# Predicción de la calidad del aire de la ciudad de Medellín y su área metropolitana mediante el uso de redes neuronales recurrentes.

**Autor: Oscar Jovanni Maestre Sanmiguel**
**Tutor: Raúl Parada Medina**
**PRA: Jordi Casas Roma**

Máster Universitario de Ciencia de Datos
UOC

# ÍNDICE

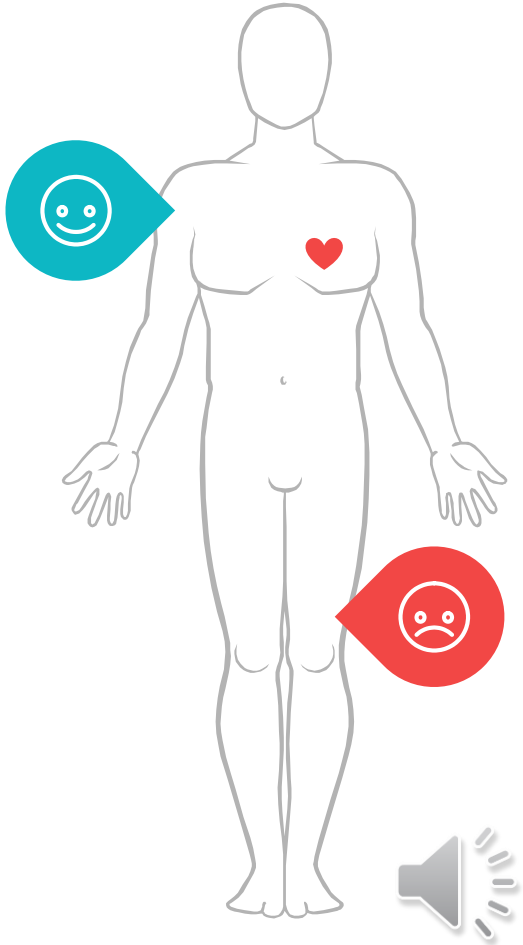

1. **Introducción**
   - Justificación
   - Objetivos
   - Metodología
2. **Diseño e implementación**
   - Origen de los datos
   - Exploración de los datos
   - Preparación de los datos
   - Modelamiento
3. **Evaluación**
   - Resultados
   - Conclusiones

# 1. INTRODUCCIÓN

# JUSTIFICACIÓN

- **La contaminación del aire**

  - Efectos negativos sobre la salud y la economía
  - Aumento de enfermedades respiratorias agudas
  - Relación con muertes prematuras
  - Deterioro de los materiales
  - Efectos sobre los seres vivos



# JUSTIFICACIÓN

## Medellín y área metropolitana

- Trastornos neuropsicológicos en sus habitantes
- Enfermedades cardiorrespiratorias
- Cáncer de pulmón relacionados con el deterioro de la calidad del aire
- Emisión de 1.230 toneladas por año de material particulado PM2.5
- Acciones reactivas por parte de las autoridades

*Figura 1. Centro de Medellín durante contingencia ambiental en el 2017*

*(Fuente: El mundo, https://www.elmundo.com/noticia/-Que-tan-danino-es-el-aire-que-se-respira-en-Medellin-/373516)*

