PROPUESTA DE TRABAJO FINAL o PROYECTO DE TESIS MAESTRÍA

La propuesta o proyecto debe contener el estado del arte del tema de investigación y la evidencia clara de la posibilidad de realizar una contribución a la solución de problema(s) complejo(s) en el área del programa.

TITULO: Una aproximación mediante técnicas de aprendizaje de máquinas para el pronóstico de la calidad del aire del Valle de Aburrá

DIRECTOR: María Constanza Torres Madroñero

PROGRAMA: Maestría en Ingeniería - Analítica

PERFIL: Profundización

AUTOR:

Manuel Alejandro De la Rosa Gómez

1234991606

aldela@unal.edu.co

3137026346

1. PLANTEAMIENTO DEL PROBLEMA

De acuerdo con la literatura consultada, se observa que, tanto a nivel de Colombia como a nivel mundial, la calidad del aire se modela principalmente mediante modelos numéricos, modelos estadísticos o modelos de aprendizaje de maquina (Machine Learning – ML) tradicionales.

Los modelos tradicionales, como el modelo atmosférico que emplea el Área Metropolitana del Valle de Aburrá (AMVA) [1], se basan en la simulación de las emisiones de material contaminante, las condiciones cambiantes de la atmósfera (velocidad del viento, precipitación, radiación) y las reacciones químicas del material contaminante con la atmósfera. Estos métodos son sensibles a las condiciones de borde a partir de las cuales se hace la simulación y, entre estaciones de medición de material particulado PM2.5, presentan un amplio rango de errores de predicción, con un MAPE (error absoluto medio porcentual) entre el 10% y el 100%, con respecto a los datos históricos obtenidos mediante medición del material particulado.

Por otro lado, se ha intentado predecir la calidad del aire con métodos estadísticos y modelos de ML tradicionales, los cuales se basan en datos históricos tanto de la concentración de material particulado como de variables meteorológicas, como se menciona en [2], [3] y [4]. Estos métodos incluyen: métodos autorregresivos como el ARIMA, regresión lineal múltiple, bosques aleatorios (Random Forest -RF) y regresores de máquinas de vectores de soportes (Support Vector Machine - SVM). De acuerdo con Zhang [5], estos métodos presentan dificultades en la modelación de las variaciones no lineales de las distintas variables. Para los casos aplicados en Colombia, se tienen valores de RMSE (error cuadrático medio) de

entre 6.50 y 15.33 µg/m3, para material particulado PM2.5. Cabe mencionar que también se han implementado modelos predictivos basados en aprendizaje profundo (Deep Learning - DL), los cuales exhiben valores de errores menores a los demás métodos, pero presentan altos requisitos computacionales y su interpretabilidad puede ser compleja, como menciona Zhang [5].

Este trabajo busca evaluar y comparar diversas técnicas incluyendo modelos estadísticos, modelos ML y DL, para plantear un modelo predictivo que minimice las métricas de error en la predicción de la concentración de material particulado PM2.5 para las distintas estaciones de medición del SIATA para el valle de Aburrá con datos capturados desde el 2015 tanto a escala horaria como diaria.

2. JUSTIFICACIÓN

Para la salud pública es de alta importancia minimizar la concentración de material particulado en el aire (en especial los materiales PM2.5 y PM10, y el ozono). De acuerdo con un estudio realizado por la Universidad de Antioquia, se estima que entre 2008 y 2019 se presentaron más de 70.000 muertes atribuibles a la contaminación por PM2.5 y ozono [6], con un promedio de 5909 muertes/año, presentando mayores frecuencias en los últimos tres años analizados (ver Figura 1). También se observó una especial vulnerabilidad en los adultos mayores de 64 años, cubriendo alrededor del 76% de las muertes, como se muestra en la Figura 2. Además, el mismo estudio menciona que se atribuyen costos de alrededor de $150 mil millones por año por muertes prematuras relacionadas con la contaminación del aire, lo cual se traduce en una pérdida de productividad durante el periodo analizado de alrededor de $1.8 billones (Tabla 1). Un modelo predictivo de la calidad del aire se muestra como una oportunidad para apoyar la toma de decisiones relacionadas con el trazado de metas y políticas de restricciones de emisiones de contaminantes con el fin de reducir la concentración de dichos contaminantes.

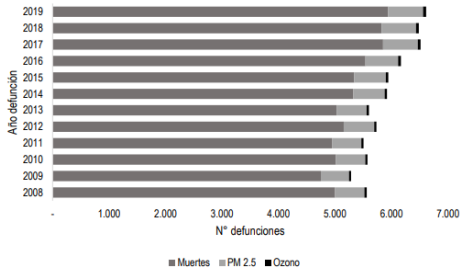


Figura 1. Distribución del número de defunciones relacionadas y atribuibles a la contaminación del aire por año entre 2008 y 2019. Tomada de [6].

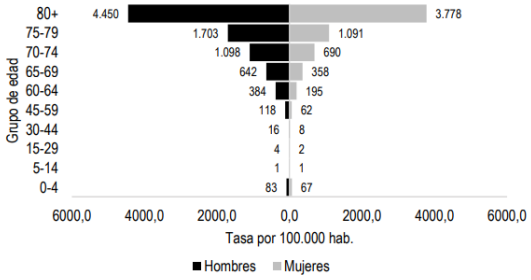


Figura 2. Distribución de las tasas de mortalidad por eventos relacionados con la contaminación del aire según grupo de edad y sexo entre 2008 y 2019. Tomada de [6].

