Informe Proyecto Deep Learning

Darwin Agudelo Hernández
Juliana Carvajal Guerra
José Alejandro Urrego Pabón

Introducción

El presente proyecto tiene como objetivo predecir la velocidad máxima del viento a 10 metros de altura en un punto específico de la región de Medellín, utilizando datos históricos proporcionados por la plataforma NASA POWER. Este objetivo responde a la necesidad de contar con herramientas predictivas que apoyen la toma de decisiones en aplicaciones relacionadas con las energías renovables, gestión de recursos naturales y planificación de infraestructuras seguras. Inicialmente, el proyecto se planteó para predecir simultáneamente la velocidad del viento en múltiples ubicaciones. Sin embargo, el enfoque evolucionó hacia la predicción de un único punto objetivo, utilizando como entrada los datos históricos de otras ubicaciones cercanas, con el fin de aprovechar tanto las relaciones espaciales como temporales en los datos disponibles.

En esta reformulación, se determinó que las coordenadas iniciales que se habían planteado contenían redundancia debido a la resolución espacial de los datos de NASA POWER, de aproximadamente 0.5°x0.5°. Esto llevó a una redefinición de las ubicaciones objetivo y a una mejor representación de la variabilidad espacial en la región. Las coordenadas finales seleccionadas abarcan seis puntos: cinco puntos de apoyo (6.75, -75.6; 6.5, -75.25; 6, -75.25; 6, -75.95; 6.5, -75.95) y un punto principal para predicción (6.25, -75.6). Cada uno cuenta con registros diarios de la velocidad máxima del viento, generando un conjunto de datos que cubre un período de 10 años (2014-2023) con 3652 observaciones por ubicación.

El modelo propuesto se basa en redes neuronales recurrentes (RNN), diseñadas específicamente para capturar patrones temporales en series de tiempo. Además, la incorporación de datos de varias ubicaciones como entradas busca enriquecer las predicciones en el punto objetivo.

Descripción de los Notebooks

Notebook 01: Preparación de Datos

El primer notebook se centra en la adquisición, limpieza y consolidación de los datos históricos de la velocidad máxima del viento (WS10M_MAX) para las seis ubicaciones de interés.

Acceso y descarga de datos

- Los archivos CSV con los datos históricos están almacenados en un repositorio de GitHub.
- Utilizando el paquete requests, se descargan los archivos directamente desde el repositorio.

Procesamiento de los archivos

- Los datos de cada archivo se cargan en un DataFrame utilizando pandas, omitiendo las primeras líneas irrelevantes.
- Las columnas originales (YEAR, MO, DY, WS10M_MAX) se renombran para garantizar uniformidad (year, month, day, WS10M_MAX_{ubicación}).
- Se genera una nueva columna Date combinando las columnas de año, mes y día, lo que permite un formato estándar para las fechas.

Consolidación de los datos

- Cada archivo procesado se almacena en un diccionario, con claves que identifican las ubicaciones y valores que contienen los respectivos DataFrames.
- Los datos se combinan en un único DataFrame utilizando la columna Date como clave de unión.

Resultados

- Se obtiene un DataFrame consolidado donde cada fila corresponde a una fecha y las columnas representan la velocidad máxima del viento para cada una de las seis ubicaciones.

Date	punto1	punto2	punto3	punto4	punto5	punto6
1/01/2014	1.93	1.35	1.01	2.17	2.55	1.41
2/01/2014	2.03	1.55	1.45	2.39	3.48	1.06
3/01/2014	1.29	1.49	1.3	1.78	2.66	1.5
...						

Notebook 02: Análisis Exploratorio

El objetivo de este notebook fue realizar un análisis detallado de los datos combinados para comprender sus características fundamentales, las relaciones entre las ubicaciones, y los patrones temporales y estacionales presentes en la velocidad máxima del viento.

Exploración inicial del dataset

El dataset tiene 3652 registros por ubicación, con velocidades máximas del viento que varían entre 0.59 m/s y 6.3 m/s. La media de la velocidad máxima en el punto objetivo (Punto 6) es de 1.68 m/s, con una desviación estándar de 0.43 m/s. Esto indica que los valores están centrados en rangos moderados, con eventos extremos poco frecuentes.

