**Logotipo

Descripción generada automáticamente**

**Pronóstico del aumento de ingresos en agrupaciones de hospitales por diagnósticos de enfermedades respiratorias basado en las condiciones metereológicas y la calidad del aire del Valle de Aburrá**

Kenneth David Leonel Triana

Juan José Naranjo Velásquez

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Universidad de Antioquia  
Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2023

| **Cita** | (Leonel Triana & Naranjo Velásquez, 2024) |
| --- | --- |
| **Referencia**  **Estilo APA 7 (2020)** | Leonel Triana, K. L., & Naranjo Velásquez, J. J. (2024). *Pronóstico del aumento de ingresos en agrupaciones de hospitales por diagnósticos de enfermedades respiratorias basado en las condiciones meteorológicas y la calidad del aire del Valle de Aburrá***.** Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia. |

**** 

Especialización en Analítica y Ciencia de Datos, CohorteX.

Centro de Investigación Ambientales y de Ingeniería (CIA).

|  | Diagrama  Descripción generada automáticamente con confianza media |
| --- | --- |

Centro de Documentación Ingeniería (CENDOI)

**Repositorio Institucional:** http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano: Julio Cesar Saldarriaga Molina

Jefe departamento: Diego José Luis Botia Valderrama

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

**Siglas, acrónimos y abreviaturas**

**SIATA** Sistema de alerta temprana de Medellín y del Valle del Aburrá

**RIPS** [Sistema de Información de Prestaciones de Salud](https://www.minsalud.gov.co/proteccionsocial/paginas/rips.aspx)

# Resumen

Colombia es el décimo primer país de la región que presenta mayor mortalidad por enfermedades respiratorias, lo que representa la cuarta causa de muerte en el país (Ministerio de Salud y Protección Social, 2022). Estas cifras pueden darse en ocasiones donde las condiciones climáticas y ambientales son dañinas para la salud (Hospital internacional de Colombia, 2024). Dadas estas cifras alarmantes, es necesario tener en cuenta la cantidad de pacientes que pueden ingresar a los hospitales en ocasiones donde existan afectaciones en las condiciones ambientales, por ello, con el fin de alertar a los establecimientos de salud acerca de la cantidad de pacientes con diagnósticos asociados a enfermedades respiratorias que ingresarán en ciertos sectores o comunas de Medellín, presentamos un modelo alimentado por variables climáticas e ingresos en hospitales, con datos proporcionados por el Sistema de Alerta Temprana de Medellín y el Valle de Aburrá (SIATA) y el Sistema de Información de Prestaciones de Salud (RIPS) en los años de 2020 a 2024. La creación del modelo predictivo es a base de la metodología CRISP-DM, lo que permite realizar el proceso ETL en los datos para ajustar el modelo de Random Forests a un R2 de 0.6. De esta forma, podemos estimar la cantidad de asistencias proyectadas en ciertos sectores de la ciudad que hayan tenido malas condiciones de calidad del aire.

**https://github.com/kennethLeonel/seminario-Udea**

Incluye al final de dicha página la dirección de los repositorios GitHub.

*Palabras clave*: machine learning, hospitales, enfermedades respiratorias, calidad del aire, SIATA, RIPS, Medellín.

# Abstract

Colombia is the eleventh country in the region with the highest mortality from respiratory diseases, which represents the fourth cause of death in the country (Ministerio de Salud y Protección Social, 2022). These figures can occur on occasions where climatic and environmental conditions are harmful to health (Hospital internacional de Colombia, 2024). Given these alarming figures, it is necessary to take into account the number of patients who may enter hospitals on occasions where environmental conditions are affected, therefore, in order to alert health establishments about the number of patients with diagnoses associated with respiratory diseases that will enter certain sectors or communes of Medellín, we present a model powered by climatic variables and hospital admissions, with data provided by the Sistema de Alerta Temprana de Medellín y el Valle de Aburrá (SIATA) and the Sistema de Información de Prestaciones de Salud (RIPS) in the years from 2020 to 2024. The creation of the predictive model is based on the CRISP-DM methodology, which allows the ETL process to be carried out on the data to adjust the Random Forests model to a R2 of 0.6. In this way, we can estimate the amount of assistance projected in certain sectors of the city that have had poor air quality conditions.

*Keywords***:**machine learning, hospitals, respiratory diseases, air quality, SIATA, RIPS, Medellín.