Año D	Mujeres				Hombres				Total			
	PM 2.5	Ozono	Total	%	PM 2.5	Ozono	Total	%	PM 2.5	Ozono	Total	%
2008	56.2	4.5	60.7	7.8	75.5	5.3	80.8	7.9	131.7	9.9	141.6	7.9
2009	50.8	4.1	54.9	7.0	74.6	5.3	79.9	7.8	125.4	9.4	134.8	7.5
2010	55.8	4.6	60.4	7.7	75.2	4.9	80.1	7.8	131.0	9.5	140.5	7.8
2011	54.3	4.4	58.7	7.5	76.3	5.4	81.6	8.0	130.6	9.8	140.4	7.8
2012	58.6	4.6	63.1	8.1	76.6	4.9	81.6	8.0	135.2	9.5	144.7	8.0
2013	57.0	4.9	62.0	7.9	75.0	5.0	80.0	7.8	132.0	9.9	142.0	7.9
2014	60.5	5.0	65.5	8.4	79.2	5.0	84.2	8.2	139.7	10.0	149.7	8.3
2015	61.7	5.2	66.9	8.6	78.6	5.6	84.2	8.2	140.3	10.8	151.1	8.4
2016	62.9	5.5	68.4	8.8	82.3	6.1	88.4	8.7	145.2	11.6	156.8	8.7
2017	67.9	6.1	74.0	9.5	85.5	6.1	91.6	9.0	153.4	12.2	165.6	9.2
2018	66.7	5.9	72.6	9.3	85.9	5.8	91.7	9.0	152.6	11.7	164.4	9.1
2019	66.7	6.1	72.8	9.3	89.1	7.3	96.4	9.4	155.8	13.3	169.2	9.4
Total	719.0	61.0	780.0	43.3	953.9	66.7	1,020.6	56.7	1,672.9	127.7	1,800.6	100.0

Tabla 1. Costos de la mortalidad atribuible a la contaminación por pm2.5 y Ozono según sexo, AMVA 2008-2019 (Miles de millones de $ de 2019). Tomada de [6]

Por otro lado, es importante destacar el cumplimiento de las metas del índice de calidad del aire que se propuso en el *Plan Integral para la Gestión de la Calidad del Aire* (PIGECA) del Valle de Aburrá de 2017, en el cual se evidencia la necesidad de disminuir la concentración promedio anual de contaminantes PM2.5 hasta 23 µg/m3 y no exceder en periodos de 24 horas dicho valor más de 15 veces en estaciones poblaciones y 56 veces en estaciones de tráfico para 2030 [7]. Cabe decir que, en algunas estaciones, como la Centro-Tráfico, ubicada en Medellín, dicho indicador ha superado el umbral permitido en 2 de los últimos 5 años [8]. Si bien esta muestra una tendencia decreciente, como se muestra en la Figura 3, es importante hacer un seguimiento a su disminución para las distintas estaciones.

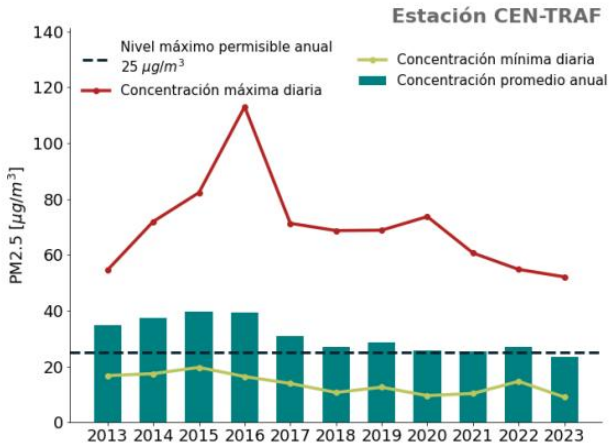


Figura 3. Variación de indicadores de concentración de contaminantes PM2.5 en la estación Centro-Tráfico entre 2013 y 2023. Tomado de [8].

3. OBJETIVOS

3.1. OBJETIVO GENERAL: Desarrollar un modelo predictivo que permita pronosticar la concentración de material particulado a partir de las mediciones de las estaciones del SIATA y la aplicación de técnicas de aprendizaje de máquina.

3.2. OBJETIVOS ESPECÍFICOS:

1. Identificar variables climáticas, geográficas o topográficas relacionadas con la concentración de material particulado PM2.5, que permitan completar la información proporcionada por las estaciones SIATA, para el desarrollo del modelo predictivo.
2. Comparar modelos estadísticos, de aprendizaje de máquinas (Machine Learning) y aprendizaje profundo, para la predicción de material particulado PM2.5 desde los datos seleccionados.
3. Proponer un modelo predictivo óptimo para la predicción de material particulado para el Valle de Aburrá empleando datos del SIATA y de otras fuentes.

4. MARCO TEÓRICO

4.1. ANTECEDENTES

4.1.1. Modelos predictivos en el mundo

En general, los modelos predictivos de calidad del aire presentes en la literatura emplean uno de los siguientes enfoques:

- Modelos numéricos. Usualmente basados en modelos de transporte y transformación química.
- Modelos estadísticos. Por ejemplo, basados en modelos autorregresivos como SARIMA.
- Modelos de aprendizaje de maquina: Modelos como bosques aleatorios (RF), máquinas de soporte vectorial (SVM), árboles de decisión.
- Modelos de aprendizaje profundo:
- Modelos híbridos: Usualmente combinando modelos estadísticos con aprendizaje profundo.

La Figura 4 y Tabla 2 presenta la línea temporal descrita por Zhang [5], en la cual se muestra la tendencia de emplear modelos de aprendizaje de maquina como regresión lineal múltiple, bosques aleatorios y máquinas de soporte vectorial a partir de 2018, y aprendizaje profundo a partir de 2019, con arquitecturas como *Gate Recurrent Units* (GRU), *Graph Neural Network* (GNN), *Long-Short Term Memory* (LSTM) y *Transformer*, mencionando también modelos híbridos que combinan distintas técnicas de aprendizaje profundo y enfoques tradicionales. El autor menciona la importancia de las variables de velocidad del viento, humedad, presión atmosférica, información del terreno y datos de tráfico. La Tabla 2 muestra los diferentes métodos que se han empleado, a distintas escalas de tiempo (predominantemente escala horaria y diaria), de los cuales destacan los casos de Feng et al. (2023) [9], que presenta el menor error a escala diaria (RMSE=0.739 μg/m3) y el de Gilik et al. (2022) [10], que presenta el menor error a escala horaria (RMSE=0.07 μg/m3).

