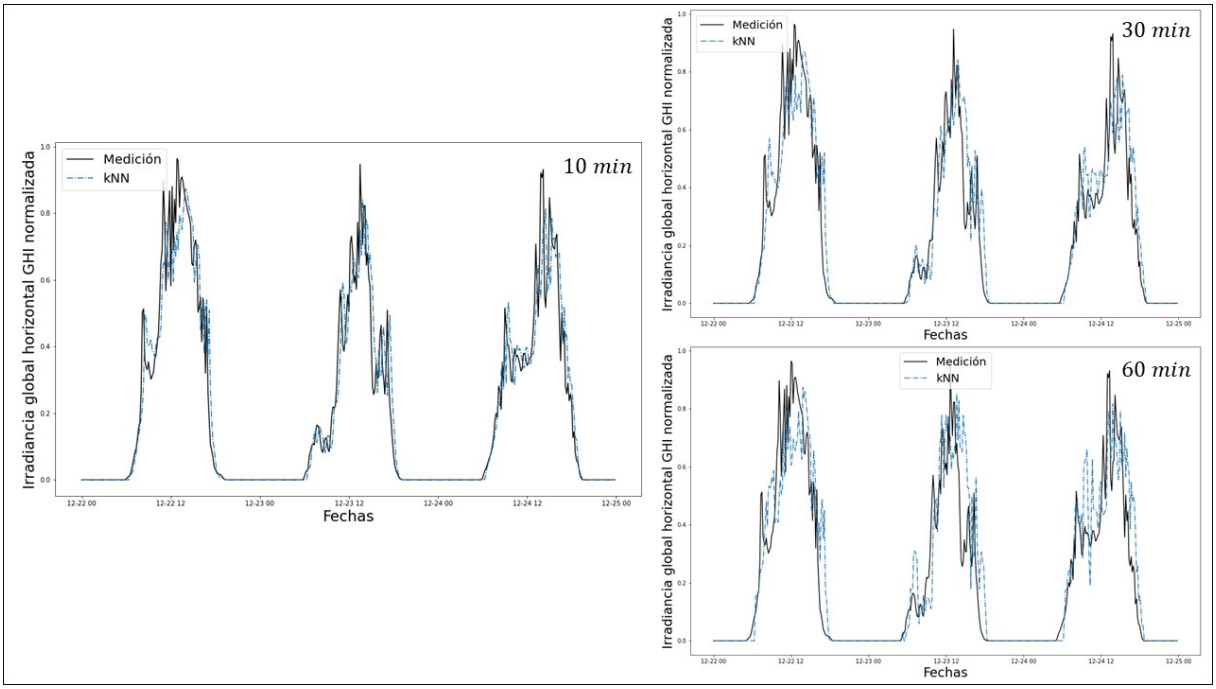

Figura 4.3: Comparación de pronósticos de irradiancia GHI de modelo kNN en días parcialmente nublados para tres horizontes de pronóstico.

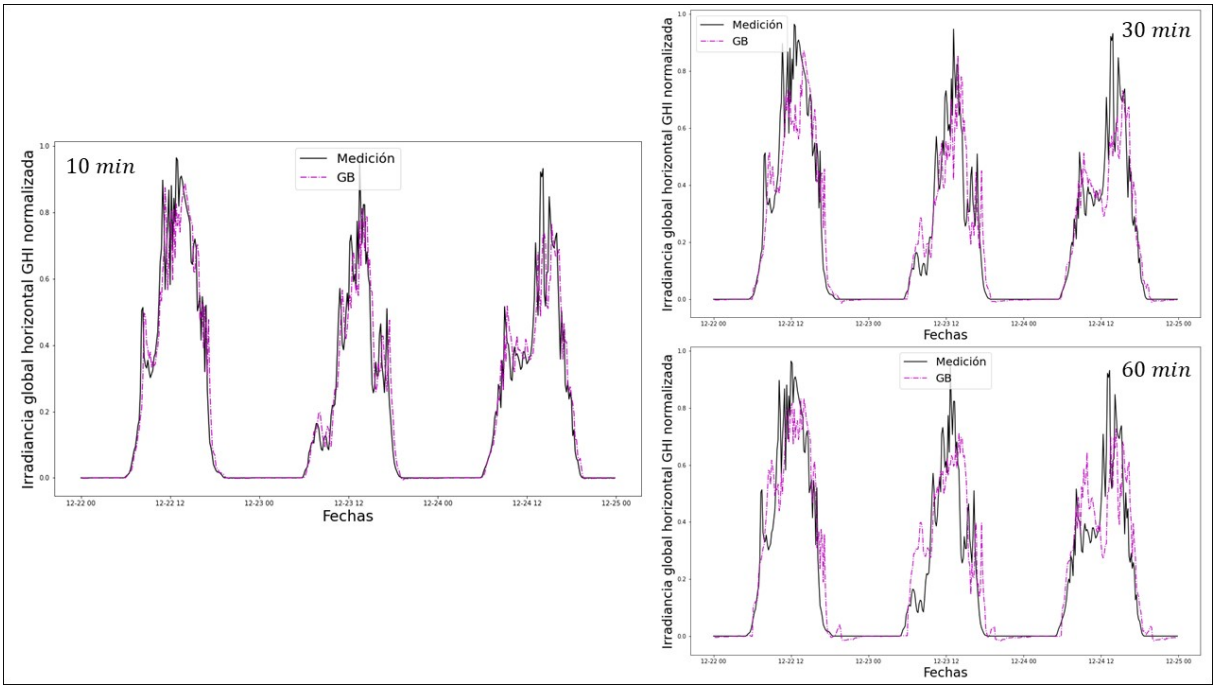

Figura 4.4: Comparación de pronósticos de irradiancia GHI de modelo GB en días parcialmente nublados para tres horizontes de pronóstico.

Finalmente, las Figuras 4.3 y 4.4 muestran la predicción de tres días parcialmente nublados del mes de diciembre para cada modelo. En la Figura 4.3, que corresponde al modelo kNN, lo que más resalta es el patrón previamente mencionado para el sesgo de manera más explícita, donde es claro ver que los picos presentes para irradiancias bajas son en su ma-

yoría sobreestimados, mientras que los picos absolutos (puntos máximos de irradiancia) son subestimados. Esto no es tan evidente para GB en la Figura 4.4, exceptuando por el pronóstico a 60 minutos, donde sí se puede observar este patrón, aunque de manera más errática.

Es importante considerar que los tres días mostrados corresponden a la situación menos favorable a predecir para todos los modelos, y que los resultados expuestos en las Tablas 4.1 y 4.2 corresponde a todo el set prueba, lo que incluye tanto días soleados como días con baja y alta nubosidad, de tal manera que se entrega un resultado más generalizado.

4.1.2. Rendimiento de pronóstico DNI

Si bien la irradiancia GHI es la componente más importante en el estudio de pronósticos solares, la DNI sigue siendo de interés, sobretodo para tecnologías de CSP, como para paneles PV con función de seguimiento solar. Así, un análisis similar al de GHI se realiza para la componente normal directa de la irradiancia, donde los mejores resultados obtenidos dadas las iteraciones a las que se sometieron los modelos kNN y GB, se muestran en las Tablas 4.3 y 4.4 respectivamente.

Tabla 4.3: Métricas estadísticas obtenidas en la evaluación de desempeño de modelo kNN para pronóstico de DNI.

Horizonte	MAE	RMSE	MBE	Skewness	FS
10 minutos	0.0137	0.0439	0.0057	0.3392	0.00 %
30 minutos	0.0252	0.0689	0.0100	0.7014	0.00 %
60 minutos	0.0426	0.1001	0.0136	0.9557	0.00 %

Tabla 4.4: Métricas estadísticas obtenidas en la evaluación de desempeño de modelo GB para pronóstico de DNI.

Horizonte	MAE	RMSE	MBE	Skewness	FS
10 minutos	0.0151	0.0454	0.0019	-0.5019	-3.58 %
30 minutos	0.0250	0.0659	0.0089	0.3198	4.35 %
60 minutos	0.0421	0.0911	0.0129	0.8815	8.99 %

Lo primero a notar es que el error MAE como el RMSE aumenta considerablemente para ambos modelos respecto a la predicción realizada para GHI, cercano a un incremento del 200 %, que bien en el global de la predicción, los valores de MAE y RMSE no son altos, sigue siendo una baja en el rendimiento importante en comparación al pronóstico de GHI. Esto no es sorpresivo, pues los datos registrados para DNI por lo general presentan un comportamiento más errático que el de su componente GHI, dado que a diferencia de esta, el DNI se mide usando un pirheliómetro, en lugar de un piranómetro, este capta irradiancia en una sola dirección (que es lo que se necesita para medir DNI) y se encuentra fijo a la estación solarimétrica. Luego con ayuda del seguidor solar de la instalación, tiene la capacidad de apuntar

en la dirección correcta, pero esto trae consigo errores sistemáticos, como la dependencia al movimiento del seguidor solar, donde una mala calibración puede dar como resultado datos no confiables, además del tamaño del lente del pirheliómetro como tal, ya que al ser pequeño, no permite un gran margen de error en la calibración y por tanto no siempre puede captar necesariamente toda la irradiancia directa.