# Descripción del problema

El aire que respiramos se compone en su mayor proporción por nitrógeno y oxígeno (Torres Jaramillo, 2021), sin embargo, según la zona geográfica y las condiciones climáticas, esta proporción puede verse afectada con partículas inhalables finas que causan efectos negativos en la función pulmonar (Agencia de Protección Ambiental de Estados Unidos, 2024), por ello, desde nuestro punto de vista, es un fenómeno que no podemos evitar (dada la distribución territorial de Medellín), pero que se puede monitorear y medir su impacto para anticiparse a complicaciones posteriores. De acuerdo con lo anterior, con el fin de generar alertas por crecimiento de cantidades de pacientes en hospitales por contaminación en el ambiente, se propone una estructuración y modelación de datos climáticos, ambientales y médicos, cuya fuente de información es el Sistema de Alerta Temprana de Medellín y el Valle de Aburrá y el [Sistema de Información de Prestaciones de Salud](https://www.minsalud.gov.co/proteccionsocial/paginas/rips.aspx) (RIPS) , donde concatenados, relacionan variables meteorológicas como temperatura, presión, humedad, vientos, entre otras, junto a registros de pacientes ingresados en hospitales por diagnósticos asociados a deficiencias respiratorias. La precisión del modelo se verifica bajo el R2 y el error cuadrático medio (MSE) en diferentes momentos del periodo de 2020 a 2024, identificando comunas y sectores de Medellín con el fin de conocer el comportamiento del aumento de pacientes por problemas respiratorios. De esta forma, se procura que los hospitales tengan la información de este incremento mediante un mapa de calor para que se preparen ante estos eventos.

## Problema de negocio

La calidad del aire ha sido un tema de abundante estudio en el Valle del Aburrá, debido a la distribución geográfica del espacio y que al estar rodeado por montañas genera una estrechez que impide el flujo libre de gases y partículas en el aire (Área Metropolitana Valle del Aburrá, 2019), por ende, estas partículas se concentran en el centro de esta misma, consecuentemente, es un efecto de convivencia que debe atacarse desde diferentes medios. Es por ello que desde finales de los años 90 se creó el Sistema de Alertas Tempranas SIATA como un proyecto para la gestión ambiental y de riesgos del Área Metropolitana del Valle de Aburrá, realizando monitoreos en tiempo real con modelamientos hidrológicos y meteorológicos para pronosticar ocurrencias de fenómenos naturales que puedan generar riesgos a la población (Universidad de EAFIT, 2020).

Según la Organización Mundial de la Salud (OMS), “Colombia es el décimo primer país de la región que presenta mayor mortalidad por enfermedades respiratorias, por lo que representa la cuarta causa de muerte en el país”, esta medida genera alertamiento, como en 2019 donde estas enfermedades causaron 534.242 defunciones, con una tasa de 35.8 defunciones por 100,000 habitantes, y en 2020 con 22.9 defunciones por cada 100,000 habitantes (Ministerio de Salud y Protección Social, 2022). Adicionalmente, según el Instituto Nacional de Salud (INS) se atribuyen 15,681 muertes asociadas a la mala calidad del aire, lo que radica en 8 mil muertes anuales, que bajo una vista financiera abarca costos de 12.2 billones de pesos, equivalente al 1.5% del PIB. Por ello, la subdirección de Salud ambiental busca formas de tener una gestión de la calidad del aire del país orientadas a la disminución de la exposición y los factores de riesgo (Minsalud, 2021).

Se debe agregar que estas enfermedades respiratorias son silenciosas y en algunos casos no tiene cura como el EPOC (BBC News Mundo, 2017). En relación con lo anterior, bajo nuestra opinión, la visualización del ser humano no permite detectar en qué ocasiones hay una mala calidad del aire, por ello es posible inhalar componentes peligrosos para los pulmones que contribuyan a desarrollar alguna de estas enfermedades, y poco a poco, de manera discreta, se desarrolle estas sin conocimiento de aparición en el cuerpo, lo que hace que sea tarde para su prevención y tratamiento. Por estos motivos es requerido anticiparse al caso de alcanzar una sobrecapacidad de los establecimientos de salud en épocas donde la calidad del aire afecte a la población.