# JUSTIFICACIÓN

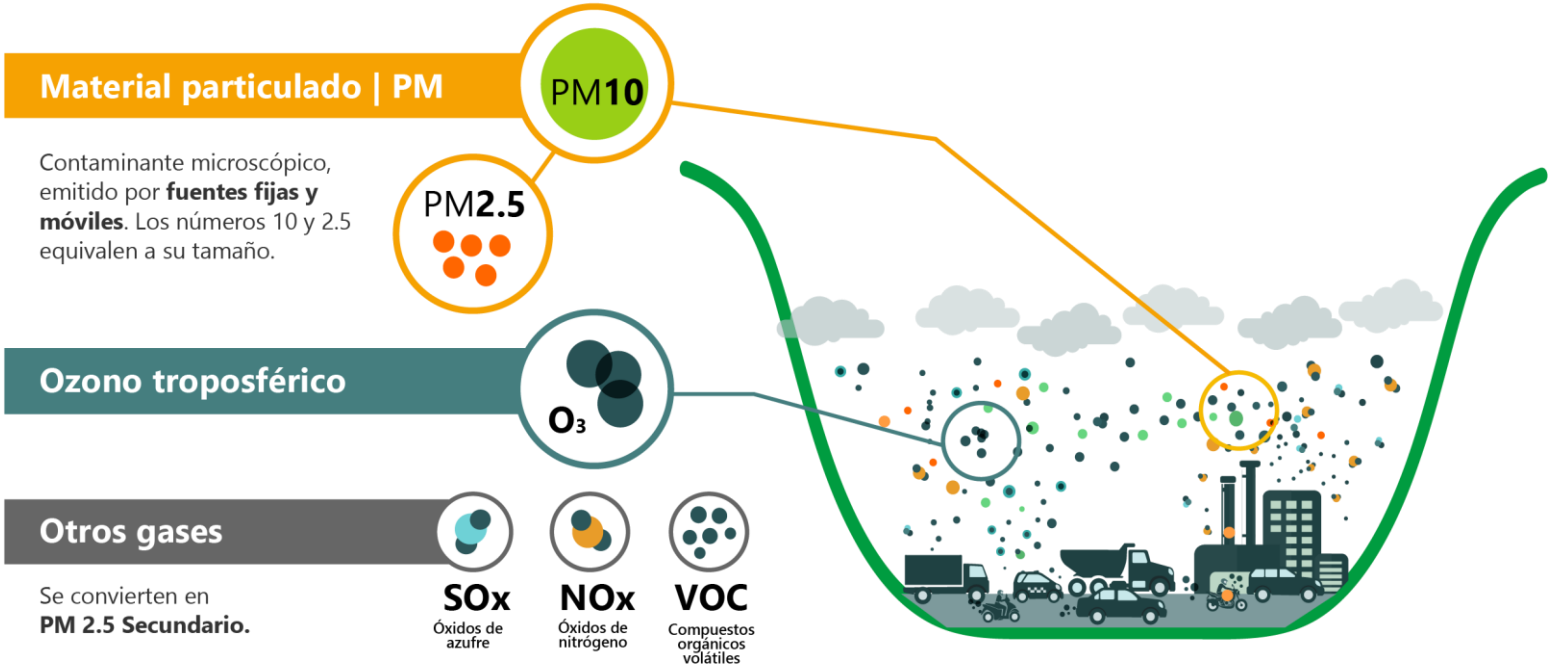



*Figura 2. Contaminantes principales del aire*

*(Fuente: Área Metropolitana del Valle de Aburrá, https://www.metropol.gov.co/ambiental/calidad-del-aire/Paginas/Generalidades/Principales-contaminantes.aspx)*

# JUSTIFICACIÓN

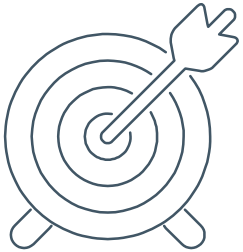
- **Causas**

  - Crecimiento demográfico
  - Uso de vehículos a combustión
  - Fábricas que emiten agentes contaminantes
  - Carencia infraestructura industrial y de transporte planeada

# OBJETIVOS

## General

- Predecir con 48 horas de anticipación los niveles de material particulado PM2.5 de la ciudad de Medellín y su área metropolitana, mediante la creación de modelos con redes neuronales recurrentes.



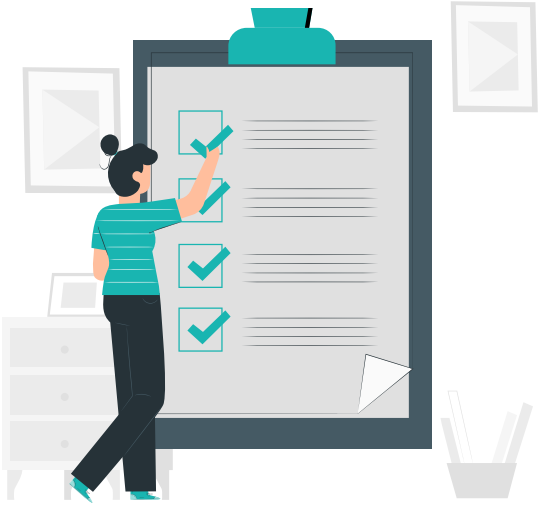
# OBJETIVOS

## Específicos

- Conocer los estudios realizados y los antecedentes sobre de calidad del aire de la ciudad de Medellín.
- Correlacionar los datos meteorológicos con los datos de contaminantes registrados por las estaciones de medición.
- Aplicar técnicas de limpieza y preprocesamiento al conjunto de datos para su posterior modelamiento.
- Crear modelos de redes neuronales recurrentes RNN, GRU, LSTM y un modelo híbrido LSTM + MLP para la predicción de series temporales.
- Evaluar los resultados de las predicciones generados por los modelos resultantes

# METODOLOGÍA

1. Estado del arte
2. Recolección y comprensión de los datos
3. Preparación de los datos
4. Modelado de los datos
5. Evaluación



# 2.
# DISEÑO E IMPLEMENTACIÓN

# ORIGEN DE LOS DATOS

- **Sistema de Alerta Temprana de Medellín y el Valle de Aburrá (SIATA)**

- Estaciones de monitoreo de automático
- Datos desde el 2015 al primer semestre del 2019
- Contaminantes del aire
- Variables meteorológicas
- Selección de tres estaciones de monitoreo
- 54 archivos CSV por estación (162 en total)