Respecto a la comparativa de ambos modelos según las métricas como tal, los resultados no difieren significativamente entre kNN y GB, inclusive, en contraste a lo ocurrido para el pronóstico de GHI, el modelo KNN vence en rendimiento para el horizonte de 10 minutos por un 3.58 % según la métrica FS, y si bien GB revierte esta situación en los otros dos horizontes de pronóstico, solo llega a obtener un 8.99 % de mejor puntaje FS en su escenario más favorable en comparación al 28.83 % obtenido bajo este mismo escenario en el pronóstico de GHI.

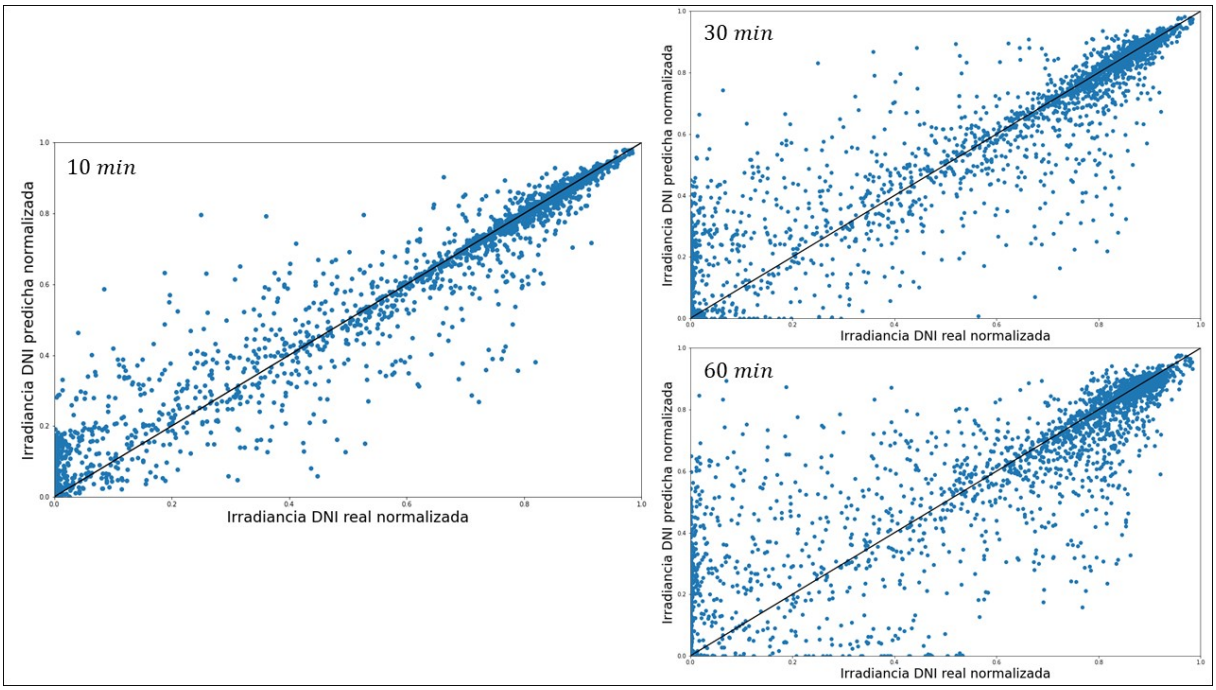

Figura 4.5: Dispersión de pronósticos DNI de modelo kNN respecto a los valores reales medidos en el mes de diciembre de 2021 para tres horizontes de pronóstico.

La Figura 4.5 muestra la dispersión del pronóstico de DNI para el modelo kNN. Lo más destacable al comparar con el gráfico de dispersión de GHI para este mismo modelo mostrado en la Figura 4.1, es que en este caso, para los tres horizontes de pronóstico, el modelo presenta claras dificultades para la predicción de irradiancias intermedias, donde hay una gran dispersión de puntos, mientras que para irradiancias altas logra un mayor desempeño, que es donde principalmente se forma el cúmulo de puntos. Esto es en gran medida consecuencia también del comportamiento errático de la variable DNI, comportamiento que queda plasmado de mejor manera en la Figura 4.7, que muestra la predicción para los mismos tres días de diciembre vistos para el caso de GHI, pero que a diferencia de este, para DNI se pueden ver una mayor cantidad de saltos abruptos en los datos reales de la irradiancia. Esto provoca

que sea más difícil para el modelo distinguir patrones. En cuanto a las irradiancias altas, si bien el comportamiento de DNI para estos días presenta muchos picos, el modelo tiene como ayuda la correlación horaria, pues el máximo de irradiancia siempre es alcanzado entre las 12:00 y las 14:00 horas, además de que este máximo es relativamente recurrente para casi todos los días dentro de una misma estación del año.

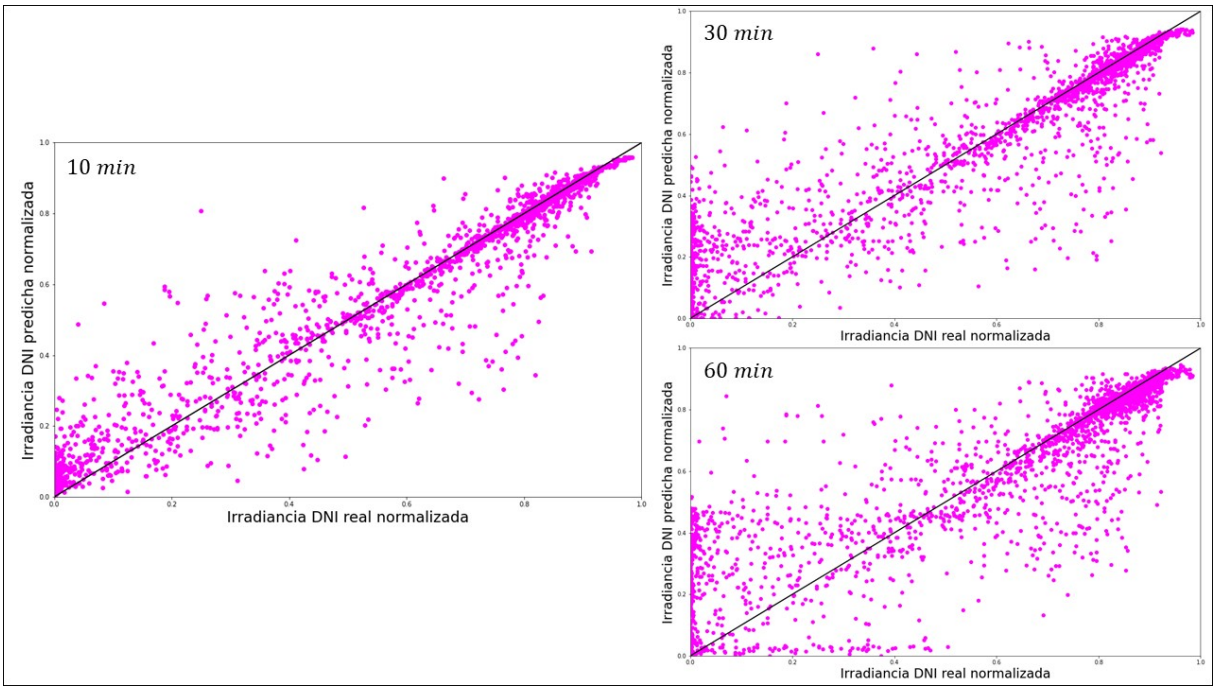

Figura 4.6: Dispersión de pronósticos DNI de modelo GB respecto a los valores reales medidos en el mes de diciembre de 2021 para tres horizontes de pronóstico.

De la misma manera, en la Figura 4.6 se observa la dispersión de pronóstico DNI para el modelo GB, donde es posible apreciar el mismo comportamiento que el visto en el pronóstico realizado por el modelo kNN, tal que hay una mayor dispersión de puntos para valores intermedios de DNI, mientras que en irradiancias altas destaca el cúmulo de puntos cercano a la línea, indicando un mejor rendimiento para las predicciones de estos valores. La razón de esto es exactamente ls misma que para el modelo kNN, es decir, a causa de la naturaleza errática de los valores registrados para kNN.