## Aproximación desde la analítica de datos

Como equipo de trabajo usaremos durante el transcurso del desarrollo la metodología CRISP-DM, una técnica reconocida ante la extracción de información valiosa a partir de volúmenes de datos, basándose en las etapas de entendimiento de negocio, entendimiento de datos, preparación de los datos, modelamiento y evaluación o validación desde el negocio para posterior despliegue del modelo. Esta forma de trabajo se toma con el fin de mitigar los inconvenientes que tengamos frente al desarrollo analítico y poder concluir los objetivos planteados ante la acción de las circunstancias encontradas, para ello, esta propuesta de solución inicia desde el entendimiento de la toma de datos con los puntos de monitoreo (entre automáticos y manuales) que usan la información de ceilómetros y radiómetros de la entidad SIATA, que se encuentran distribuidos a lo largo de todas las zonas del valle de aburrá, estos son importantes porque son los encargados de tomar los datos meteorológicos y de la calidad del aire en diferentes medidas de control. A su vez, en esta solución se implementa la información del número de atenciones totales de pacientes por enfermedades relacionadas con los problemas respiratorios, los cuales tienen información encriptada para el cuidado de la información de los pacientes que han entrado a consulta por estos motivos. A ambas fuentes de información se realiza el respectivo análisis exploratorio para encontrar tipos de variables, tipos de datos, relación de las características con el fenómeno, entre otras. Dicho tratamiento con el fin de preparar los datos para una misma base de trabajo y tener registros únicos entre ambas fuentes de información, lo que nos permite manipularlos para el traslado en el análisis de modelado bajo Machine Learning y finalmente el despliegue del modelo mediante una herramienta visual o mapa de calor que permita observar los posibles puntos de la ciudad donde se tengan mayores concentraciones de pacientes con estas condiciones respiratorias. Por ello, el modelo tendrá como finalidad la estimación de cantidades de pacientes que ingresarán a los establecimientos médicos por condiciones médicas relacionadas con enfermedades respiratorias según las condiciones climáticas y ambientales de los puntos de senseo del SIATA. Esto permitirá observar los sitios de mayores cantidades de pacientes para alertar y preparar los centros médicos, de tal forma que se disminuya las tasas de mortalidad por estas causas al ser tratadas con prioridad.

## Origen de los datos

Los datos utilizados en el estudio se obtuvieron desde bases de datos libres presentadas por el Valle del Aburrá y Medellín, mediante el portal de descargas del Sistema de Alerta Temprana de Medellín y el Valle de Aburrá SIATA junto con información de ingresos en hospitales del RIPS Medellín, donde se seleccionaron datos meteorológicos, ambientales y médicos. Desde el año 2010 el SIATA ha desarrollado una estrategia regional desde el conocimiento científico, el desarrollo tecnológico y la innovación, para identificar y pronosticar la ocurrencia de fenómenos naturales y antrópicos que alteren las condiciones ambientales de la región, o que puedan generar riesgos a la población; esto a partir de monitoreo en tiempo real y modelación hidrológica y meteorológica ajustada al territorio (Area Metropolitana Valle de Aburrá, n.d.), bajo este recorrido de tiempo y reconocimiento del Área Metropolitana del Valle del Aburrá lo convierte una entidad confiable para la selección de datos. Adicionalmente, cuentan dentro de sus características con un parámetro de categorización que da evidencia de si el dato es confiable o erróneo, lo que valida el uso de datos para el proyecto. Dentro de estas características contamos con variables como temperatura, humedad, presión, vientos, precipitación, partículas de tamaño de 2.5 micras hasta 10 micras, partículas por billón de gases como ozono, dióxido de carbono, monóxido de carbono, azufre, entre otras. Por otro lado, RIPS Medellín es el Sistema de Información de Prestaciones de Salud, el cual recopila datos mínimos y básicos que el Sistema General de Seguridad Social en Salud requiere para los procesos de dirección, regulación y control, y como soporte de la venta de servicio, cuya denominación, estructura y características se ha unificado y estandarizado para todas las entidades, los datos presentados se basan en identificación del prestador del servicio de salud, del usuario que lo recibe, de la prestación del servicio y del motivo que originó su prestación: diagnóstico y causa externa (Ministerio de Salud y Protección Social - República de Colombia, n.d.). De esta forma, dentro de los campos de la tabla poseemos información de los pacientes encriptados, edad, tipo usuario, establecimiento médico de ingreso, diagnóstico asociado a ingreso, fecha de ingreso y fecha de salida de la IPS; bajo este respaldo seleccionamos la base como datos confiables y para uso dentro del proyecto. Por ende, la propuesta se enfoca en la relación de los datos meteorológicos, ambientales y médicos presentados por las entidades anteriores, usando la fecha de cada medición y toma de datos como llave de unión para la generación del dataset de alimento para el modelado.