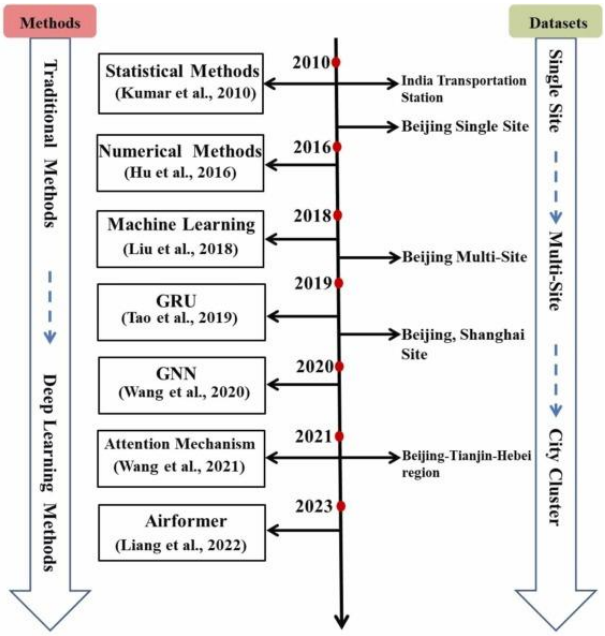


Figura 4. Diversos modelos predictivos para el pronóstico de la calidad del aire entre 2010 y 2023. Tomado de [5].

Año	Referencia	Área de estudio	Método	Desempeño
2017	Li et al.[11]	China	Geoi-DBN	MAE=8.54 RMSE=13.03 (h=24)
2018	Lin et al.[12]	Beijing, Los Angeles	GC-DCRNN	Beijing: MAE=36.62, RMSE=53.31 (h=24) Los Angeles: MAE=16.03, RMSE=21.08 (h=24)
2019	Tao et al.[13]	Beijing	CBGRU	MAE=10.47, RMSE=14.53 (h=1)
2019	Wu et al.[14]	North China	MSSTN	MAE=17.50, RMSE=29.87 (h=24)
2020	Wang et al.[15]	China	$PM_{2.5}$-GNN	MAE=15.91, RMSE=20.16 (h=72)
2020	Wang et al.[16]	Beijing	Attention-based seq2seq	MAE=16.11, RMSE=22.08 (h=24)
2021	Zhang et al.[17]	Beijing	VMDBiLSTM	MAE=7.12, RMSE=9.29 (h=1)
2021	Jin et al.[18]	Beijing	MTMC-NLSTM	MAE=0.88, RMSE=1.17 (h=1)
2021	Retta et al.[19]	Beijing	TCN	MAE=0.91, RMSE=2.18 (h=1)
2021	Han et al.[20]	Beijing, Shanghai	MasterGNN	Beijing: MAE=27.45, MAPE=0.548 (h=1) Shanghai: MAE=16.51, MAPE=26.5 (h=1)
2021	Huang et al.[21]	Beijing	SpAttRNN	MAE=57.36, RMSE=82.41, MAPE=0.698 (h=24)

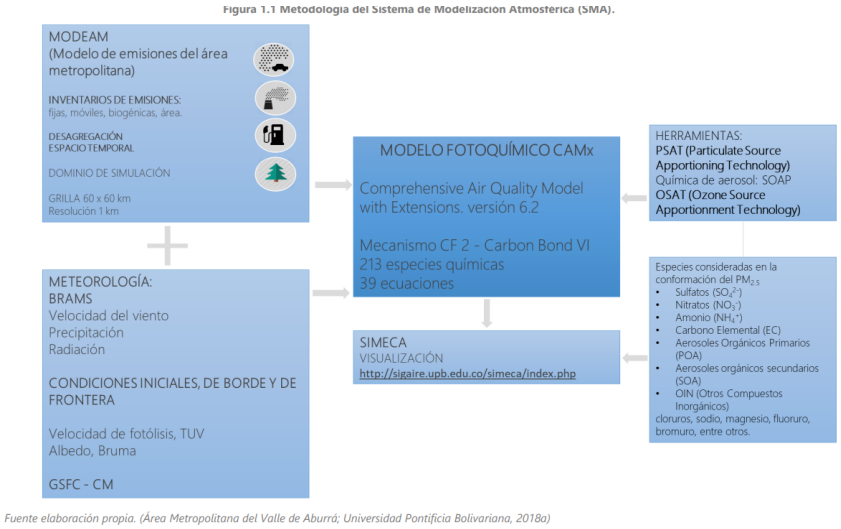
Año	Referencia	Área de estudio	Método	Desempeño
2021	Yeo et al.[22]	Seoul	CNN-GRU	RMSE=10.08 (h=1 año)
2021	Ge et al.[23]	Beijing	MST-GCN	MAE=18.44, RMSE=26.49 (h=24)
2021	Wang et al.[24]	Beijing, Tianjin	ATGCN	Beijing: MAE=11.25, RMSE=18.61 (h=1) Tianjin: MAE=14.27, RMSE=20.90 (h=1)
2021	Padhi et al.[25]	Beijing	BERT	RMSE=32.80 (h=1)
2022	Pruthi et al.[26]	Delhi	Nature-inspired deep learning	RMSE=10.80 (h=24)
2022	Xiao et al.[27]	China	DP-DDGCN	MAE=12.09, RMSE=15.19 (h=18)
2022	Gilik et al.[10]	Barcelona	CNN+LSTM	RMSE=0.07 (h=1)
2022	Faraji et al.[28]	Tehran	3DCNN-GRU	RMSE=15.21 (h=24)
2022	Huang et al.[29]	Beijing	EMD–IPSO–LSTM	MAE=4.02, RMSE=7.11, MAPE=8.07 (h=1)
2023	Fang et al.[30]	Beijing	IVLSTM-MCMR	MAE=31.31, RMSE=33.04 (h=24)
2023	Zhao et al.[31]	Shanghai	SAC-QBSO-BiLSTM	MAE=10.16, RMSE=13.23, MAPE=25.53 (h=1)
2023	Iskandaryan et al.[32]	Madrid	A3T-GCN	MAE=15.91, RMSE=19.85 (h=24)
2023	Zhang et al.[33]	Beijing	TDGTN	MAE=11.06, RMSE=18.51, MAPE=22.91 (h=1)
2023	Zhang et al.[34]	Beijing	Sparse attention-based transformer networks	MAE=11.13, RMSE=19.04 (h=1)
2023	Elbaz et al.[35]	Saudi Arabia	ResNet-ConvLSTM	MAPE=13.57(h=1)
2023	Feng et al.[9]	Yuzhong, China	EnAutoformer	MAE=0.538, RMSE=0.739 (h=24)
2023	Luo et al.[36]	Baoding, China	ARIMA-WOA-LSTM	MAE=5.572, RMSE=10.34 (h=1)
2023	Liang et al.[37]	China	Airformer	MAE=16.03, RMSE=32.36 (h=24)