Cabe mencionar que si bien las métricas relacionadas al sesgo, MBE y *Skewness*, principalmente la primera, podrían indicar una leve tendencia a sobreestimar los valores para practicamente todos los escenarios, dados los valores positivos de MBE, estos siguen estando muy cercanos a cero. Adicionalmente, esto tampoco es un patrón que se observe en los gráficos de dispersión de las Figuras 4.5 y 4.6, sino que más bien se aprecia un comportamiento relativamente compensado entre subestimación y sobreestimación.

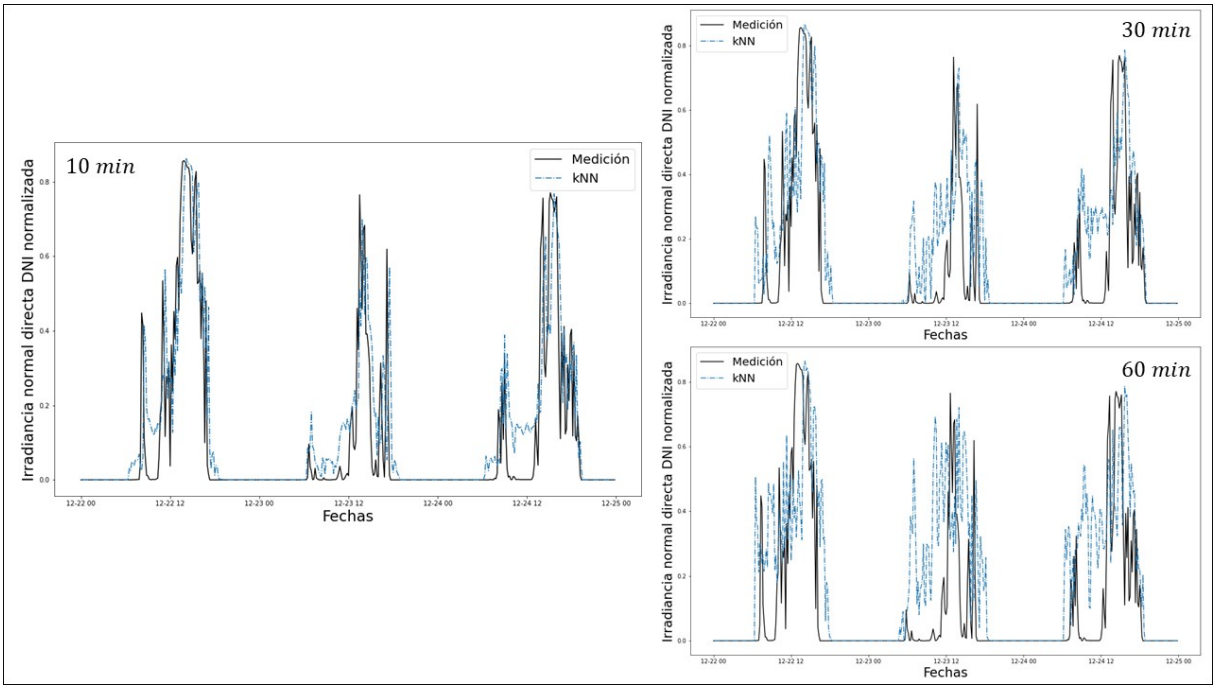

Figura 4.7: Comparación de pronósticos de irradiancia DNI de modelo kNN en días parcialmente nublados para tres horizontes de pronóstico.

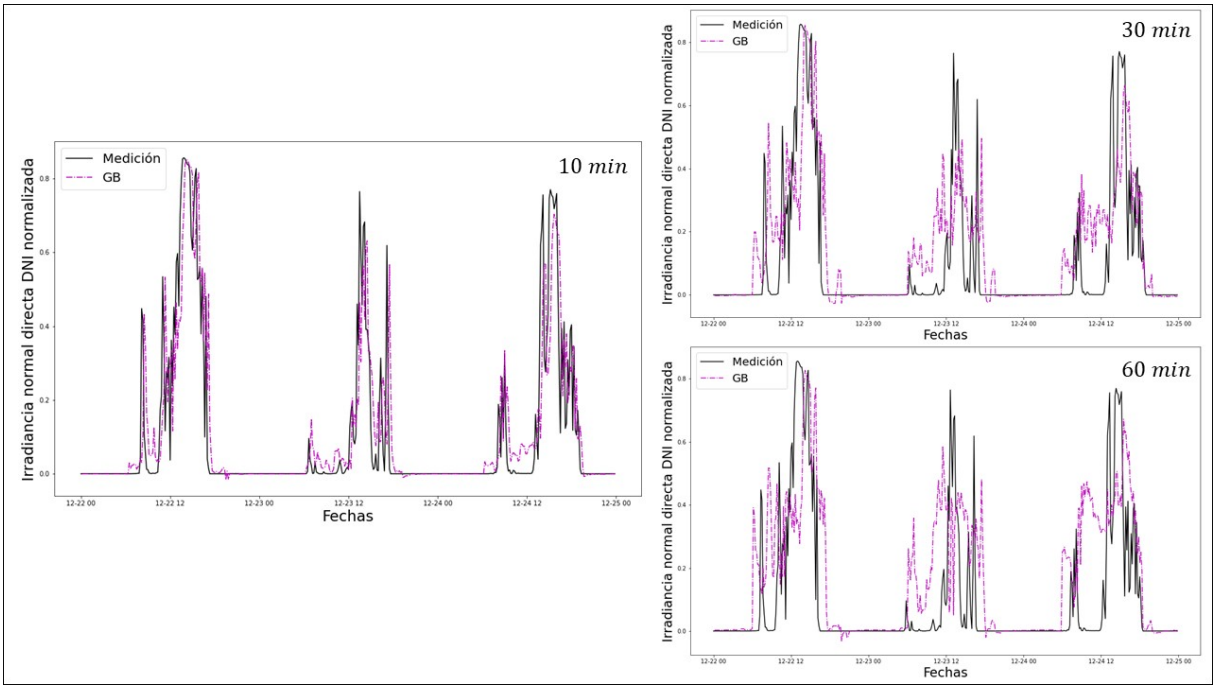

Figura 4.8: Comparación de pronósticos de irradiancia DNI de modelo GB en días parcialmente nublados para tres horizontes de pronóstico.

Por último, haciendo énfasis en las Figuras 4.7 y 4.8, se observa de manera mayoritaria que los puntos más difíciles de predecir corresponden a los espacios intermedios previos a un cambio abrupto en la irradiancia, lo cual se alinea con lo discutido previamente respecto a los gráficos de dispersión. Este fenómeno es cada vez más marcado a medida que se incrementa

el horizonte de pronóstico, lo cual es coherente con lo esperado.

4.2. Análisis de modelos FFNN y LSTM

Las redes neuronales generalmente son predilectas para abordar problemas que involucren series de tiempo, esto dado que son capaces de aproximar comportamientos no lineales, añadiendo versatilidad a los modelos.

A continuación se presentan los resultados obtenidos para las redes FFNN y LSTM.

4.2.1. Rendimiento de pronóstico GHI

En función de las iteraciones realizadas para el pronóstico de irradiancia GHI, los mejores resultados conseguidos por los modelos FFNN y LSTM se muestran en las Tablas 4.5 y 4.6 respectivamente.

Tabla 4.5: Métricas estadísticas obtenidas en la evaluación de desempeño de modelo FFNN para pronóstico de GHI.

Horizonte	MAE	RMSE	MBE	Skewness	FS
10 minutos	0.0066	0.0200	-0.0019	-1.0245	8.58 %
30 minutos	0.0110	0.0327	0.0085	1.5407	0.65 %
60 minutos	0.0270	0.0409	-0.0127	-0.6179	21.37 %

Tabla 4.6: Métricas estadísticas obtenidas en la evaluación de desempeño de modelo LSTM para pronóstico de GHI.