Relaciones entre ubicaciones: Matriz de correlación

Se generó una matriz de correlación que muestra relaciones importantes entre las velocidades del viento en diferentes ubicaciones. Los puntos cercanos tienen correlaciones más altas:

- Puntos 4 y 5 tienen una correlación de 0.88, lo que indica patrones de viento similares en estas áreas.
- El Punto 6 presenta correlaciones más moderadas con otros puntos (ej., 0.47 con el Punto 1), indicando que combina patrones de varias ubicaciones.

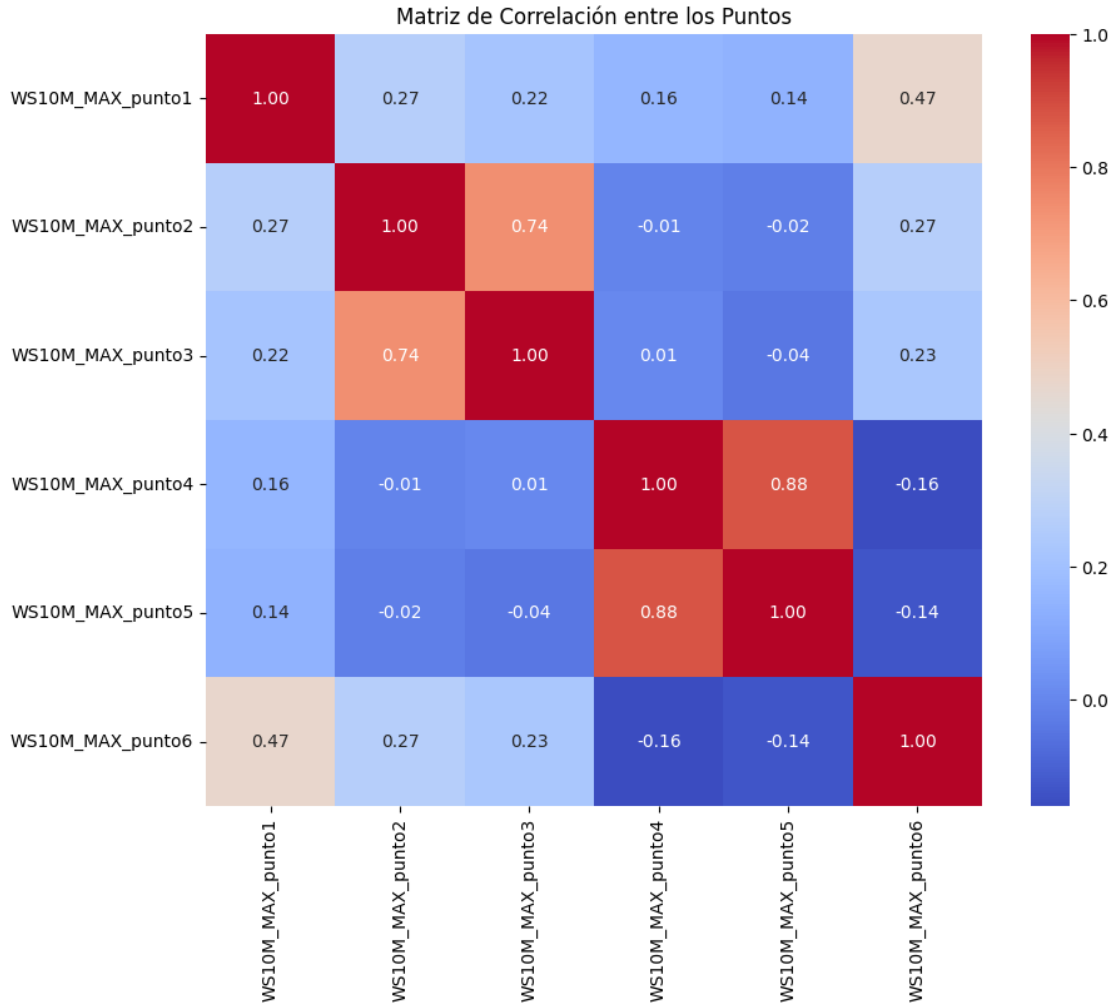

Figura 1. Matriz de correlación de Pearson.