Tabla 2. Resumen de modelos descritos por Zhang [5] y sus indicadores de desempeño. Entre paréntesis se menciona la escala de las mediciones.

4.1.2. Modelos predictivos aplicados a Medellín y otras ciudades colombianas

A nivel de Colombia, como se muestra en la Tabla 3, se destaca el uso de modelos de aprendizaje de maquina en la ciudad de Bogotá, como es el caso de las máquinas de vectores de soporte, empleadas por los autores Mogollón, Casallas y Vidal (2021) [38], en las cuales evidencian un RMSE de 9.30 µg/m3. Por otro lado, los autores Simanca y Blanco (2021) [2], emplearon bosques aleatorios, evidenciando un RMSE de 15.33 µg/m3. Por último, Celis y Casallas (2022) [3] entrenaron un modelo de red neuronal LSTM (memoria a corto-largo plazo), del cual se obtuvo un RMSE de 4.5 µg/m3, reduciendo significativamente el error con respecto a los modelos anteriores. En ambos casos, se emplearon como variables predictoras las concentraciones de PM10, O3, NO2 y CO. Adicionalmente, el modelo de Mogollón, Casallas y Vidal [38] tiene en cuenta las concentraciones de SO2, y variables meteorológicas tales como velocidad del viento, radiación solar, precipitación, temperatura, humedad y presión atmosférica.

Por otro lado, el Área Metropolitana del Valle de Aburrá en su PIGECA (Plan Integral de la Gestión del Aire del Valle de Aburra) implementa un modelo numérico para la predicción de la concentración de contaminantes en la atmósfera, para el cual en el caso de PM2.5 se obtuvo un RMSE de entre 1.96 y 25.98 µg/m3 para las diferentes estaciones de la región. Este modelo emplea diferentes etapas, dentro de las que incluye un modelo de emisiones (MODEAM), un modelo de variables meteorológicas (BRAMS) y un modelo de transporte y transformación química (CAMx) [1].



Fuente elaboración propia. (Área Metropolitana del Valle de Aburrá; Universidad Pontificia Bolivariana, 2018a)

Figura 5. Metodología del Sistema de Modelización Atmosférica del Valle de Aburrá. Tomada de [1].

Por último, se menciona el trabajo realizado por Santa, Vélez y Patiño (2023) [4], para la predicción de la concentración de PM2.5 en Medellín para la estación "Casa de justicia de Itagüí" usando datos de septiembre de 2016 a marzo de 2020, empleando un modelo ARIMA (con un RMSE de 6.50 µg/m3) y una red neuronal artificial con arquitectura de Perceptrón multicapa (con un RMSE de 6.90 µg/m3), en ambos casos obteniendo un RMSE menor al obtenido por el AMVA (RMSE=25.97 µg/m3). Para este trabajo se tuvo en cuenta las concentraciones de PM10, O3, NO2 y CO, así como temperatura, presión, velocidad y dirección del viento, radiación solar, precipitación y humedad.

Año	Referencia	Área de estudio	Método	Desempeño
2023	S. Santa, D. Velez, G. Patino [4]	Medellín	Comparación entre Modelo ARIMA y Red Neuronal Artificial (ANN)	**ARIMA:** RMSE: 6.50 µg/m3 **ANN:** RMSE: 6.90 µg/m3
2022	N. Celis, A. Casallas, E. A. López-Barrera, et al. [3]	Bogotá	Red neuronal convolucional de memoria a corto y largo plazo (1D-BD-LSTM-NN)	RMSE: 4.5 µg/m3
2021	F. Simanca, F. Blanco, et al. [2]	Bogotá	Random Forest	RMSE: 15.33 µg/m3
2021	C. Mogollón, A. Casallas, S. Vidal, et al. [38]	Bogotá	Máquina de vectores de soporte (SVM)	RMSE: 9.30 µg/m3 R=0.65
2020	J. Palacio, et al. [1]	Valle de Aburrá	Modelo MODEAM (fuentes fijas, móviles -LEAP- y fuentes de área) + BRAMS-CAMx (transporte y transformación química).	Para PM2.5: RMSE: 1.96-25.98 µg/m3 MAPE: 10%-100%

Tabla 3. Modelos predictivos para el pronóstico de calidad del aire en Colombia entre 2020 y 2024. Elaboración propia.