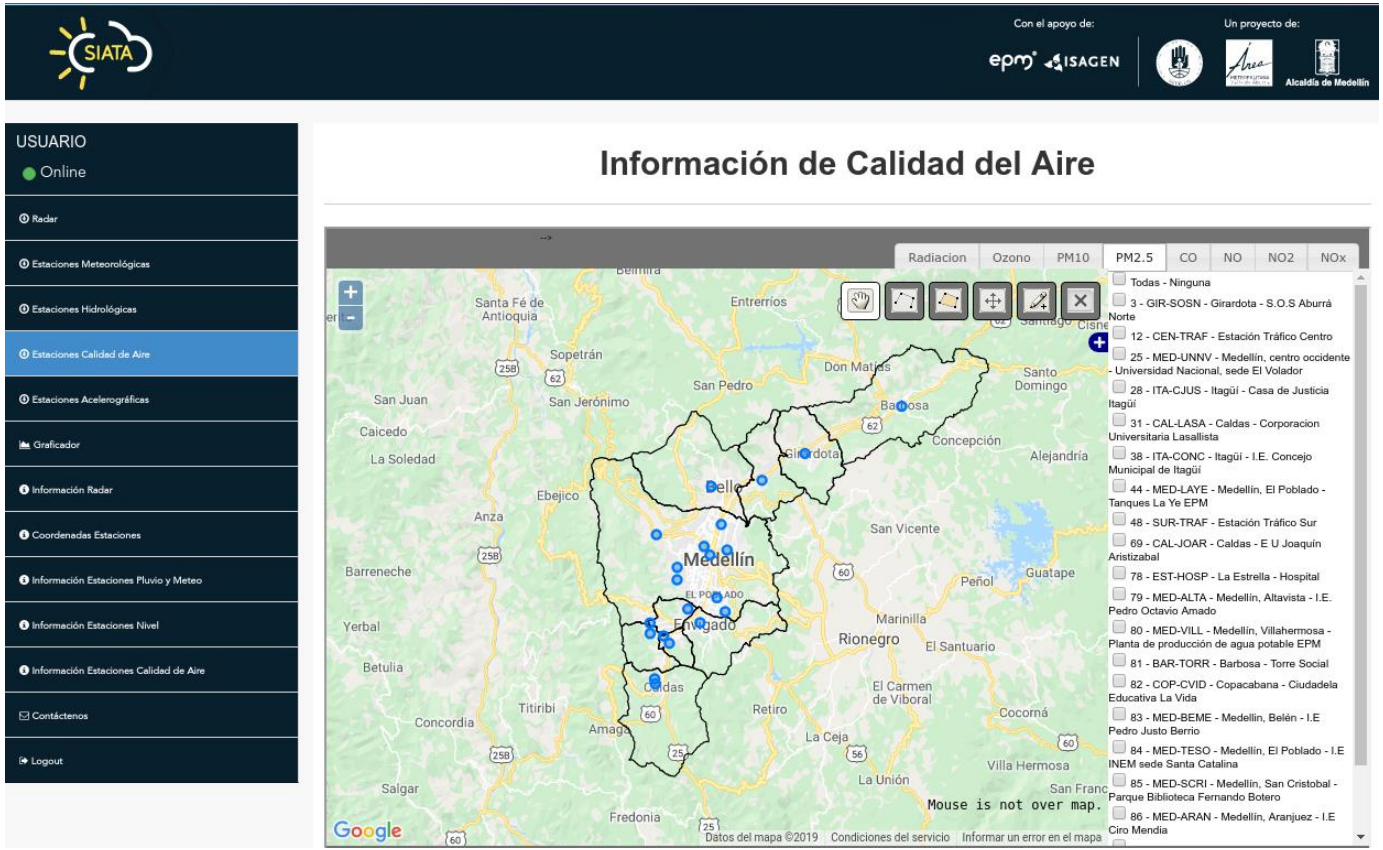



*Figura 3. Red de monitoreo de la calidad del aire vista desde del portal SIATA*

# ORIGEN DE LOS DATOS

| Código | Estación | Lugar |
|---|---|---|
| 25 | Universidad Nacional (UNAL), sede El Volador | Medellín |
| 31 | Corporación Universitaria Lasallista | Caldas |
| 28 | Casa de la justicia | Itagüí |

*Tabla 1. Estaciones seleccionadas en el trabajo*

# EXPLORACIÓN DE LOS DATOS

## Resumen de los datos

- Serie temporal por periodos horarios
- Resumen estadístico descriptivo
- 39408 registros por estación

## Calidad de los datos

- Identificación de *outliers*
- Datos erróneos o faltantes

| Contaminantes | Nomenclatura | Unidades |
|---|---|---|
| Material Particulado menor a 1 micra | pm1 | $\mu g/m^3$ |
| Material Particulado menor a 2.5 micras | pm25 | $\mu g/m^3$ |
| Material Particulado menor a 10 micras | pm10 | $\mu g/m^3$ |
| Ozono | ozono | ppb |
| Monóxido de Carbono | co | ppm |
| Monóxido de Nitrógeno | no | ppb |
| Dióxido de Nitrógeno | no2 | ppb |
| Óxidos de Nitrógeno | nox | ppb |
| Dióxido de Azufre | so2 | ppb |

*Tabla 2. Variables de contaminantes*

| Variable | Nomenclatura | Unidades |
|---|---|---|
| Temperatura del aire a 10 metros | taire10 ssr | °C |
| Humedad del aire a 10 metros | haire10 ssr | % |
| Velocidad del viento | vviento sssr | m/s |
| Dirección del viento | dviento ssr | grados |
| Radiación | rglobal ssr | $W/m^2$ |
| Precipitación | pliquida ssr | mm |
| Presión | p ssr | mmHg |

*Tabla 3. Variables meteorológicas*

# PREPARACIÓN DE LOS DATOS

## Limpieza y selección

- Selección de las variables meteorológicas y el contaminante PM2.5
- Eliminación variables cuyos valores nulos o erróneos superaban el 10% respecto del total.

## Imputación

- Método híbrido (Interpolación linear e imputación por la mediana)

## Transformación

- Normalización por la diferencia

## Análisis estadístico

- Resumen estadístico descriptivo
- Uso de gráficos estadísticos
- Agrupación por periodos de tiempo
- Correlaciones