Tendencias temporales

Se identificaron patrones consistentes en las velocidades mensuales del viento. Todas las ubicaciones presentan fluctuaciones cíclicas, con máximos recurrentes en determinados periodos del año. La Figura 2 muestra cómo los picos y valles varían entre puntos, destacando la estacionalidad.

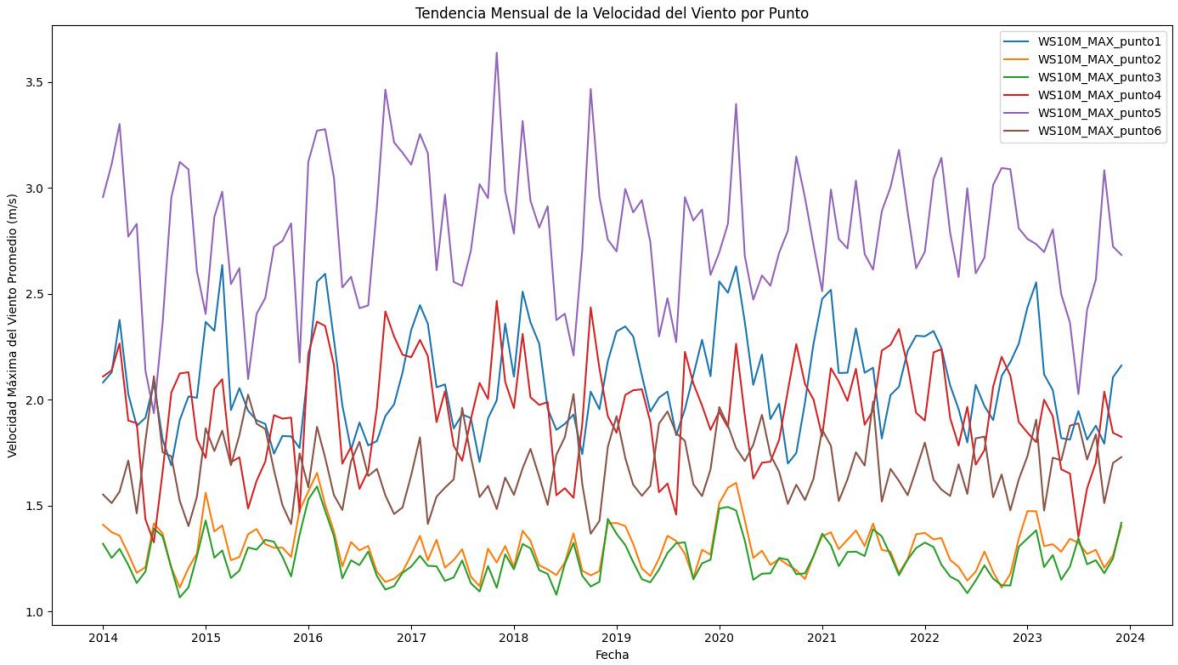

Figura 2. Tendencia de la velocidad del viento durante el tiempo de muestreo.

Análisis estacional en el punto 6

El promedio mensual revela que los meses de verano (enero, julio y agosto) tienen mayores velocidades, mientras que abril y noviembre registran valores más bajos.

Autocorrelación

El análisis de autocorrelación del Punto 6 evidencia dependencias temporales significativas, especialmente en los primeros días, lo que justifica el uso de modelos secuenciales como RNN.

Distribución de velocidades

La distribución de la velocidad máxima del viento en el Punto 6 tiene una forma asimétrica, con la mayoría de los valores concentrados entre 1.5 m/s y 2 m/s.

Notebook 03: Preprocesamiento de Datos

Este notebook transforma los datos originales en un formato adecuado para el modelo RNN. Incluye:

- Normalización de las variables numéricas al rango [0, 1].
- Estructuración de los datos en secuencias temporales con ventanas de 30 días y horizonte de pronóstico de 7 días.
- División en conjuntos de entrenamiento (80%) y prueba (20%).

Resultado: Conjuntos X (características) e y (objetivo) listos para el entrenamiento del modelo.