## Métricas de desempeño

Se debe tener en cuenta que el proyecto se basa en la creación del modelo analitico que permita la estimación de los ingresos de pacientes en las IPS según las variables meteorológicas, por ello, la variable a predecir es un dato cuantitativo continuo, por lo que el modelo debe pertenecer a un esquema por Regresión. De esta manera, seleccionamos la raíz del error cuadrático medio (RMSE) porque proporciona una interpretación directa del error en las mismas unidades que la variable objetivo, lo que facilita identificar el impacto práctico de las desviaciones en las predicciones; además, usaremos el Coeficiente de Determinación R² que da una mejor vista porcentual de qué tan bien el modelo explica la variabilidad de los datos, ofreciendo una evaluación del ajuste general del modelo, siendo ideal para comunicar la eficacia del modelo a nivel técnico. Según los resultados obtenidos bajo las métricas del modelado, se procede a complementar con indicadores como la capacidad de planificación operativa, que evalúa si las predicciones permiten gestionar eficientemente recursos hospitalarios, la reducción de costos operativos derivados de optimizar la asignación de personal e insumos, y el impacto en la atención al paciente, al reducir tiempos de espera y prevenir saturaciones. Podemos decir que al integrar estas métricas buscaremos no sólo que el modelo sea preciso para la determinación de las capacidades de pacientes de una forma confiable, sino también útil para la toma de decisiones estratégicas en salud pública con tal de evitar generar esfuerzos humanos y recursos innecesarios.

# Objetivos

## Objetivo general

Identificar posibles escenarios en los que se aumenten las capacidades de atención a pacientes de los sectores o comunas de Medellín por diagnósticos de enfermedades respiratorias respecto a las condiciones meteorológicas y ambientales del Valle del Aburrá, con el fin de generar pronósticos prescriptivos y de alertamiento para la preparación de estos establecimientos ante situaciones de alta demanda por malas condiciones climáticas.

## Objetivos específicos

Extraer y transformar datos crudos del portal de descargas SIATA y base de datos de RIPS desde años 2020 a 2024 para la generación de un dataset con granularidad de registros diarios.

Realizar una exploración de datos bajo análisis descriptivo para identificar valores atípicos, distribuciones y correlaciones de las variables de entrada y salida.

Implementar modelos de machine learning con la finalidad de generar el pronóstico de la cantidad de pacientes que se tendrá en un sector o comuna de Medellín dependiendo de las condiciones climáticas del territorio.

Analizar la evaluación de los modelos generados bajo métricas estadísticas para la selección del modelo que se adapte con mayor precisión a los datos manejados.

Desplegar el modelo seleccionado en una herramienta visual para consulta.

# Datos

## Datos originales

Los datos utilizados en el estudio se obtuvieron mediante el portal de descargas del Sistema de Alerta Temprana de Medellín y el Valle de Aburrá SIATA junto con información de ingresos en hospitales del RIPS Medellín, donde se seleccionaron datos meteorológicos, ambientales y médicos. Estos son suministrados mediante archivos CSV (Comma-Separated Values) que facilitan la exploración, imputación y manipulación de ellos para generar nuevos datos categóricos y numéricos que nos permiten comprender el comportamiento de nuestros datos. Además, dado que se debía realizar la extracción de datos por estación de monitoreo, se generaron dataframes por cada variable para la identificación de ellas. Como ejemplo se muestra la sección de [código 1](#bookmark=kix.9xk4nw41jr05) donde se recolecta la información de los archivos csv de la calidad del aire en DataFrames de trabajo.