# PREPARACIÓN DE LOS DATOS

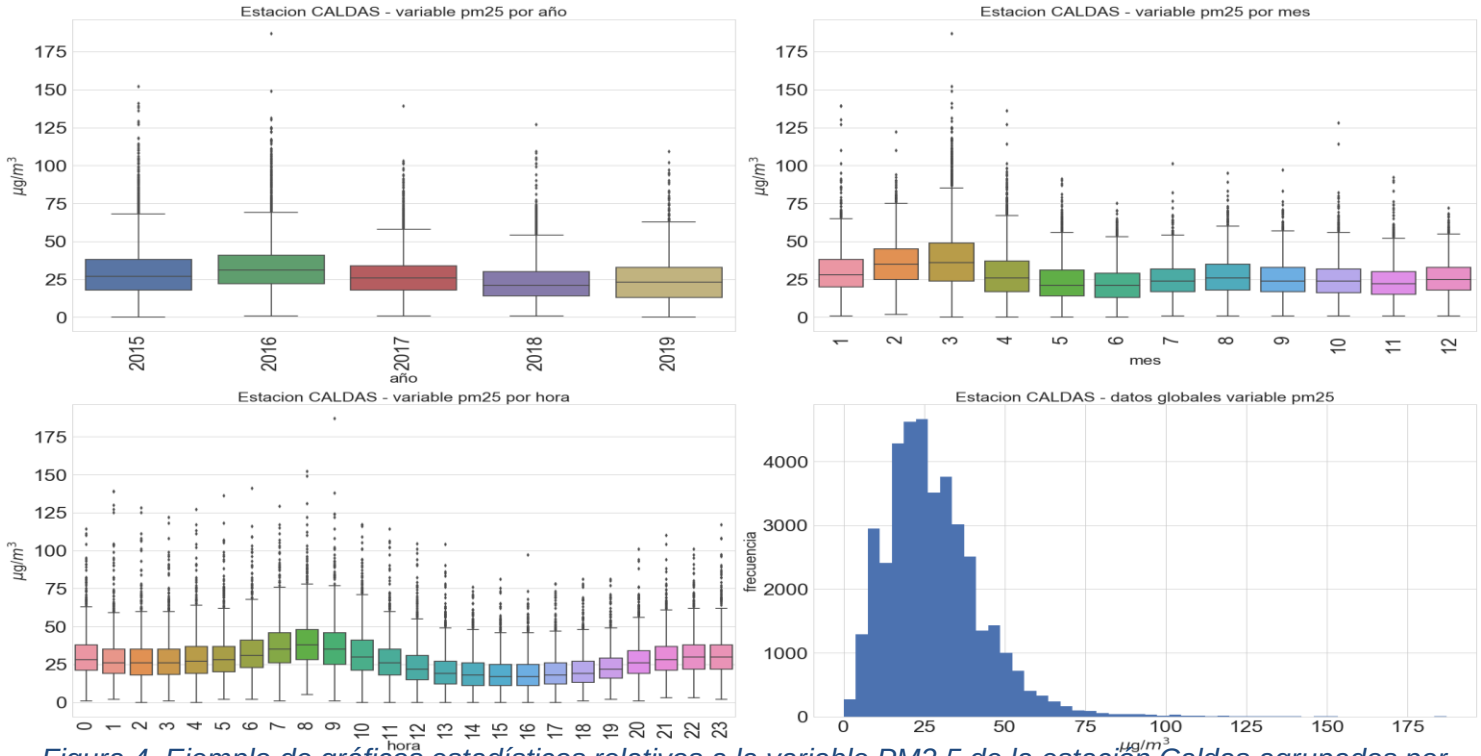



*Figura 4. Ejemplo de gráficos estadísticos relativos a la variable PM2.5 de la estación Caldas agrupados por periodos*

# PREPARACIÓN DE LOS DATOS

- **Descomposición de la variable PM2.5**
  - Tendencia
  - Estacionalidad
  - Existencia de comportamientos cíclicos
  - Periodos de 24h

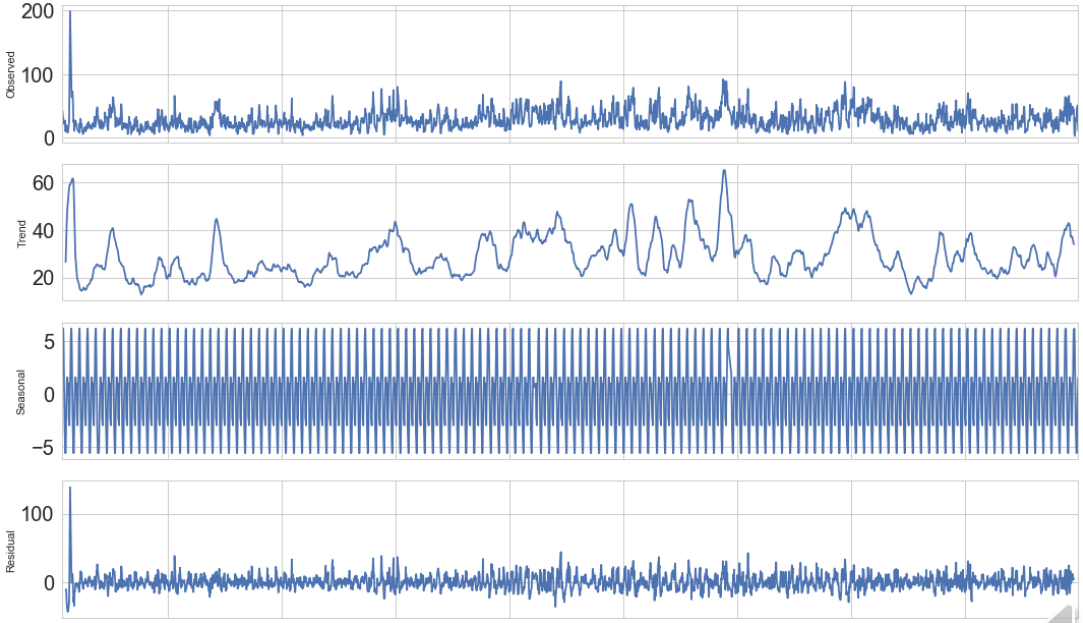



*Figura 5. Descomposición de la serie temporal relacionada al PM2.5*

# MODELAMIENTO

## Técnicas

- RNN
- GRU
- LSTM
- LSTM+MLP

## Hiperparámetros

- Grid Search
- Épocas (20 y 50)
- Batch (32, 64, y 128)
- Neuronas (16, 32 y 64)
- Función de activación (relu y tanh)

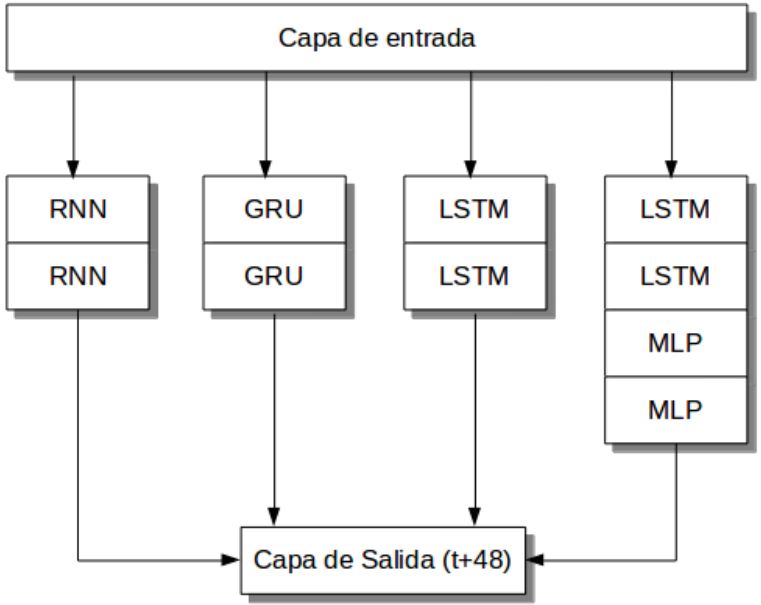



*Figura 6. Esquema de los cuatro modelos elegidos para las predicciones de PM2.5*