Horizonte	MAE	RMSE	MBE	Skewness	FS
10 minutos	0.0004	0.0011	-0.0004	-0.2359	94.97 %
30 minutos	0.0012	0.0022	0.0006	0.0872	93.32 %
60 minutos	0.0198	0.0345	0.0195	0.1792	33.67 %

Lo primero a analizar son las métricas de error MAE y RMSE, en conjunto con el FS, el cual vuelve a tomar como referencia el modelo kNN. Si bien el rendimiento de ambos modelos es satisfactorio, con valores similares en las métricas de FFNN en relación a los obtenidos para GB, el resultado de la red LSTM es excepcional, superando con creces a lo logrado por sus competidores. En particular, si se compara el desempeño de la red FFNN en el pronóstico a 10 minutos, esta supera al método de referencia en un 8.58 % según la métrica FS, mientras que la red LSTM lo supera en un 94.97 %. Esta proporción de mejora se mantiene para el horizonte de pronóstico de 30 minutos, pues presenta un FS de 93.32 %, lo cual no ocurre para FFNN, donde sucede un fenómeno atípico acorde a la tendencia vista en los otros modelos GB y ahora LSTM, y es que el FS entregado por la red neuronal tradicional es tan

solo de 0.65 %, es decir, para el horizonte de predicción de 30 minutos, el rendimiento de este modelo y del modelo de referencia es muy similar, al menos en términos de RMSE, lo cual para esta instancia posiciona al modelo FFNN por debajo del modelo GB y por supuesto del desempeño observado para LSTM.

En cuanto al horizonte de pronóstico de 60 minutos, esta vez el modelo FFNN sí supera de manera satisfactoria a kNN, con un FS de 21.37 %, aunque inferior al 28.83 % correspondiente a GB y también al de la red LSTM, que nuevamente se desempeña mejor que el resto de modelos, con un 33.67 % de FS.

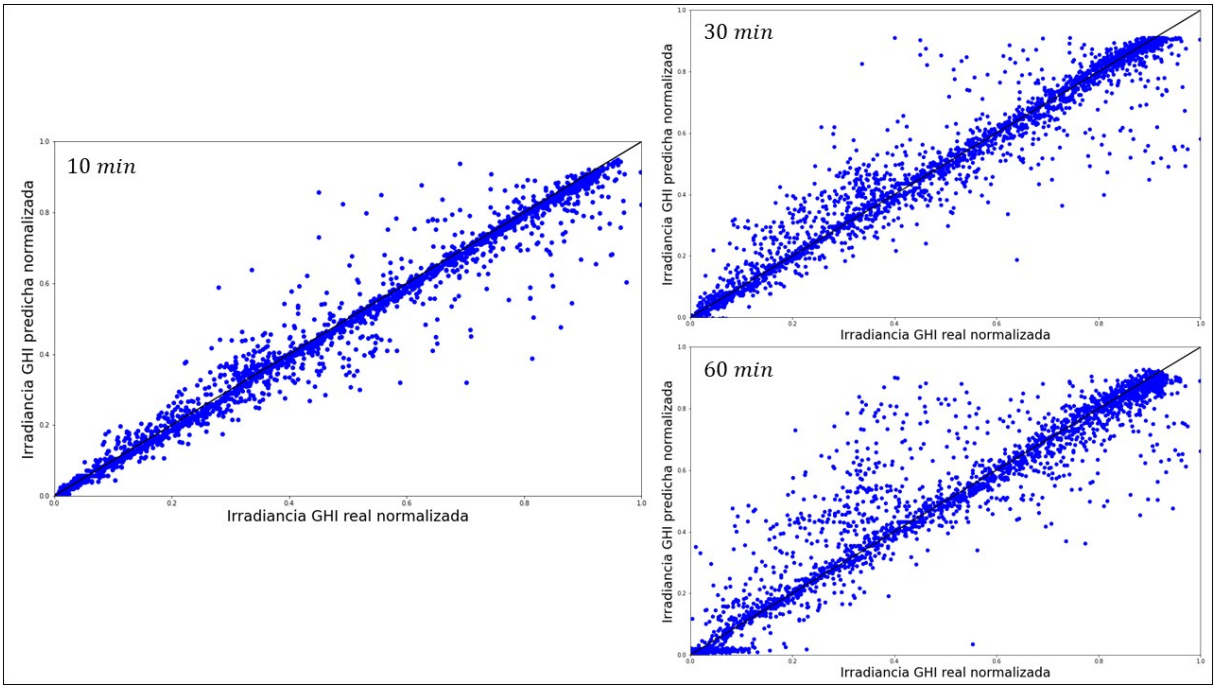

Figura 4.9: Dispersión de pronósticos GHI de modelo FFNN respecto a los valores reales medidos en el mes de diciembre de 2021 para tres horizontes de pronóstico.

En la Figura 4.9 se observan los gráficos de dispersión de la predicción realizada por la red FFNN. Se aprecia una gran concentración de puntos acompañando a la recta diagonal, lo cual reitera la idea del buen desempeño del modelo, donde adicionalmente también se puede observar que la dispersión aumenta proporcionalmente al incrementar el horizonte de pronóstico.

Respecto al sesgo del modelo, se tiene una pequeña oscilación entre subestimación y sobreestimación acorde a la métrica MBE, pero en general estos valores son muy bajos. Por lo que tampoco es posible extraer un patrón claro de sesgo, de tal manera que este puede considerarse insignificante.

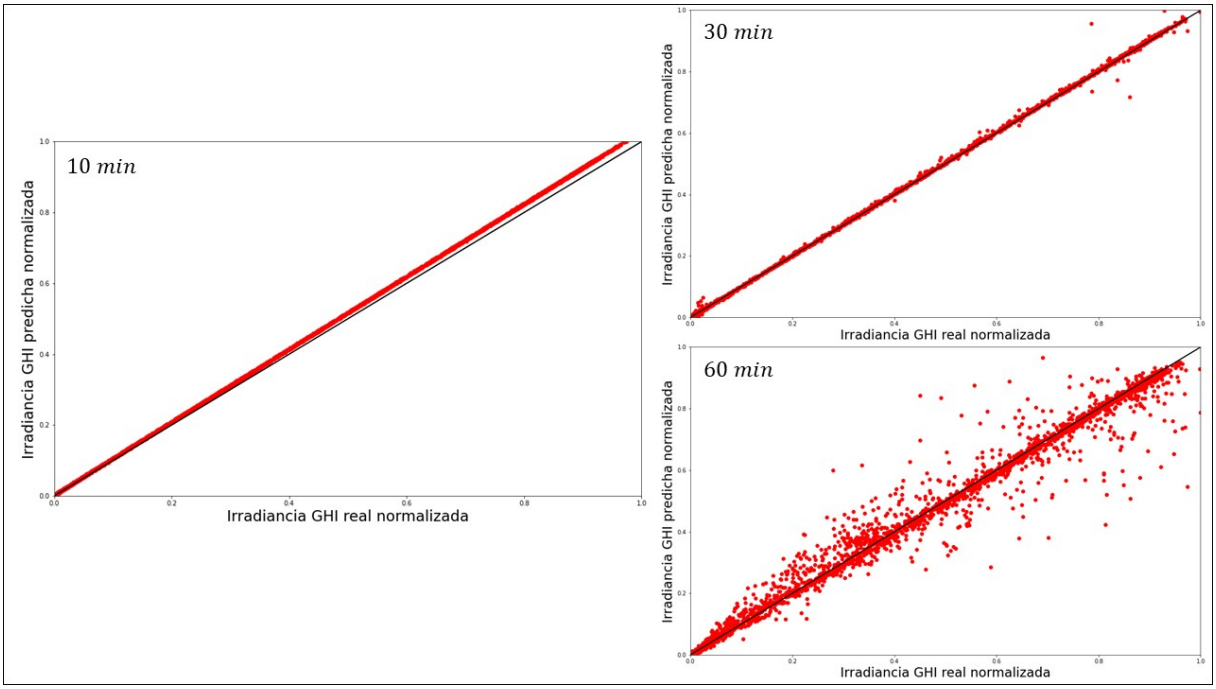

Figura 4.10: Dispersión de pronósticos GHI de modelo LSTM respecto a los valores reales medidos en el mes de diciembre de 2021 para tres horizontes de pronóstico.

Al observar las métricas obtenidas para el modelo LSTM, queda claro que presenta un desempeño sublime respecto a sus competidores, lo cual se comprueba con sus gráficos de dispersión mostrados en la Figura 4.10, de los cuales se puede apreciar una correlación prácticamente perfecta para los pronósticos con horizontes de 10 y 30 minutos, mientras que la predicción con el horizonte de 60 minutos presenta una dispersión existente, pero acotada en comparación al resto de modelos. Un detalle adicional, es que si bien las métricas no indican sesgo significativo en los datos, el escenario con el horizonte de pronóstico de 10 minutos, dentro de su gran resultado, distingue una desviación leve, pero existente, a sobreestimar irradiancias altas, lo cual no genera un impacto en los resultados, pues numéricamente esta desviación es muy pequeña y no es un fenómeno recurrente en ninguna otra configuración, con lo que se puede considerar como un caso aislado.