4.2. MARCO CONCEPTUAL

- **Pronóstico:** Pronosticar es realizar un enunciado sobre el valor futuro de una variable de interés, fundamentado ya sea por el análisis de datos históricos disponibles, por el juicio de expertos en el tema o por una combinación de ambas cosas [39].
- **Machine Learning (Aprendizaje de Maquinas):** Rama de la inteligencia artificial que involucra la creación de modelos mediante el entrenamiento de un algoritmo para hacer predicciones o tomar decisiones a partir de datos [40].
- **Aprendizaje supervisado:** Forma de aprendizaje de máquina que consiste en entrenar algoritmos para clasificar o predecir datos a partir de datos etiquetados o históricos [40].
- **Deep Learning (Aprendizaje profundo):** Forma de aprendizaje de máquina basada en las redes neuronales artificiales multicapa [40].
- **Red neuronal:** Una red neuronal es un procesador distribuido masivamente en paralelo compuesto por unidades de procesamiento simples que tiene una propensión natural a almacenar conocimientos experienciales y ponerlos a disposición para su uso. Se asemeja al cerebro en dos aspectos: La red adquiere conocimiento de su entorno a través de un proceso de aprendizaje y las fortalezas de las conexiones interneuronales, conocidas como pesos sinápticos, se utilizan para almacenar el conocimiento adquirido [41].
- **Red neuronal recurrente:** Es una arquitectura de red neuronal que permite transmitir la salida de las neuronas de una capa como entradas a las neuronas de una capa anterior [41].
- **LSTM:** Es una arquitectura de red neuronal recurrente que incluye células de memoria y puertas (como la puerta de olvido y la puerta de entrada). Esta estructura permite retener información por varios pasos de tiempo, solucionando el problema del desvanecimiento del gradiente [42].

- **RMSE:** Dado un conjunto de datos verdaderos $\{y_i\}$ y un conjunto de datos pronosticados $\{y_i'\}$, se define el RMSE o Raíz cuadrada del error cuadrático medio como:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2}$$

  Esta métrica de error posee la característica de tener las mismas unidades que la variable objetivo del pronóstico, lo cual permite facilita su interpretabilidad. Entre menor sea su valor, se considera que es mejor el desempeño del modelo predictivo [43].
- **MAE:** Dado un conjunto de datos verdaderos $\{y_i\}$ y un conjunto de datos pronosticados $\{y_i'\}$, se define el MAE o Error absoluto promedio como:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_i'|$$

  Esta métrica se expresa en las unidades de la variable objetivo al cuadrado. Entre menor sea su valor, se considera que es mejor el desempeño del modelo predictivo [43].
- **R^2:** Dado un conjunto de datos verdaderos $\{y_i\}$, el promedio de los datos verdaderos $\overline{y}$ y un conjunto de datos pronosticados $\{y_i'\}$, se define el coeficiente de determinación (R^2) como:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i')^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

  Esta métrica siempre se encuentra en el rango [0,1] y usualmente se usa para describir qué tan bien se ha ajustado el modelo de pronóstico. En general, se interpreta como la proporción de varianza explicada por el modelo predictivo (numerador) y la varianza total en la respuesta de los datos (denominador). Entre más cercano sea el valor del R^2 a 1, se considera mejor el desempeño del modelo de pronóstico [43].

5. METODOLOGÍA A UTILIZAR

5.1 BASE DE DATOS

Para la construcción del modelo predictivo se propone el uso de la base de datos de medición de la calidad del aire del SIATA, la cual se reporta a final de cada mes. Estos datos se descargan desde el geoportal del SIATA [14] y cuentan con 24 estaciones con datos de la concentración del material particulado PM2.5 para cada hora. En general, para cada estación se reporta la siguiente información:

- Fecha y hora [YYYY-MM-DD HH:MM:SS]
- Concentración de PM2.5 [µg/m3]
- Concentración de PM10 [µg/m3]
- Concentración de PM1 [µg/m3]
- Concentración de monóxido de nitrógeno (NO) [ppb]
- Concentración de óxidos de nitrógeno (NOx) [ppb]
- Concentración de Ozono (O3) [ppb]
- Concentración de monóxido de carbono (CO) [ppm]
- Concentración de dióxido de azufre (SO2) [ppb]

- Dirección del viento [grados]
- Velocidad del viento [m/s]
- Humedad del aire a 10 metros [%]
- Temperatura del aire a 10 metros[°C]
- Radiación [W/m2]
- Precipitación [mm]
- Presión [mmHg]

Cada dato de concentración de contaminantes y variables meteorológicas reporta un valor de calidad del dato, codificado como se muestra en la tabla 4:

Valor	Significado
1	Dato válido
-1	Dato válido del operador anterior
1.8 – 2.5	Dato cuestionable
2.6 – 3.9	Dato malo
>= 4.0	Dato faltante

Tabla 4. Codificación de calidad de los datos del SIATA. Tomado del geoportal del SIATA

Cabe mencionar que el SIATA también tiene como convención, cuando el equipo de medición está fuera de servicio, reportar el valor del dato como -9999 y la calidad como 1.

Por otro lado, es necesario aclarar que no todas las estaciones miden todas estas variables, por lo cual se propone aproximar dichas variables faltantes con las mediciones de estaciones cercanas.

Se propone establecer los siguientes conjuntos de datos:

- **Entrenamiento:** Datos del 01/01/2019 al 31/12/2021
- **Prueba:** Datos del 01/01/2022 al 31/12/2023
- **Validación:** Datos del 01/01/2024 al 31/12/2024

5.2 DISEÑO METODOLÓGICO

La Tabla 5 presenta el resumen del diseño metodológico para el desarrollo del proyecto. Se buscar desarrollar un modelo predictivo a partir de modelos de regresión temporal, el cual tendrá como variable objetivo la concentración de material particulado PM2.5.

Para este trabajo se usará el lenguaje de programación Python. Para las tareas relacionadas con la carga y procesamiento de datos se empleará la librería **pandas** (ver Tabla 5 – actividades 1.1 a 1.4)**.** Para el entrenamiento de modelos y optimización de hiperparámetros (las tareas 1.5 a 1.7 y todas las tareas relacionadas con el objetivo 2) se usarán las librerías **statsmodels** (para el modelo SARIMAX), **scikit-learn** (para el modelo de regresión lineal múltiple y Random Forest) y **Tensorflow** (para el modelo de perceptrón multi-capa, red neuronal LSTM y red neuronal con arquitectura Transformer).