# MODELAMIENTO

## Conjunto de datos

- Ajuste de datos para aprendizaje supervisado
- 90% de datos de entrenamiento
- 10 % datos pruebas

## Herramientas

- Keras
- Sckit-learn
- Pandas
- Numpy

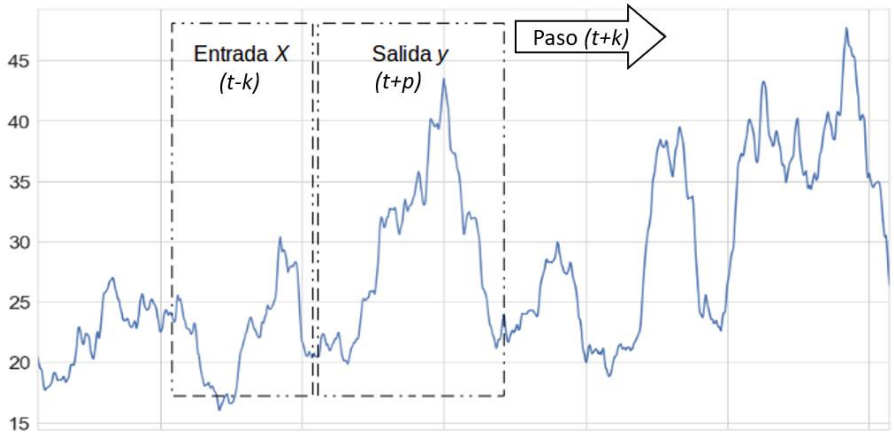



*Figura 7. Ejemplo de creación de datos para modelos supervisados a partir de una serie temporal mediante la ventana deslizante*

# 3. EVALUACIÓN

# RESULTADOS

- **Métrica de evaluación**
  - RMSE

- **Datos de entrenamiento**
  - PM2.5 y variables meteorológicas
  - Sólo PM2.5

- **Modelos**
  - Selección de los mejores modelos para cada técnica

# RESULTADOS

## Modelos con variables meteorológicas

- Modelos con celdas GRU tuvieron mejor desempeño en las predicciones de las estaciones UNAL y Caldas
- El modelo híbrido tuvo mejores predicciones en la estación Itagüí
- Las diferencias del RMSE entre los modelos GRU, LSTM e híbrido son pequeñas
- Los modelos RNN tuvieron un menor desempeño

| Estación Universidad Nacional (UNAL), sede El Volador | | | | | | |
|---|---|---|---|---|---|---|
| Modelo | Épocas | Batch | Activación | Neuronas | RMSE Entrenamiento | RMSE Predicciones |
| RNN | 50 | 32 | tanh | 16 | 0.0565 | 0.0545 |
| GRU | 50 | 32 | tanh | 32 | **0.0535** | **0.0478** |
| LSTM | 50 | 32 | tanh | 32 | 0.0541 | 0.0490 |
| LSTM+MLP | 50 | 64 | tanh | 64 | 0.0539 | 0.0487 |

*Tabla 4. Mejores modelos para la estación UNAL con datos meteorológicos*

| Estación Corporación Universitaria Lasallista de Caldas | | | | | | |
|---|---|---|---|---|---|---|
| Modelo | Épocas | Batch | Activación | Neuronas | RMSE Entrenamiento | RMSE Predicciones |
| RNN | 50 | 32 | Tanh | 16 | 0.0562 | 0.0584 |
| GRU | 50 | 32 | Tanh | 16 | **0.0529** | **0.0530** |
| LSTM | 50 | 32 | Tanh | 32 | 0.0534 | 0.0554 |
| LSTM+MLP | 50 | 32 | Tanh | 16 | 0.0535 | 0.0556 |

*Tabla 5. Mejores modelos para la estación Caldas con datos meteorológicos*

| Estación Casa de la Justicia de Itagüí | | | | | | |
|---|---|---|---|---|---|---|
| Modelo | Épocas | Batch | Activación | Neuronas | RMSE Entrenamiento | RMSE Predicciones |
| RNN | 50 | 32 | Tanh | 16 | 0.0597 | 0.0556 |
| GRU | 50 | 32 | Tanh | 32 | 0.0568 | 0.0517 |
| LSTM | 50 | 64 | Tanh | 64 | **0.0563** | 0.0518 |
| LSTM+MLP | 50 | 64 | Tanh | 64 | 0.0570 | **0.0515** |

*Tabla 6. Mejores modelos para la estación Itagüí con datos meteorológicos*

# RESULTADOS

## Modelos sin variables meteorológicas

- Modelos LSTM tuvieron mejor desempeño en las predicciones
- Para todos los casos los mejores modelos LSTM e híbridos usaron el mayor numero de neuronas.

| Estación Universidad Nacional (UNAL), sede El Volador | | | | | | |
|---|---|---|---|---|---|---|
| Modelo | Épocas | Batch | Activación | Neuronas | RMSE Entrenamiento | RMSE Predicciones |
| RNN | 50 | 32 | tanh | 16 | 0.0541 | 0.0507 |
| GRU | 50 | 64 | tanh | 32 | 0.0527 | 0.0473 |
| LSTM | 50 | 64 | tanh | 64 | **0.0525** | **0.0468** |
| LSTM+MLP | 50 | 32 | tanh | 64 | 0.0528 | 0.0472 |