La capacidad de pronóstico de este modelo muestra ser tan eficiente, que se podría llegar a pensar que el algoritmo sufre de *overfitting*, pero esto se descarta inmediatamente al ver la predicción expuesta en los gráficos de la Figura 4.10. Esta predicción es realizada con datos nunca antes vistos por el modelo, ya que corresponden al mes de diciembre, que como se hace mención en la sección de metodología, estos fueron dejados apartados del set de entrenamiento con la finalidad de comprobar si el modelo efectivamente es capaz de generalizar el problema a resolver o no.

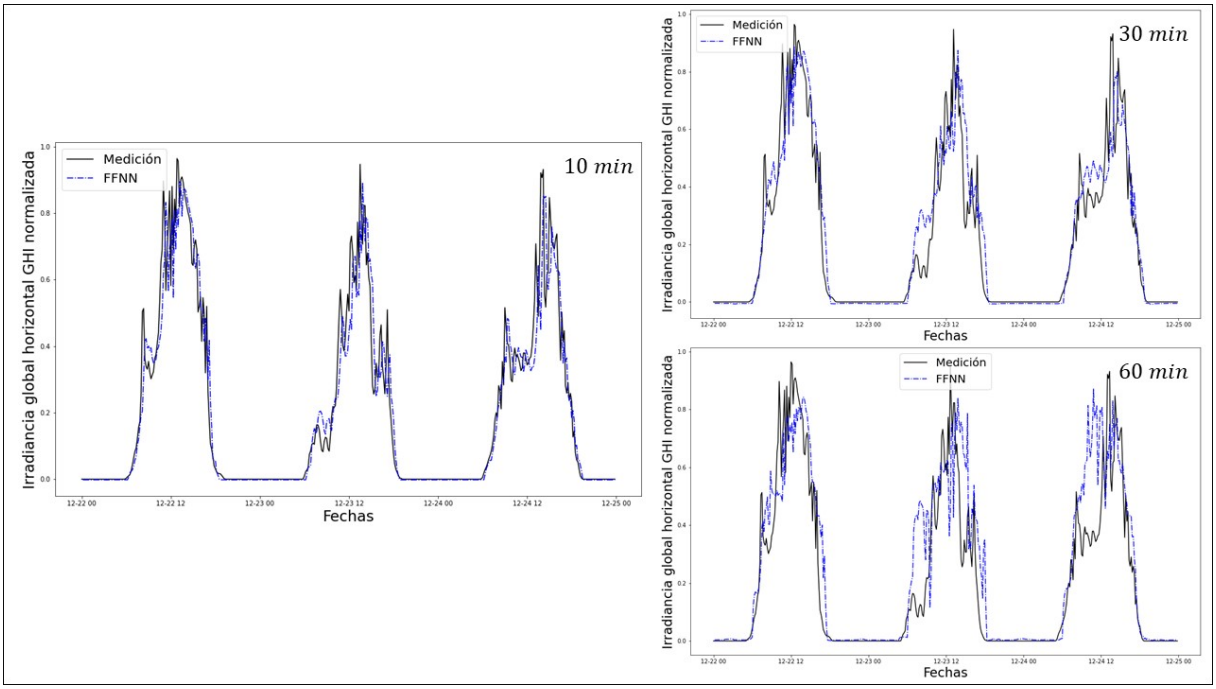

Figura 4.11: Comparación de pronósticos de irradiancia GHI de modelo FFNN en días parcialmente nublados para tres horizontes de pronóstico.

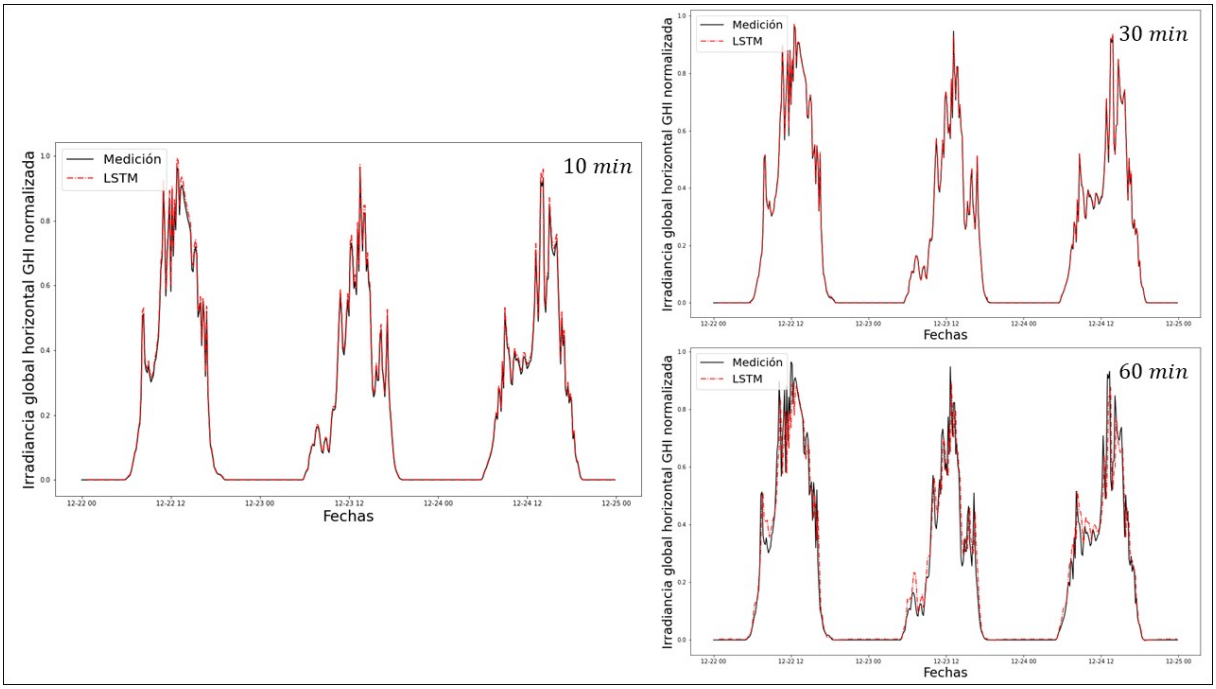

Figura 4.12: Comparación de pronósticos de irradiancia GHI de modelo LSTM en días parcialmente nublados para tres horizontes de pronóstico.

Finalmente, las Figuras 4.11 y 4.12 reafirman el buen desempeño de ambos modelos, inclusive en días nublados. No obstante, se observa que a medida que aumenta el horizonte de pronóstico, el modelo FFNN (Figura 4.11) presenta complicaciones para predecir los picos bajos e intermedios de irradiancia, lo cual no ocurre de manera significativa en el modelo

LSTM (Figura 4.12), ratificando así su buen desempeño aun en el escenario más complejo como lo son los días nublados con un mayor horizonte de pronóstico.

4.2.2. Rendimiento de pronóstico DNI

Por último, queda por analizar cómo se desempeñaron los modelos de redes neuronales al tener que predecir irradiancia DNI, la cual para los otros modelos demostró ser más compleja de pronosticar que la GHI.

De acuerdo a las iteraciones realizadas para el pronóstico de DNI, los mejores resultados obtenidos para los modelos FFNN y LSTM se muestran en las Tablas 4.7 y 4.8 respectivamente.

Tabla 4.7: Métricas estadísticas obtenidas en la evaluación de desempeño de modelo FFNN para pronóstico de DNI.

Horizonte	MAE	RMSE	MBE	Skewness	FS
10 minutos	0.0167	0.0433	0.0073	0.3476	1.26 %
30 minutos	0.0232	0.0621	0.0039	1.1668	9.88 %
60 minutos	0.0401	0.0905	0.0195	1.1037	9.59 %

Tabla 4.8: Métricas estadísticas obtenidas en la evaluación de desempeño de modelo LSTM para pronóstico de DNI.

Horizonte	MAE	RMSE	MBE	Skewness	FS
10 minutos	0.0023	0.0036	-0.0025	-0.2202	91.81 %
30 minutos	0.0022	0.0026	0.0009	2.3581	96.21 %
60 minutos	0.0175	0.0454	-0.0033	-0.0750	54.70 %

Como era esperado, el error MAE y RMSE aumenta para ambos modelos respecto de sus equivalentes para el pronóstico de GHI, a excepción del MAE correspondiente al pronóstico de DNI con horizonte de 60 minutos para LSTM, el cual disminuye ligeramente respecto al obtenido para el pronóstico de irradiancia global.

Respecto al desempeño en relación al método de referencia, FFNN presenta un rendimiento levemente superior acorde a la métrica FS, con valores de 1.26 % para el horizonte de 10 minutos, 9.88 % para el horizonte de 30 minutos y 9.59 % para la ventana de predicción correspondiente a 60 minutos. Por otro lado, la red LSTM nuevamente supera notablemente a los otros tres métodos para todos los horizontes de tiempo. Esto se hace tomando como base comparativa la métrica FS, la cual muestra un resultado del 91.81 % superior al modelo de referencia en el horizonte de tiempo más corto, 96.21 % en el horizonte intermedio y un 54.70 % para el último escenario.