#Extraer datos de calidad del aire en las estaciones # 69 Caldas, 86 Aranjuez, 28 Itagüí

def get\_data (url):

"""

Función utilizada para la obtención de datos en tipo DataFrame según las bases de datos del repositorio

In[0]: url del dataset en el repositorio

Out[0]: DataFrame con los datos extraídos

"""

return pd.read\_csv(url)

# Datos de la calidad del aire tomados desde red SIATA de monitoreo de calidad del aire en Valle de Aburrá

cal\_aire\_ita = '28\_itagui/estacion\_data\_calidadaire\_28\_20240301\_20240331.csv'

cal\_aire\_cal = '69\_Caldas/estacion\_data\_calidadaire\_69\_20240301\_20240331.csv'

cal\_aire\_ara = '86\_Aranjuez/estacion\_data\_calidadaire\_86\_20240301\_20240331.csv'

cal\_aire\_bel = '83\_Belen/estacion\_data\_calidadaire\_83\_20240101\_20240131.csv'

cal\_aire\_ita = get\_data(cal\_aire\_ita)

cal\_aire\_cal = get\_data(cal\_aire\_cal)

cal\_aire\_ara = get\_data(cal\_aire\_ara)

cal\_aire\_bel = get\_data(cal\_aire\_bel)

Código 1. Generación de dataframes por variable de calidad del aire.

De esta extracción de archivos se nota que los archivos de variables meteorológicas (humedad, precipitación, presión, temperatura y vientos) tienen registros por minuto (figura 1), mientras que los archivos de la calidad del aire muestran archivos por hora (figura 2). Adicionalmente, se nota que los registros proporcionados presentan una clasificación que indica la calidad del dato tomado en la columna “Calidad”.

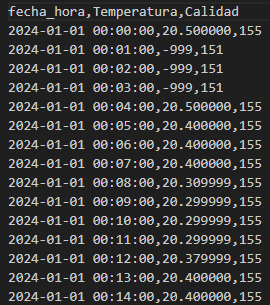


Figura 1. Datos de condiciones meteorológicas por minuto.

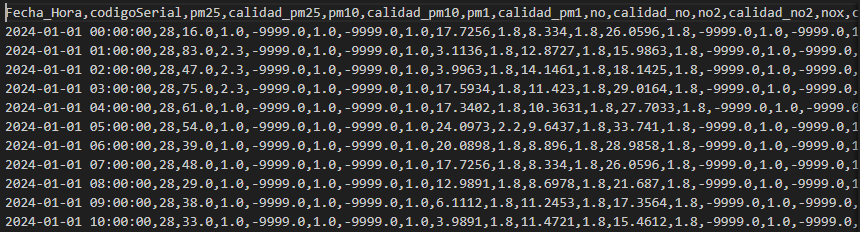


Figura 2. Datos de condiciones ambientales por hora.

Por otro lado, los datos proporcionados por RIPS Medellín generan registros encriptados de ingresos de pacientes declarados con diagnósticos asociados a enfermedades respiratorias, con ellos se tiene una granularidad de registros diarios desde 2020 a 2024, una muestra se tiene en la Figura 3.

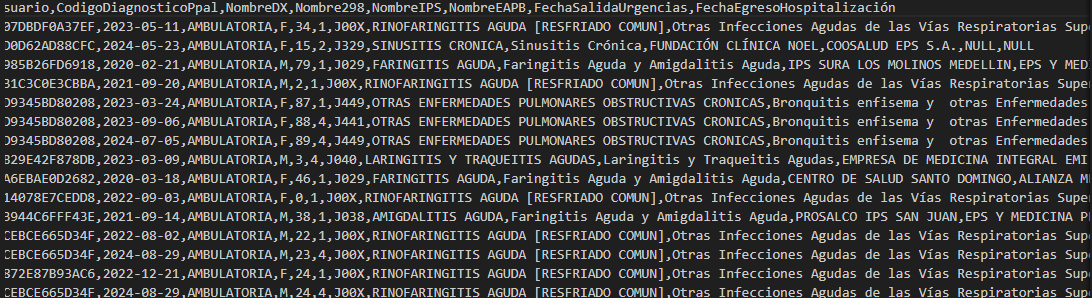


Figura 3. Datos médicos de pacientes con diagnósticos de enfermedades respiratorias.

De esta forma, se recopilan todos los archivos de datos mostrados en una misma tabla (ver [Anexo 1](#bookmark=id.jtolkt8uv24r)) donde se puede observar el nombre del archivo, descripción del archivo, peso de almacenamiento de datos, número de filas y número de columnas de cada uno de estos descargados desde el portal del SIATA y la información proporcionada por la entidad RIPS. Además, para el conocimiento de las características conjuntas de estos archivos se tiene el [Anexo 2](#bookmark=id.bl38ujveyp1e) donde se identifican los nombres de los campos, las descripciones de estos y los tipos de datos almacenados en ellos.