*Tabla 7. Mejores modelos para la estación UNAL sin datos meteorológicos*

| Estación Corporación Universitaria Lasallista de Caldas | | | | | | |
|---|---|---|---|---|---|---|
| Modelo | Épocas | Batch | Activación | Neuronas | RMSE Entrenamiento | RMSE Predicciones |
| RNN | 50 | 32 | tanh | 16 | 0.0534 | 0.0544 |
| GRU | 50 | 32 | tanh | 64 | 0.0522 | 0.0533 |
| LSTM | 50 | 32 | tanh | 64 | **0.0521** | **0.0523** |
| LSTM+MLP | 50 | 32 | tanh | 64 | 0.0526 | 0.0544 |

*Tabla 8. Mejores modelos para la estación Caldas sin datos meteorológicos*

| Estación Casa de la Justicia de Itagüí | | | | | | |
|---|---|---|---|---|---|---|
| Modelo | Épocas | Batch | Activación | Neuronas | RMSE Entrenamiento | RMSE Predicciones |
| RNN | 50 | 32 | tanh | 16 | 0.0577 | 0.0504 |
| GRU | 50 | 32 | tanh | 64 | 0.0556 | 0.0501 |
| LSTM | 50 | 32 | tanh | 64 | 0.0551 | **0.0500** |
| LSTM+MLP | 50 | 64 | tanh | 64 | **0.0547** | 0.0524 |

*Tabla 9. Mejores modelos para la estación Itagüí sin datos meteorológicos*

# RESULTADOS

- Los modelos LSTM obtuvieron mejor rendimiento sin variables meteorológicas respecto a los demás.
- Los modelos GRU fueron  más estables en las predicciones, y obtuvo un buen rendimiento con variables meteorológicas
- Los modelos RNN obtuvieron un menor desempeño en la mayoría de los casos.
- El modelo híbrido tuvo mejores predicciones en la estación Itagüí con datos meteorológicos.
- Las diferencias del RMSE entre los diferentes modelos fueron pequeñas.

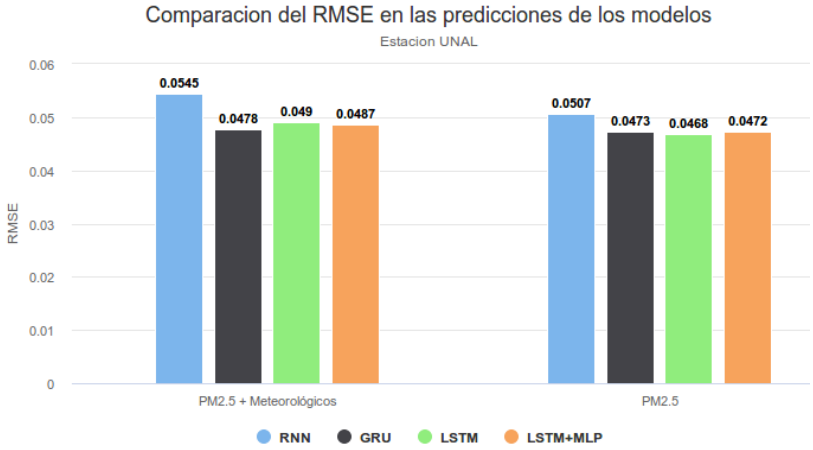



*Figura 8. Comparación de los resultados de las predicciones relacionados a la estación UNAL*

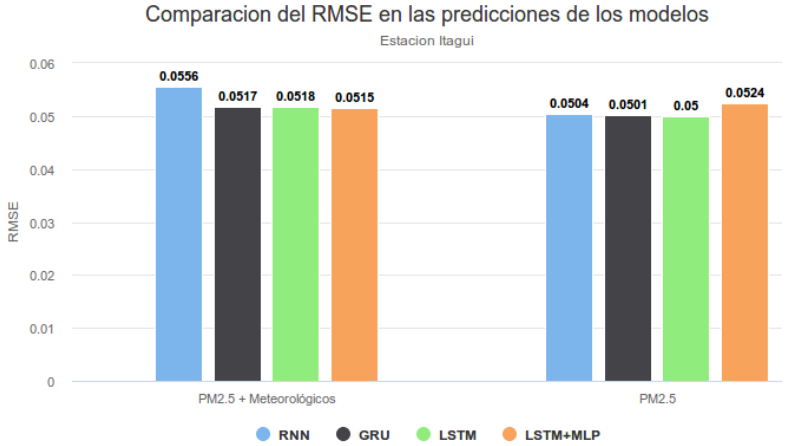



*Figura 10. Comparación de los resultados de las predicciones relacionados a la estación Itagüí*

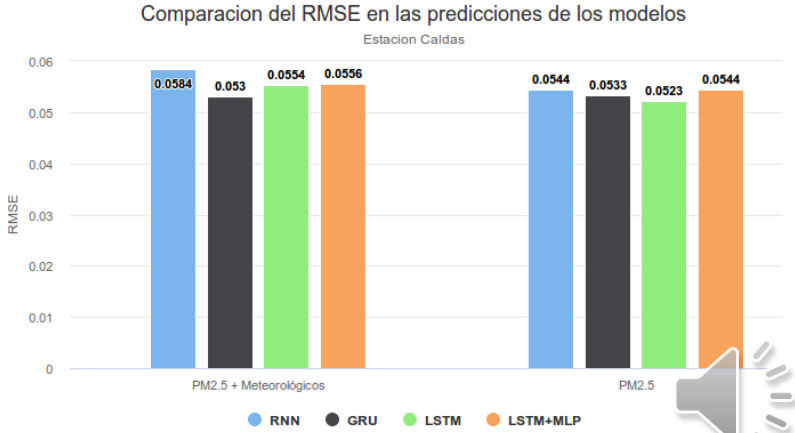



*Figura 9. Comparación de los resultados de las predicciones relacionados a la estación Caldas*