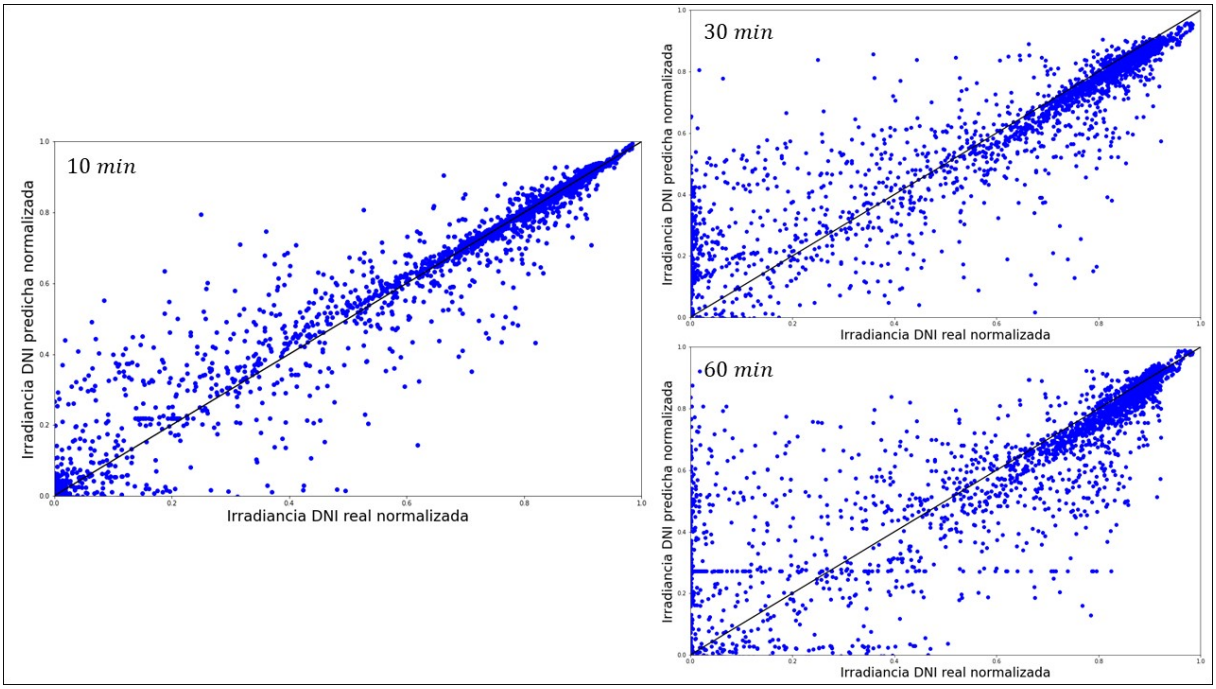

Figura 4.13: Dispersión de pronósticos DNI de modelo FFNN respecto a los valores reales medidos en el mes de diciembre de 2021 para tres horizontes de pronóstico.

En relación a los gráficos de dispersión del modelo FFNN, mostrados en la Figura 4.13, se puede observar que de la misma manera que ocurrió para kNN y GB, este modelo presenta una gran dispersión de puntos para irradiancias intermedias, lo cual no ocurría en los pronósticos de GHI. Nuevamente los cúmulos cercanos a la recta se observan para irradiancias altas, este fenómeno ocurre principalmente dada la información adicional de la correlación horaria para los valores máximos de irradiancia, idea previamente discutida en el pronóstico de DNI de los modelos kNN y GB.

Algo llamativo que es posible apreciar en el gráfico para el horizonte a 60 minutos es que el modelo FFNN presenta una predilección a asignar un valor cercano a 0.3 para una cierta combinación de *features*, a tal punto de que destaca visualmente una línea de puntos perfectamente horizontal para este valor predicho. Esto corresponde claramente a un error de reconocimiento de patrones en el modelo.

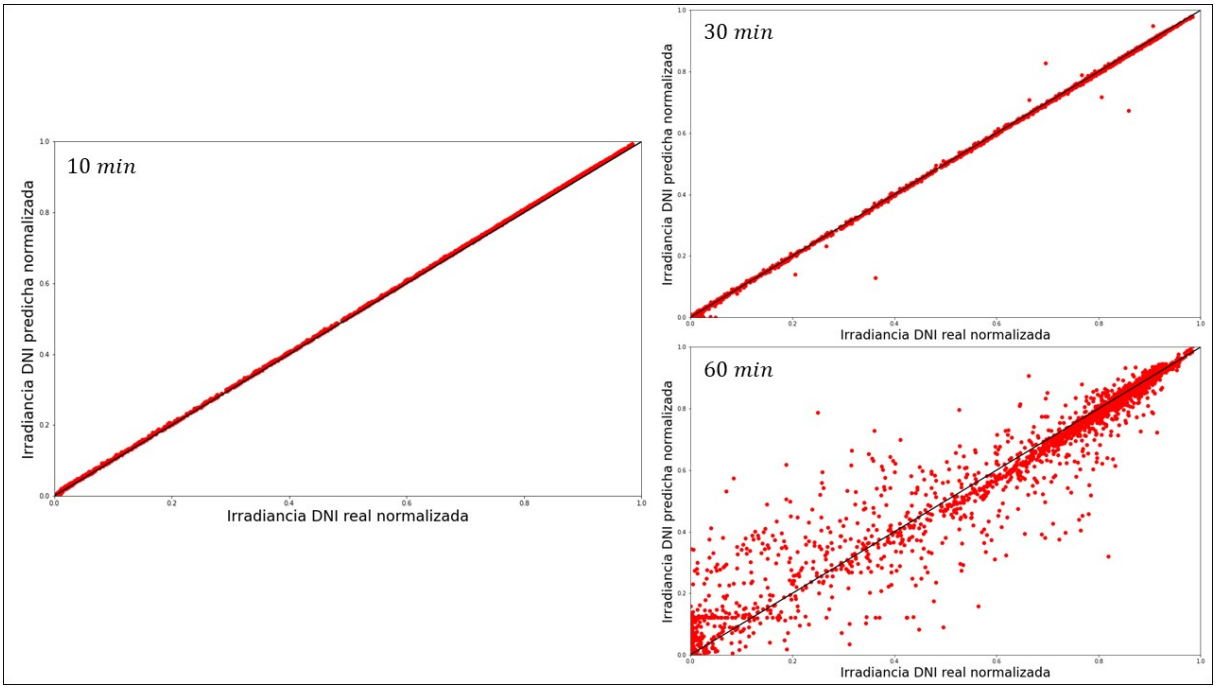

Figura 4.14: Dispersión de pronósticos DNI de modelo LSTM respecto a los valores reales medidos en el mes de diciembre de 2021 para tres horizontes de pronóstico.

Reafirmando los resultados expuestos por las métricas MAE y RMSE, la Figura 4.16 muestra la dispersión del pronóstico de DNI realizado por el modelo LSTM, El patrón es prácticamente el mismo que el visto para la predicción GHI con horizontes de 10 y 30 minutos, los cuales presentan una correlación muy elevada, sin embargo, se logra observar un aumento de dispersión para el horizonte de 60 minutos. Esto representa un comportamiento similar a los otros modelos para este mismo escenario, donde se desempeña mejor en irradiancias altas y no tanto así para las irradiancias intermedias. Evidentemente, la dispersión sigue siendo más acotada para LSTM que en sus modelos rivales, pero es un patrón que esta red no había mostrado previamente en otras configuraciones, tanto de GHI como de DNI. Adicionalmente, si se es riguroso, es posible observar para este mismo horizonte de pronóstico nuevamente el fenómeno de predilección de un valor puntual visto en la red FFNN, esta vez más cercano a 0.1 y no abarcando todo el espectro, pero lo suficiente para visualizarse una línea horizontal en este valor.

Con relación al sesgo, se repite lo visto para el pronóstico de GHI, donde no hay una tendencia clara acorde a las métricas MBE y *Skewness*, reincidiendo en que ambos modelos no presentan un sesgo significativo.