Notebook 04: Modelo RNN

En este notebook se diseñó, entrenó y validó un modelo basado en redes neuronales recurrentes (RNN) para predecir la velocidad máxima del viento:

Arquitectura del modelo

- Capa LSTM para procesar secuencias temporales.
- Dropout y regularización L2 para prevenir sobreajuste.
- Capa densa para producir una salida de 7 valores.

Entrenamiento

- Función de pérdida: MSE.
- Optimizador: Adam.
- Early Stopping para detener el entrenamiento cuando la pérdida de validación no mejora.

Descripción de la solución

La solución propuesta para predecir la velocidad máxima del viento en el Punto 6 combina un preprocesamiento riguroso de los datos con una arquitectura de Red Neuronal Recurrente (RNN) optimizada para modelar relaciones espaciales y temporales.

Preprocesamiento y generación de datos secuenciales

El preprocesamiento se diseñó para estructurar y preparar los datos de entrada en un formato adecuado para las RNN:

- Todas las variables (WS10M_MAX de los seis puntos) se escalaron al rango [0, 1] utilizando MinMaxScaler.
- Esto asegura que las diferencias en magnitudes entre ubicaciones no interfieran con el aprendizaje del modelo y mejora la convergencia del optimizador.
- Se implementó una función personalizada para crear ventanas deslizantes de 30 días (window_size) como entrada (X) y un horizonte de predicción de 7 días (forecast_horizon) como salida (y).
- Este formato permite que el modelo aprenda patrones temporales a mediano plazo y realice predicciones multivariadas.
- El 80% de los datos se utilizó para el entrenamiento y el 20% restante para la evaluación.

Resultado: Los datos fueron transformados en matrices de entrada (X) y salida (y) con las dimensiones:

- Entrenamiento: X=(2892, 30, 5) y y=(2892, 7).
- Prueba: X=(724, 30, 5) y y=(724, 7).

Arquitectura del modelo

El modelo de predicción fue diseñado para capturar relaciones complejas tanto en el tiempo como entre ubicaciones mediante una arquitectura basada en RNN.

- Capa LSTM:
  - Unidades: 32.
  - Activación: tanh para gestionar eficientemente relaciones no lineales en los datos temporales.
  - Regularización L2: Con un parámetro de penalización de 0.001, esta técnica ayuda a prevenir el sobreajuste al restringir los valores de los pesos.
- Dropout:
  - Tasa: 30% de las unidades se apagan aleatoriamente durante el entrenamiento, reduciendo el riesgo de coadaptación entre las unidades.
- Capa Densa (salida):
  - Unidades: 7, correspondientes al horizonte de predicción de 7 días.
  - Regularización L2: También con un parámetro de 0.001, para mejorar la estabilidad de las predicciones.
- Hiperparámetros:
  - Optimizador: Adam (aprendizaje adaptativo eficiente).
  - Función de pérdida: MSE (Error Cuadrático Medio), seleccionada por penalizar más los errores grandes, críticos en un problema climático.
  - Tamaño del lote: 32.
  - Épocas: 50 (aunque se utilizó Early Stopping para detener el entrenamiento anticipadamente si la pérdida de validación no mejoraba después de 5 épocas).

Durante el entrenamiento, se observó una convergencia adecuada de las pérdidas de entrenamiento y validación, como se muestra en la Figura 3. La implementación de Early Stopping monitoreó la pérdida de validación y detuvo el entrenamiento cuando no hubo mejoras después de 5 épocas, previniendo el sobreajuste y reduciendo el tiempo de cómputo.