# RESULTADOS

## Mejor modelo estación UNAL

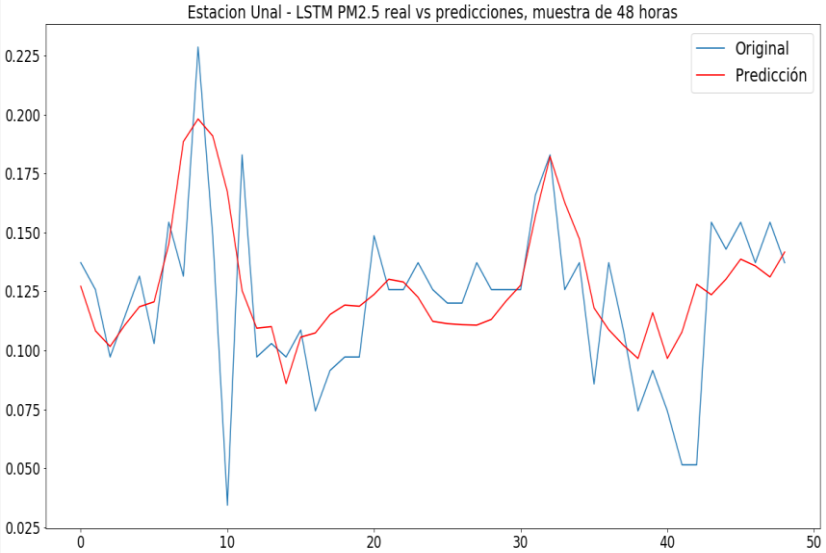



*Figura 11. Muestra de una predicción de PM2.5 de las próximas 48 horas usando el mejor modelo (LSTM) de la estación UNAL sin datos meteorológicos.*

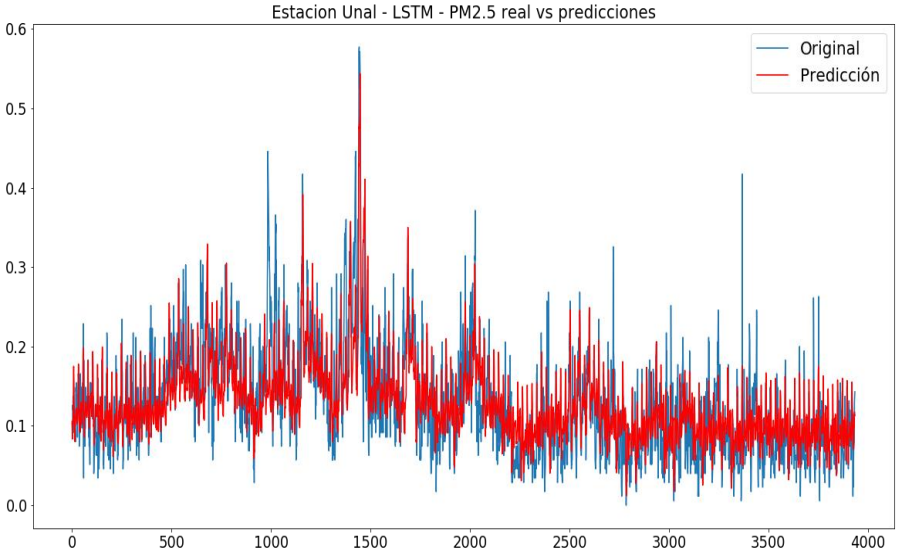



*Figura 12. Unión de todas las predicciones de PM2.5 del mejor modelo (LSTM) relacionado a la estación UNAL sin datos meteorológicos.*

# RESULTADOS

## Mejor modelo estación Caldas

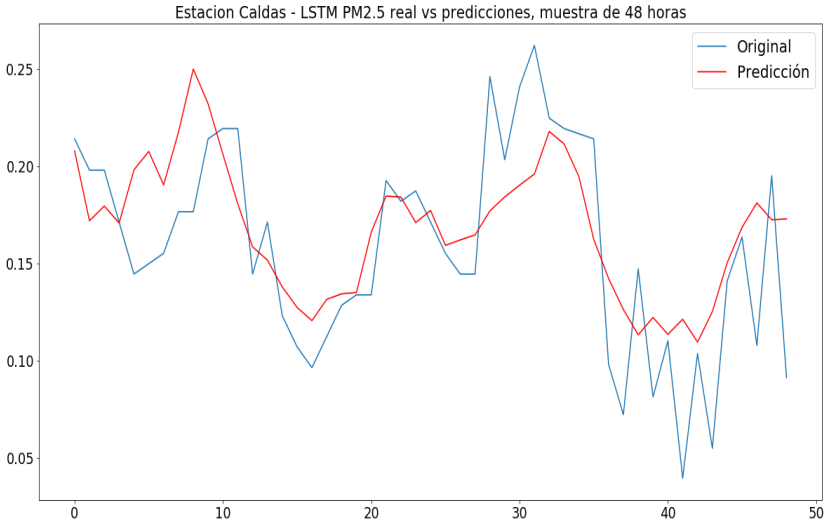



*Figura 13. Muestra de una predicción de PM2.5 de las próximas 48 horas usando el mejor modelo (LSTM) de la estación Caldas sin datos meteorológicos.*

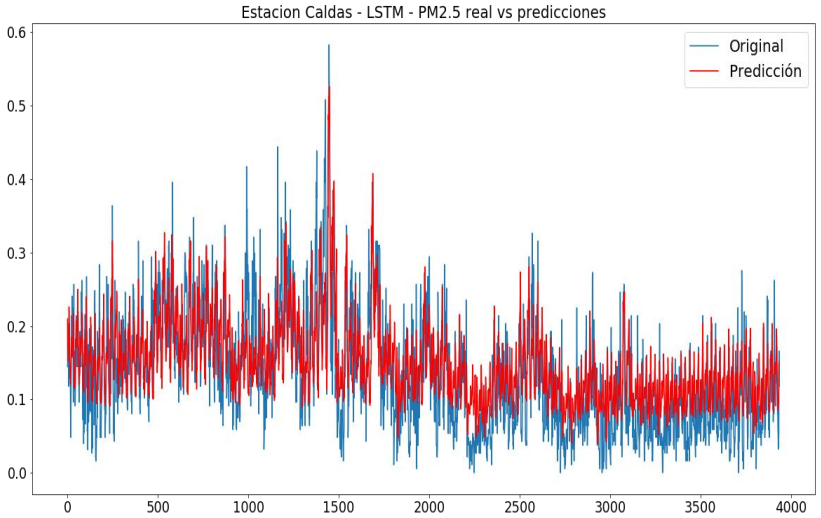



*Figura 14. Unión de todas las predicciones de PM2.5 del mejor modelo (LSTM) relacionado a la estación Caldas sin datos meteorológicos.*