Se plantea para el primer objetivo específico, el cual corresponde a **Identificar variables climáticas, geográficas o topográficas relacionadas con la concentración de material particulado PM2.5, que permitan completar la información proporcionada por las estaciones SIATA**, las siguientes tareas: la descarga, consolidación y almacenamiento de los datos relacionados con la concentración de material particulado PM2.5 del SIATA y otras fuentes (tarea 1.1); el tratamiento de datos faltantes (tarea 1.2), el tratamiento de datos atípicos (tarea 1.3), la normalización de los datos (tarea 1.4), la preparación y entrenamiento de un primer modelo predictivo tipo Random Forest con todas las variables recopiladas (tarea 1.5); a partir de este modelo se calculará el puntaje de importancia de cada variable (tarea 1.6) y se elegirán las variables que expliquen el 90% de la varianza de los datos (tarea 1.7). Se tendrá como resultado un dataset procesado con únicamente las variables explicativas relevantes.

Para el segundo objetivo específico, el cual consiste en **Comparar modelos estadísticos, de aprendizaje de máquinas (Machine Learning) y aprendizaje profundo, para la predicción de material particulado PM2.5 desde los datos seleccionados,** se proponen las siguientes tareas: particionamiento de los datos en conjunto de entrenamiento, conjunto de prueba y conjunto de validación (tarea 2.1); la preparación y entrenamiento de distintos modelos predictivos (tarea 2.2): SARIMAX, Regresión lineal múltiple, Random Forest, Perceptrón multi-capa, Red neuronal LSTM, Red neuronal con arquitectura Transformer; para cada modelo se llevará a cabo una optimización de hiperparámetros (tarea 2.3) y se calcularán las métricas de desempeño (tarea 2.4) a partir de R2 y el RMSE. Como resultado se tendrá un reporte con el desempeño de cada modelo, que será el insumo principal para la elección del modelo óptimo.

Para el tercer objetivo, el cual es **Proponer un modelo predictivo óptimo para la predicción de material particulado para el Valle de Aburrá empleando datos del SIATA y de otras fuentes**, se llevará a cabo la elección del modelo óptimo (tarea 3.1), se validará el modelo con el conjunto de validación (tarea 3.2) y se presentarán los resultados del trabajo de comparación de modelos y validación del modelo óptimo (tarea 3.3). Como resultado se obtendrá el reporte de resultados de este trabajo final.

5.3 DISEÑO LÓGICO

Objetivo específico	Actividades	Resultado
1. Identificar variables climáticas, geográficas o topográficas relacionadas con la concentración de material particulado PM2.5, que permitan completar la información proporcionada por las estaciones SIATA, para el desarrollo del modelo predictivo.	1.1. Obtención de datos 1.2. Tratamiento de valores faltantes 1.3. Tratamiento de valores atípicos 1.4. Normalización 1.5. Preparación y entrenamiento de primer modelo de predicción con Random Forest. 1.6. Cálculo de puntaje de importancia de variables, a partir del modelo Random Forest 1.7. Elección de las variables que expliquen el 90% de la varianza de los datos	Dataset limpio con únicamente las variables más correlacionadas con la concentración de material particulado PM2.5
2. Comparar modelos estadísticos, de aprendizaje de máquinas (Machine Learning) y aprendizaje profundo, para la predicción de material particulado PM2.5 desde los datos seleccionados.	2.1. Particionamiento de datos en conjunto de entrenamiento, pruebas y validación 2.2. Entrenamiento de modelos con dataset procesado: 2.2.1. SARIMAX 2.2.2. Regresión lineal múltiple 2.2.3. Random Forest 2.2.4. Perceptrón multi-capa 2.2.5. Red neuronal LSTM 2.2.6. Red neuronal con arquitectura Transformer 2.3. Optimización de hiperparámetros 2.4. Cálculo de métricas de desempeño: 2.4.1. RMSE 2.4.2. R2	Reporte con el desempeño de cada modelo de aprendizaje de máquina
3. Proponer un modelo predictivo óptimo para la predicción de material particulado para el Valle de Aburrá empleando datos del SIATA y de otras fuentes.	3.1. Elección de modelo óptimo 3.2. Validación del modelo óptimo 3.3. Presentación de resultados	Modelo predictivo de concentración de material particulado PM2.5

Tabla 5. Diseño lógico para el desarrollo del proyecto.

6. ALCANCES DEL TRABAJO

Como se muestra en la Tabla 5, se espera al finalizar el trabajo:

- Identificar diferentes fuentes de información, además de SIATA, que proporcionen datos relevantes para el pronóstico del material particulado
- Entender el desempeño de diferentes técnicas desde métodos estadísticos, modelos de aprendizaje de maquina y aprendizaje profundo para el pronóstico de material particulado
- Obtener un modelo predictivo optimo que permita la estimación del material particulado en el Valle de Aburrá.

7. PLAN DE TEMAS Y CRONOGRAMA

Actividad	Mes 1	Mes 2	Mes 3	Mes 4	Mes 5	Mes 6	Mes 7	Mes 8
1.1								
1.2								
1.3								
1.4								
1.5								
1.6								
1.7								
2.1								
2.2								
2.3								
2.4								
3.1								
3.2								
3.3								

8. BIBLIOGRAFÍA Y FUENTES DE INFORMACIÓN

[1] Área Metropolitana del Valle de Aburrá (2020). Plan Integral para la gestión de la calidad del aire – PIGECA 2017-2030. Alcance 2: Asesoría, transferencia de conocimientos científicos, complementación y seguimiento a las acciones asociadas a la gobernanza del PIGECA. 2.5. Simulaciones de calidad del aire. Tomado de: https://www.metropol.gov.co/ambiental/calidad-del-aire/Biblioteca-aire/Estudios-calidad-del-aire/Modelizacion-Calidad-del-Aire.pdf

[2] Blanco, F., Lemus, J., Cerón, W., Carreño H, P., Rozo, J., Guerrero, L., & Simanca, F. (2021). Air Quality Index Prediction Model for the City of Bogotá, D.C. Retrieved from https://www.researchgate.net/publication/361262864_Air_Quality_Index_Prediction_Model_for_the_City_of_Bogota_DC.