## Datasets

Usando los datos extraídos desde su forma cruda y por cada dataframe se procede a realizar una serie de manipulaciones para la estructuración del dataset de trabajo, entre ellas se tiene:

* Eliminación de columnas no significativas en los dataframes de calidad aire de las estaciones escogidas. Para ello entonces para cada característica se evaluó su valor máximo como mínimo para ver la consistencia de esos datos , para ver el resultado ver figura 4.



Figura 4 Valores mínimos y máximos columnas calidad aire.

Como se puede observar la única característica que su valor mínimo y máximo no tenía valores -999.0 era la partícula “pm25” indicándonos que los sensores del SIATA solo toma los valores de esta variable, por ende procedemos a no tener en el estudio los demás contaminantes como lo son “co” o “so2” entre otras. Logrando que tengamos en estos dataframes de calidad de aire la siguiente estructura como se detalla en la figura 5 a continuación.

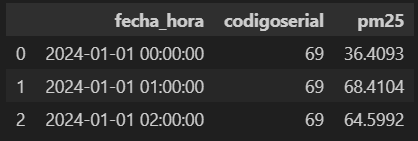


Figura 5 estructura dataframes calidad aire.

* Imputación de registros con valores negativos (-999.0) en los dataframes meteorológicos, para ello se dejó como primer paso esos valores negativos a que pasaran a ser nulos (NaN) ya que se usa la librería “KNNImputer ” y requiere que los datos estén de esa forma para hacer la debida transformación, por temas computacionales ese proceso se realiza solo una vez y se guarda el resultado en archivos con extensión csv.

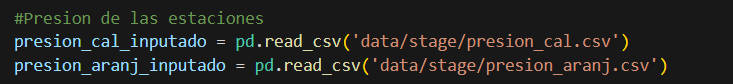


Figura 6 importar archivos meteorológicos proceso imputación.

Como se detalla en la figura 6 se lee los archivos meteorológicos ya con el proceso de imputación realizado evitando que se eliminaran esos registros nulos.

* Granularidad de los dataframes para ello se observó en los datos lo siguiente:
  + Los dataframes de calidad aire tenían la granularidad por hora
  + Los dataframes meteorológicos tenían una granularidad por minuto
  + Se determinó que se dejara la granularidad por día

Para que tengamos la misma consistencia agregamos tres nuevas columnas a partir de la fecha de cada registro en los dataframes estas son año, mes y día y agrupamos para que pudiéramos cumplir con la granularidad optada se puede observar el resultado de uno de los dataframes.



Figura 7 dataframe con granularidad por día.

* Creación de columna con nombre “día\_semana” esto nos ayuda a saber si el registro de los dataframes fue en el inicio de la semana , mediana semana o si fue el fin de semana, siendo una variable categórica y se aplicó en los dataframes de calidad aire y meteorológicos en la figura 8 se puede detallar el resultado.



Figura 8 dataframe con columna dia\_semana

* Creación columna categórica ‘estacion” en los dataframes de calidad aire con el fin de eliminar posteriormente la columna con nombre “codigoserial” ya que esta genera ruido en los datos.



Figura 9 dataframe con columna estación.

* Creación columna binaria que determina en cada registro si el día es festivo o no donde 1 significa que lo es, esto fue logrado usando la librería “Holidays” el resultado se puede ver en la siguiente figura.

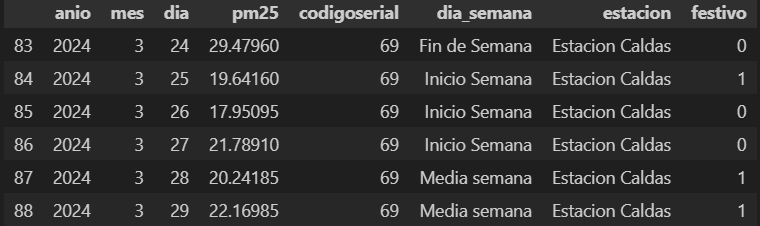


Figura 10 Creación columna festivo

Después de haber logrado todo lo anterior en los dataframes de calidad aire y los meteorológicos, ya podíamos concatenarlos para tener un único dataset logrando un resultado que se observa en la figura 11.