# RESULTADOS

## Mejor modelo estación Itagüí

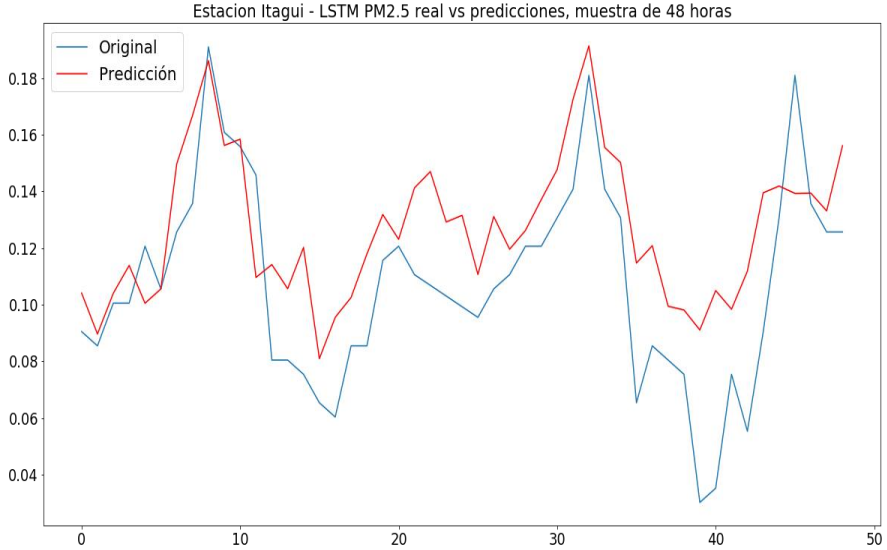



*Figura 15. Muestra de una predicción de PM2.5 de las próximas 48 horas usando el mejor modelo (LSTM) de la estación Itagüí sin datos meteorológicos.*

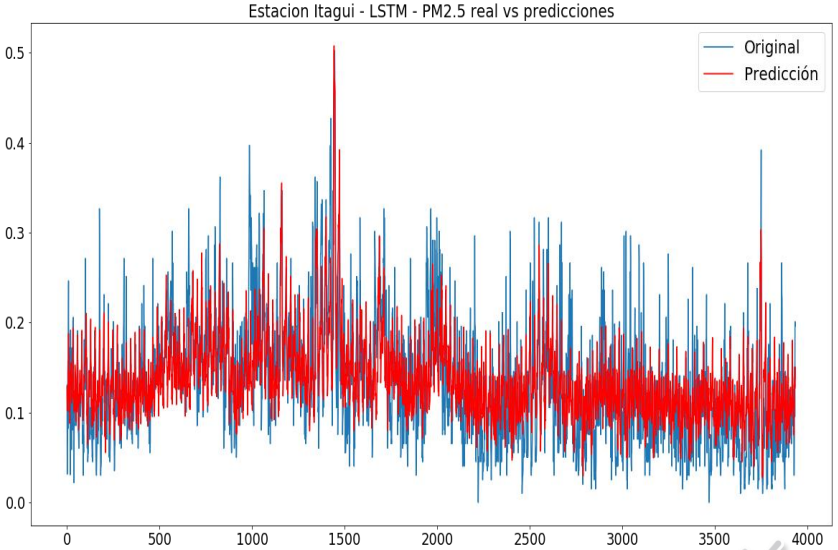



*Figura 16. Unión de todas las predicciones de PM2.5 del mejor modelo (LSTM) relacionado a la estación Itagüí sin datos meteorológicos.*

# CONCLUSIONES

- Los modelos tuvieron un buen ajuste en las predicciones del PM2.5 de cada estación.
- Los modelos LSTM entrenados sin variables meteorológicas predominaron con mejor rendimiento en las predicciones para las tres estaciones.
- En términos generales el modelo con menor rendimiento en las predicciones fue el RNN.
- Los modelos con celdas GRU tuvieron resultados mas estables en las predicciones, presentando una buena relación entre el desempeño y consumo de recursos.
- El entrenamiento de los modelos híbridos (LSTM+MLP) demanda mas tiempo y recursos computacionales.

# CONCLUSIONES

## Objetivos y metodología

- Se lograron los objetivos propuestos en el presente trabajo.
- La planificación y metodología fue adecuada.
- Se seleccionaron las estaciones con suficientes datos y que estuviesen separadas en diferentes municipios del área metropolitana.
- Fue necesario utilizar una GPU para entrenar los modelos y combinar los parámetros con *Grid Search.*

# CONCLUSIONES

## Líneas de trabajo futuro

- Explorar diversas fuentes de datos meteorológicos y de contaminantes para complementar los datos existentes.
- Expandir las predicciones hacia otro tipo de contaminantes del aire.
- Construir un sistema de predicciones programado que permita entrenarse en tiempo real, brindando predicciones a medida que reciba nuevos datos.
- Se podrían considerar modelos para detección y predicción de anomalías que detecten niveles nocivos de los contaminantes del aire.
- Usar  otro tipo de modelos que no requieran un alto volumen de datos de entrenamiento y extender el estudio a otras estaciones de medición.
- Integrar otras variables, como información del tráfico e imágenes satelitales.
- Explorar e implementar otros métodos de imputación para datos.

# GRACIAS