[3] Celis, N., Casallas, A., López-Barrera, E. A., Martínez, H., Peña Rincón, C. A., Arenas, R., & Ferro, C. (2022). Design of an early alert system for PM2.5 through a stochastic method and machine learning models. Environmental Science & Policy, 127, 241-252. https://doi.org/10.1016/j.envsci.2021.10.030

[4] Santa, S., Velez, D., & Patino, G. (2023). Analysis and prediction of PM2.5 in Medellin based on seasonal time series. In 2023 IEEE Colombian Caribbean Conference (C3), 1-6. Barranquilla, Colombia. https://doi.org/10.1109/C358072.2023.10436185

[5] Zhang, Z., Zhang, S., Chen, C., & Yuan, J. (2024). A systematic survey of air quality prediction based on deep learning. Alexandria Engineering Journal, 93, 128-141. https://doi.org/10.1016/j.aej.2024.03.031

[6] Área Metropolitana del Valle de Aburrá (2021). Mortalidad atribuible a PM2.5 y ozono en los municipios del Valle de Aburrá y sus costos económicos, 2008-2019. Medellín, Colombia. Tomado de: https://www.metropol.gov.co/ambiental/calidad-del-aire/Biblioteca-aire/Salud-publica/Mortalidad-Atribuible-Contaminacion-2008-2019.pdf.

[7] Área Metropolitana del Valle de Aburrá (2022). Informe Final: Valorar los impactos económicos y sociales de medidas establecidas en el Plan Integral de Gestión de la calidad del aire – PIGECA y las implicaciones que tiene para un territorio la incorporación de escenarios de riesgo por contaminación atmosférica. Medellín, Colombia. Tomado de https://www.metropol.gov.co/ambiental/calidad-del-aire/Biblioteca-aire/Estudios-calidad-del-aire/Informe-Final-Evaluacion-Economica-Social-PIGECA.pdf.

[8] Área Metropolitana del Valle de Aburrá (2024). Informe Anual de Calidad del Aire 2023. Enero 31 de 2024. Medellín, Colombia. https://www.metropol.gov.co/ambiental/calidad-del-aire/informes_red_calidaddeaire/Informe-Anual-Aire-2023.pdf .

[9] Feng et al. (2023). A hybrid deep learning framework for air quality prediction with spatial autocorrelation during the COVID-19 pandemic. Sci Rep 13, 1015. https://doi.org/10.1038/s41598-023-28287-8.

[10] Gilik et al. (2022). Air quality prediction using CNN+LSTM-based hybrid deep learning architecture. Environ Sci Pollut Res 29, 11920–11938. https://doi.org/10.1007/s11356-021-16227-w.

[11] Li, T., Shen, H., Yuan, Q., Zhang, X., & Zhang, L. (2017). Estimating ground-level PM2.5 by fusing satellite and station observations: A geo-intelligent deep learning approach. Geophysical Research Letters, 44, 11,985–11,993. https://doi.org/10.1002/2017GL075710.

[12] Lin, Y., Mago, N., Gao, Y., Li, Y., Chiang, Y., Shahabi, C., & Ambite, J. (2018). Exploiting spatiotemporal patterns for accurate air quality forecasting using deep learning. In Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '18), 359–368. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3274895.3274907

[13] Tao, Q., Liu, F., Li, Y., & Sidorov, D. (2018). Air Pollution Forecasting Using a Deep Learning Model Based on 1D Convnets and Bidirectional GRU. IEEE Access, 7, 76690-76698. https://doi.org/10.1109/ACCESS.2019.2921578

[14] Wu, Z., Wang, Y., & Zhang, L. (2019). MSSTN: Multi-Scale Spatial Temporal Network for Air Pollution Prediction. In 2019 IEEE International Conference on Big Data (Big Data), 1547-1556. Los Angeles, CA, USA. https://doi.org/10.1109/BigData47090.2019.9005574

[15] Wang, S., Li, Y., Zhang, J., Meng, Q., Meng, L., & Gao, F. (2020). PM2.5-GNN: A Domain Knowledge Enhanced Graph Neural Network For PM2.5 Forecasting. In Proceedings of the 28th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '20), 163–166. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3397536.3422208

[16] Wang, H., Li, X., Wang, D., Zhao, J., He, H., & Peng, Z. (2020). Regional prediction of ground-level ozone using a hybrid sequence-to-sequence deep learning approach. Journal of Cleaner Production, 253, 119841. https://doi.org/10.1016/j.jclepro.2019.119841

[17] Zhang, Z., Zeng, Y. & Yan, K (2021). A hybrid deep learning technology for $PM_{2.5}$ air quality forecasting. *Environ Sci Pollut Res* **28**, 39409–39422. https://doi.org/10.1007/s11356-021-12657-8

[18] Jin, N., Zeng, Y., Yan, K., & Ji, Z. (2021). Multivariate Air Quality Forecasting With Nested Long Short Term Memory Neural Network. IEEE Transactions on Industrial Informatics, 17(12), 8514-8522. https://doi.org/10.1109/TII.2021.3065425

[19] Retta, S., Yarramsetti, P., Kethavath, S. (2021). Comprehensive Analysis of Deep Learning Approaches for PM2.5 Forecasting. In: Chaki, N., Pejas, J., Devarakonda, N., Rao Kovvur, R.M. (eds) Proceedings of International Conference on Computational Intelligence and Data Engineering. Lecture Notes on Data Engineering and Communications Technologies, vol 56. Springer, Singapore. https://doi.org/10.1007/978-981-15-8767-2_27

[20] Han, J., Liu, H., Zhu, H., Xiong, H., & Dou, D. (2021). Joint Air Quality and Weather Prediction Based on Multi-Adversarial Spatiotemporal Networks. Proceedings of the AAAI Conference on Artificial Intelligence, 35(5), 4081-4089. https://doi.org/10.1609/aaai.v35i5.16529