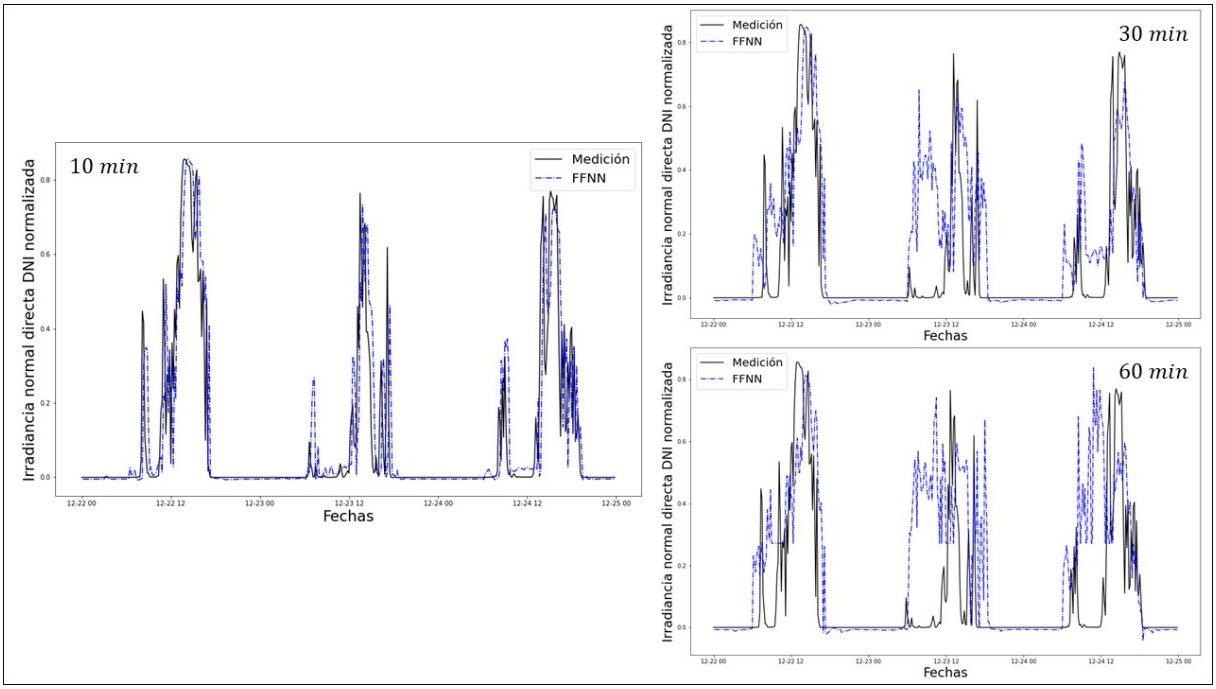

Figura 4.15: Comparación de pronósticos de irradiancia DNI de modelo FFNN en días parcialmente nublados para tres horizontes de pronóstico.

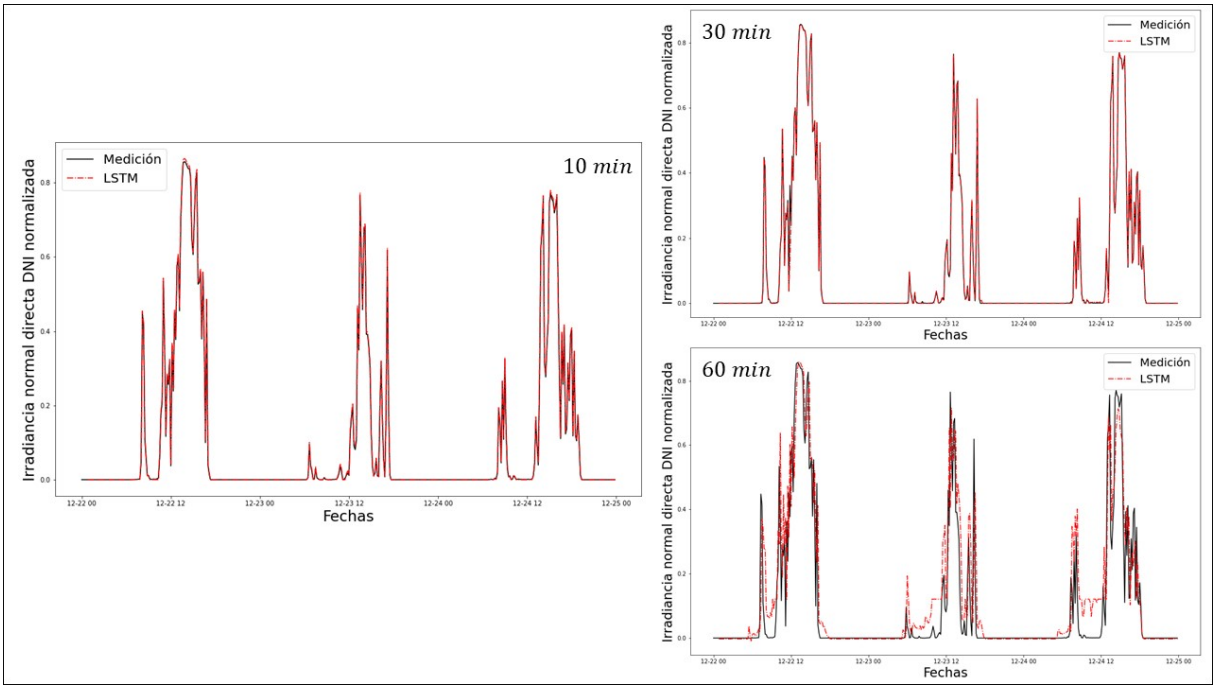

Figura 4.16: Comparación de pronósticos de irradiancia DNI de modelo LSTM en días parcialmente nublados para tres horizontes de pronóstico.

Finalmente la Figura 4.15 evidencia la dificultad de predecir valores de DNI intermedios por parte del modelo FFNN, principalmente para los horizontes de 30 y 60 minutos. Luego, en cuanto a la Figura 4.16, esta muestra el buen rendimiento de la predicción LSTM incluso en días nublados, teniendo pequeñas diferencias en algunos de los picos de irradiancia del

pronóstico a 60 minutos, más notorias que las presentes en el pronóstico de GHI realizado por el mismo modelo, pero sin mayor relevancia, sobretodo al comparar con los resultados obtenidos para los otros modelos.

4.3. Resumen y comparación

Con el objetivo de facilitar una comparación global entre todos los modelos estudiados en el presente trabajo de título, se exponen las Tablas 4.9 y 4.10 con un resumen de las métricas obtenidas en el pronóstico de irradiancia GHI y DNI respectivamente, donde como puntos relevantes y acorde a lo visto durante toda la sección de resultados, el modelo que predomina sobre el resto es la arquitectura LSTM, con un rendimiento muy superior a sus modelos rivales. Adicionalmente, tanto GB como la red FFNN se desempeñan de mejor manera que el modelo kNN en casi todos los escenarios con rendimientos variados entre sí, particularmente para GHI el modelo de ensamble de árboles de decisión toma la ventaja exceptuando en el horizonte de pronóstico a 10 minutos, pero en el pronóstico de DNI domina la red neuronal tradicional, no obstante sus resultados son similares y de todas maneras no se acercan a lo realizado por la red LSTM.

Tabla 4.9: Resumen de resultados entregados por las métricas estadísticas en el pronóstico de GHI para tres horizontes de pronóstico.

Horizonte	Modelo	MAE	RMSE	MBE	Skewness	FS
10 minutos	kNN	0.0073	0.0219	-0.0012	-1.3230	0.00 %
	GB	0.0071	0.0216	-0.0009	-1.5528	1.38 %
	FFNN	0.0066	0.0200	-0.0019	-1.0245	8.58 %
	LSTM	0.0004	0.0011	-0.0004	-0.2359	94.97 %
30 minutos	kNN	0.0143	0.0329	-0.0009	-0.1418	0.00 %
	GB	0.0131	0.0298	-0.0035	-1.6153	9.53 %
	FFNN	0.0110	0.0327	0.0085	1.5407	0.65 %
	LSTM	0.0012	0.0022	0.0006	0.0872	93.32 %
60 minutos	kNN	0.0252	0.0520	-0.0011	0.3509	0.00 %
	GB	0.0202	0.0370	-0.0040	-0.1875	28.83 %
	FFNN	0.0270	0.0409	-0.0127	-0.6179	21.37 %
	LSTM	0.0198	0.0345	0.0195	0.1792	33.67 %

Tabla 4.10: Resumen de resultados entregados por las métricas estadísticas en el pronóstico de DNI para tres horizontes de pronóstico.