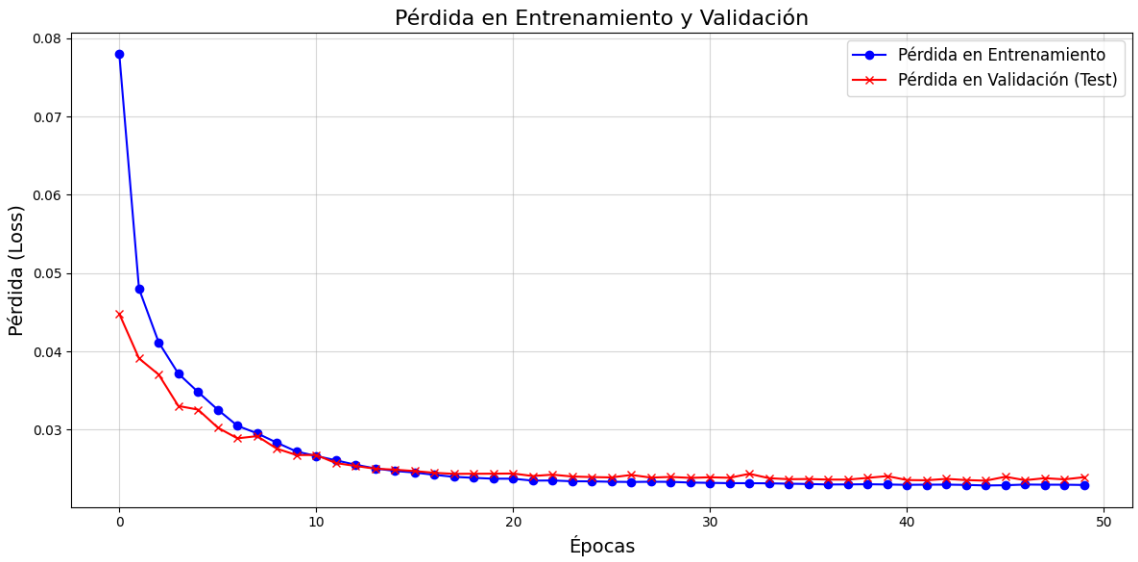

Figura 3. Tendencia de la velocidad del viento durante el tiempo de muestreo.

Resultados y métricas de desempeño

MAE (Error absoluto medio)

La métrica de desempeño elegida para la predicción de las velocidades del viento en varios puntos de la ciudad será el MAE (Error Absoluto Medio) que nos permite comparar y cuantificar la diferencia entre las velocidades del viento reales (Las reportadas por la NASA) y las obtenidas mediante la red neuronal recurrente **Shabbir et al 2021**, cabe aclarar que las velocidades del viento en esta zona al ser relativamente bajas (2-4 m/s), se podrían tolerar diferencias hasta de una unidad, es decir:

- MAE<1 (Excelente) Si las magnitudes de las variables estudiadas son pequeñas
- MAE entre 1 y 5 (Aceptable) Dependiendo de las magnitudes
- MAE>5 (Inaceptable) Motivo de revisión y mejora

Al realizar la comparación mediante el error absoluto medio (MAE) se obtuvo una diferencia de 0.1213 entre los valores reales y los predichos con el modelo, lo cual corresponde a una red neuronal excelente.

```
MAE: 0.12136279662634696
```

Figura 4. Error absoluto medio obtenido a partir de la red neuronal recurrente.

Acceso a los datos y ejecución

Los datos fueron extraídos de la plataforma de la nasa “NASA POWER” de donde se tomaron varios puntos ya antes explicados, estos puntos fueron acomodados en varios archivos .csv que se encuentran en los distintos repositorios del equipo, además, se combinaron los datos de todo los csv en un archivo llamado “combined_wind_data.csv” con el objetivo de llamarlo en los notebooks posteriores.

NOTA: Se recomienda ejecutar los notebooks en el orden que indica su subíndice, es decir, 0.1 antes que el 0.2 y así sucesivamente hasta llegar al 0.4 que contiene el modelo entrenado de la red neuronal recurrente.

Conclusiones y futuras mejoras

- Se concluye por medio el MAE obtenido (Error absoluto medio) de 0.1213 aproximadamente es bueno para el rango de los datos de las velocidades del viento (1.2 - 3) ya que indica que las predicciones del modelo están relativamente cerca a los valores reales, el éxito final de esta red neuronal recurrente radica en que al ser datos de entrenamiento tan variados según el punto que se elija de Medellín, los resultados seguirán ese patrón no contante en el recurso eólico.

Referencias

- Mohamed, M., Rehman, S., Nuha, H., Islam, M. S., & Schulze, F. H. (2021). Accuracy of wind speed predictability with heights using Recurrent Neural networks. FME Transactions, 49(4), 909.