[21] Huang, Y., Ying, J., & Tseng, V. (2021). Spatio-attention embedded recurrent neural network for air quality prediction. Knowledge-Based Systems, 233, 107416. https://doi.org/10.1016/j.knosys.2021.107416

[22] Yeo, I., Choi, Y., Lops, Y. et al (2021). Efficient PM2.5 forecasting using geographical correlation based on integrated deep learning algorithms. Neural Comput & Applic 33, 15073–15089. https://doi.org/10.1007/s00521-021-06082-8

[23] Ge, L., Wu, K., Zeng, Y. *et al* (2021). Multi-scale spatiotemporal graph convolution network for air quality prediction. *Appl Intell* **51**, 3491–3505. https://doi.org/10.1007/s10489-020-02054-y

[24] Wang, C., Zhu, Y., Zang, T., Liu, H., & Yu, J. (2021). Modeling Inter-station Relationships with Attentive Temporal Graph Convolutional Network for Air Quality Prediction. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM '21), 616–634. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3437963.3441731

[25] Padhi, I., et al. (2021). Tabular Transformers for Modeling Multivariate Time Series. In ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 3565-3569. Toronto, ON, Canada. https://doi.org/10.1109/ICASSP39728.2021.9414142

[26] Pruthi, D., & Liu, Y. (2022). Low-cost nature-inspired deep learning system for PM2.5 forecast over Delhi, India. Environment International, 166, 107373. https://doi.org/10.1016/j.envint.2022.107373

[27] Xiao, X., Jin, Z., Wang, S., Xu, J., Peng, Z., Wang, R., Shao, W., & Hui, Y. (2022). A dual-path dynamic directed graph convolutional network for air quality prediction. Science of The Total Environment, 827, 154298. https://doi.org/10.1016/j.scitotenv.2022.154298

[28] Faraji, M., Nadi, S., Ghaffarpasand, O., Homayoni, S., & Downey, K. (2022). An integrated 3D CNN-GRU deep learning method for short-term prediction of PM2.5 concentration in urban environment. Science of The Total Environment, 834, 155324. https://doi.org/10.1016/j.scitotenv.2022.155324

[29] Huang, Y., Yu, J., Dai, X., Huang, Z., & Li, Y. (2022). Air-Quality Prediction Based on the EMD–IPSO–LSTM Combination Model. Sustainability, 14(9), 4889. https://doi.org/10.3390/su14094889

[30] Fang, W., Zhu, R., & Lin, J. C.-W. (2023). An air quality prediction model based on improved Vanilla LSTM with multichannel input and multiroute output. *Expert Systems with Applications, 211,* 118422. https://doi.org/10.1016/j.eswa.2022.118422

[31] Zhao, Z., Wu, J., Cai, F. *et al* (2023). A hybrid deep learning framework for air quality prediction with spatial autocorrelation during the COVID-19 pandemic. *Sci Rep* **13**, 1015. https://doi.org/10.1038/s41598-023-28287-8

[32] Iskandaryan, D., Ramos, F., & Trilles, S. (2023). Graph Neural Network for Air Quality Prediction: A Case Study in Madrid. IEEE Access, 11, 2729-2742. https://doi.org/10.1109/ACCESS.2023.3234214

[33] Zhen, Z., Shiqing, Z., Xiaoming, Z., Linjian, C., & Jun, Y. (2022). Temporal Difference-Based Graph Transformer Networks For Air Quality PM2.5 Prediction: A Case Study in China. Frontiers in Environmental Science, 10. https://doi.org/10.3389/fenvs.2022.924986

[34] Zhang, Z., & Zhang, S. (2023). Modeling air quality PM2.5 forecasting using deep sparse attention-based transformer networks. International Journal of Environmental Science and Technology, 20, 13535–13550. https://doi.org/10.1007/s13762-023-04900-1

[35] Elbaz, K., Hoteit, I., Shaban, W. M., & Shen, S.-L. (2023). Spatiotemporal air quality forecasting and health risk assessment over smart city of NEOM. Chemosphere, 313, 137636. https://doi.org/10.1016/j.chemosphere.2022.137636

[36] Luo, J., & Gong, Y. (2023). Air pollutant prediction based on ARIMA-WOA-LSTM model. Atmospheric Pollution Research, 14(6), 101761. https://doi.org/10.1016/j.apr.2023.101761

[37] Liang, Y., Xia, Y., Ke, S., Wang, Y., Wen, Q., Zhang, J., Zheng, Y., & Zimmermann, R. (2023). AirFormer: Predicting Nationwide Air Quality in China with Transformers. Proceedings of the AAAI Conference on Artificial Intelligence, 37(12), 14329-14337. https://doi.org/10.1609/aaai.v37i12.26676

[38] Mogollón-Sotelo, C., Casallas, A., Vidal, S. et al. (2021). A support vector machine model to forecast ground-level PM2.5 in a highly populated city with a complex terrain. Air Qual Atmos Health 14, 399–409. https://doi.org/10.1007/s11869-020-00945-0.

[39] Gallegos., M. E. (2013). Métodos de pronósticos para negocios. Editorial digital Tecnológico de Monterrey. Instituto Tecnológico y de Estudios Superiores de Monterrey, México. Obtenido de: http://prod77ms.itesm.mx/podcast/EDTM/P196.pdf

[40] IBM. (n.d.). What is artificial intelligence (AI)? Retrieved from https://www.ibm.com/topics/artificial-intelligence.

[41] Haykin, Simon (2009). Neural Networks and Learning Machines. Third Edition. Pearson. Prentice Hall.

[42] Patterson, J., Gibson, A. (2017) Deep Learning. A practitioner's approach. First Edition. O'Reilly.

[43] James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An Introduction to Statistical Learning: with Applications in Python.

Fecha de presentación: 10/02/2025

Nota: El profesor responsable debe ingresar en el SIA la calificación del proyecto o la propuesta.