Horizonte	Modelo	MAE	RMSE	MBE	Skewness	FS
10 minutos	kNN	0.0137	0.0439	0.0057	0.3392	0.00 %
	GB	0.0151	0.0454	0.0019	-0.5019	-3.58 %
	FFNN	0.0167	0.0433	0.0073	0.3476	1.26 %
	LSTM	0.0023	0.0036	-0.0025	-0.2202	91.81 %
30 minutos	kNN	0.0252	0.0689	0.0100	0.7014	0.00 %
	GB	0.0250	0.0659	0.0089	0.3198	4.35 %
	FFNN	0.0232	0.0621	0.0039	1.1668	9.88 %
	LSTM	0.0022	0.0026	0.0009	2.3581	96.21 %
60 minutos	kNN	0.0426	0.1001	0.0136	0.9557	0.00 %
	GB	0.0421	0.0911	0.0129	0.8815	8.99 %
	FFNN	0.0401	0.0905	0.0195	1.1037	9.59 %
	LSTM	0.0175	0.0454	-0.0033	-0.0750	54.70 %

Capítulo 5

Conclusiones

Para finalizar, se exponen las conclusiones que surgen del trabajo de título realizado. La primera de ellas es que se puede establecer que los cuatro modelos estudiados: kNN, GB, FFNN y LSTM, presentan una capacidad de pronóstico de irradiancia satisfactoria para horizontes de pronóstico a muy corto plazo, particularmente menores a 30 minutos, en el caso de que estos sean optimizados correctamente. Los modelos ven afectada su capacidad de pronóstico a medida que aumenta la ventana de tiempo a predecir, lo cual es esperable. A pesar de esto, uno de los métodos estudiados es capaz de generar un pronóstico suficientemente confiable y preciso para un horizonte de pronóstico de 60 minutos, y dicho método corresponde a la red LSTM.

La arquitetura LSTM propuesta es la que logra los mejores resultados respecto a sus rivales para todos los escenarios estudiados, incluyendo un desempeño notable para los días parcialmente nublados, tanto en pronóstico de GHI como de DNI. Para el pronóstico de GHI se obtuvo un MAE de 0.0004 para el horizonte de tiempo de 10 minutos del modelo LSTM, lo que representa un error del 0.04 %. Mientras que para el escenario más complejo, es decir, el horizonte de tiempo de 60 minutos, el MAE equivale a un error de pronóstico de 1.98 %, lo cual sigue estando en un rango bastante bajo. De la misma manera, en el pronóstico de irradiancia directa, el MAE obtenido muestra un error que va desde un 0.23 % hasta un 1.75 % de error de pronóstico, desde el menor horizonte de tiempo al mayor, respectivamente. En este aspecto, la gran diferencia que presenta este método, respecto de la red FFNN, y por tanto también de los otros modelos kNN y GB, es su capacidad de recursión de los *outputs* obtenidos, en conjunto con su memoria selectiva, de tal manera que es posible afirmar que estos atributos le otorgan al modelo una capacidad para identificar, extraer e interpretar patrones superior que el de sus competidores, al trabajar con datos correspondientes a series de tiempo de irradiancia.

Ahora bien, haciendo alusión a que el método kNN fue designado como modelo de referencia, dada la simpleza de su construcción y menor tiempo de cómputo, se esperaba que los otros tres modelos excedieran su desempeño, lo cual fue ratificado para cada escenario en todos los algoritmos, a excepción de un caso aislado, correspondiente al pronóstico de DNI a 10 minutos realizado por el método GB, que presenta un FS de -3.58 %. No obstante, sigue manteniéndose un resultado en un rango muy similar al método de referencia, afectado directamente por el comportamiento propio de la variable DNI, que viene asociada con una mayor cuota de incertidumbre que la irradiancia global.

Es importante recalcar la diferencia entre pronosticar irradiancia GHI y DNI. Para todos los modelos estudiados y bajo todos los horizontes de tiempo pronosticados, la predicción de irradiancia directa demostró ser menos confiable, al presentar una distribución de error mayor que el resto de pronósticos realizados para irradiancia global. La principal razón de que ocurra este fenómeno es la naturaleza errática de los datos registrados para DNI en comparación a los de su otra componente, lo cual es causado por las características físicas del equipo de medición junto a restricciones angulares de la propia irradiancia normal directa.

Otro aspecto fundamental a revisar era si los modelos presentaban sesgo al momento de realizar las predicciones, ya sea a subestimar o sobreestimar la irradiancia real. Según los resultados obtenidos para MBE y *Skewness*, ninguno de los modelos presenta un sesgo significativo, o al menos, no con un patrón reconocible, a excepción del modelo kNN, el cual si bien sus métricas no indican una tendencia de sesgo, desprende un comportamiento donde el modelo muestra una sobreestimación para irradiancias bajas y por el contrario, una subestimación de irradiancias altas (ver Figuras 4.1, 4.3, 4.5, 4.7), lo cual se compensa al calcular errores, originando que los valores de MBE se mantengan exhibiendo un resultado neutro para este modelo.

Es sustancial notar que los resultados expuestos están limitados al set de datos utilizados. Adicionalmente, existen consideraciones que podrían ser de utilidad para trabajos futuros, entre ellas, optimizar a un mayor nivel la exploración de hiperparámetros de los modelos, dado que en el presente trabajo, para varios de estos parámetros se optó por valores predeterminados por simplicidad, de tal manera que los resultados no necesariamente muestran el mejor *output* posible. Otra consideración a analizar es el realizar el entrenamiento de modelos con otra normalización, como por ejemplo la aplicación del índice de cielo claro k_c, que consiste en la irradiancia real respecto a la irradiancia correspondiente a las condiciones óptimas de cielo despejado para ese instante de tiempo, lo que es una práctica común en trabajos relacionados con el pronóstico de energía solar, y comparar los resultados de usar esta metodología con los obtenidos en este trabajo, donde se utiliza el valor máximo de cada una de las componentes de la irradiancia para su respectiva normalización.

Finalmente, en lo que respecta a los algoritmos estudiados, particularmente a las redes neuronales, acorde a la literatura, se tiene la premisa de que son más efectivas al combinarse con conjuntos grandes de datos, pues se generan más parámetros entrenables, con lo que incorporar más variables a la data podría resultar beneficioso. Trabajos relacionados como el mencionado previamente realizado por Pedro et al. [3], muestra los resultados al agregar adicionalmente a la información de variables ambientales, imágenes satelitales a sus modelos (kNN y GB), lo cual puede resultar de interés si se aplica a modelos de redes neuronales recurrentes, como lo es la LSTM.

Bibliografía

[1] Asociación Chilena de Energías Renovables y Almacenamiento AG (ACERA). (2022). Estadísticas Sector de generación de energía eléctrica renovable septiembre 2022. ACERA.

[2] Wang, F., Mi, Z., Su, S., and Zhao, H. (2012). Short-term solar irradiance forecasting model based on artificial neural network using statistical feature parameters. Energies, 5(5), 1355-1370.

[3] Pedro, H. T., Coimbra, C. F., David, M., and Lauret, P. (2018). Assessment of machine learning techniques for deterministic and probabilistic intra-hour solar forecasts. Renewable Energy, 123, 191-203.

[4] Gensler, A., Henze, J., Sick, B., and Raabe, N. (2016, October). Deep Learning for solar power forecasting—An approach using AutoEncoder and LSTM Neural Networks. In 2016 IEEE international conference on systems, man, and cybernetics (SMC) (pp. 002858-002865). IEEE.

[5] Inman, R. H., Pedro, H. T., and Coimbra, C. F. (2013). Solar forecasting methods for renewable energy integration. Progress in energy and combustion science, 39(6), 535-576.

[6] Kelleher, D. J. and Tierney, B. (2021). Ciencia de datos (MIT Press / Conocimientos Esenciales) (Spanish Edition). Ediciones UC.

[7] Cai, Y. L., Ji, D., and Cai, D. (2010, June). A KNN Research Paper Classification Method Based on Shared Nearest Neighbor. In NTCIR (pp. 336-340).

[8] Hastie, T. (2009). Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning.

[9] Maind, S. B., and Wankar, P. (2014). Research paper on basic of artificial neural network. International Journal on Recent and Innovation Trends in Computing and Communication, 2(1), 96-100.

[10] Patterson, J., and Gibson, A. (2017). Deep learning: A practitioner's approach. " O'Reilly Media, Inc.".

[11] Elsworth, S., and Güttel, S. (2020). Time series forecasting using LSTM networks: A symbolic approach. arXiv preprint arXiv:2003.05672.

[12] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

[13] Wang, W., and Lu, Y. (2018, March). Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model. In IOP conference series: materials science and engineering (Vol. 324, No. 1, p. 012049). IOP Publishing.

Anexo

A continuación se adjunta el acceso a un repositorio que contiene el set de datos y el desarrollo del código utilizado para la realización del presente trabajo de título, con la finalidad de que el lector tenga la posibilidad de profundizar en caso de que así lo requiera.

Acceso GitHub: https://github.com/joseignmoya/Solar-forecasting-models