Figura 11 DataSet concatenado calidad aire

Después de ello se realizó una limpieza de datos esto con dos procesos que se explicaran brevemente:

* Se calculó los valores atípicos mediante el rango intercuartil esto mediante nuestra variable “pm25”, nos quedamos con todos los registros que estuvieran en la intersección del valor inferior y superior de este proceso.
* Se encontraron valores atípicos mediante la librería “LocalOutlierFactor” que nos permite identificar cuales son los índices que generan ruido a partir de los parámetros establecido que fueron:
  + n\_neighbors : 5
  + contamination : 0.05 (5%)

Los índices que se identificaron por esté método fueron eliminados del dataSet

Para el DataSet de los datos médicos se realizaron diferentes procesos que se describirán a continuación.

* Se tenía en el dataSet más de tres millones de registros de pacientes con problemas respiratorios entre los años 2020 al 2024.
* Se hizo una agrupación de los datos a partir de la columna “NombreIPS” que nos indica el nombre de la IPS para saber cuántas atenciones médicas atendió durante el periodo anterior ya descrito y que estas atenciones fueran más de 1020 dando el siguiente resultado que fue guardado en un archivo csv para su posterior manipulación con nombre “nombresIPS.csv”:

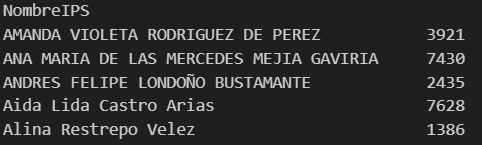


Figura 12 atenciones por IPS.

* En el archivo “nombresIPS.csv” se realizó un proceso manual que nos permitiera tener una columna llamada “comuna\_ips” esto con el fin de saber en qué comuna de Medellín la IPS presta el servicio de salud dando el siguiente resultado:

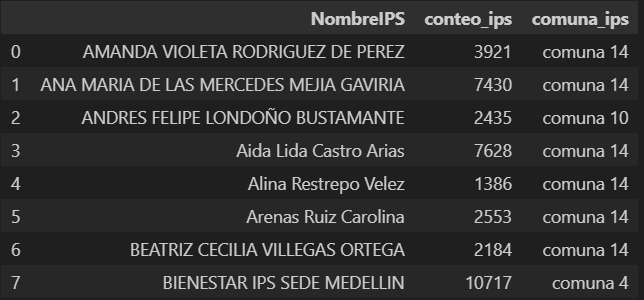


Figura 13 columna comuna\_ips para cada cada IPS.

* Por lo anterior, podemos hacer un merge del archivo “nombresIPS.csv” con el de los datos médicos para tener lo siguiente:

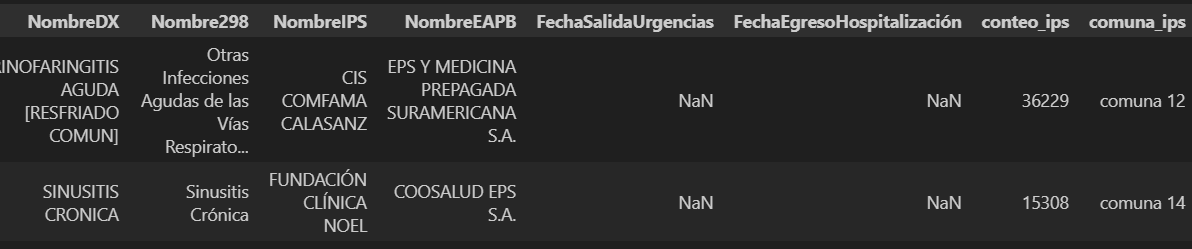


Figura 14 concatenación para tener columnas nuevas en dataset datos médicos.

* Al ver las columnas del dataset se eliminaron las columnas que no son pertinentes para el estudio quedando el dataframe de la siguiente forma:

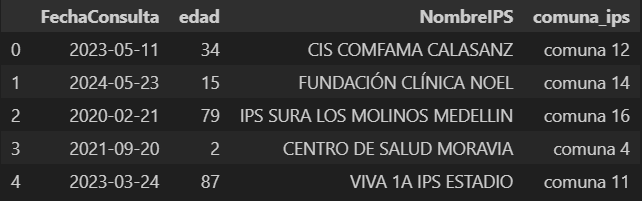


Figura 15 dataset datos médicos con columnas de estudio

* A partir de lo anterior se procede a crear las columnas año, mes y día a partir de la variable “FechaConsulta” que posteriormente se eliminaría. En la figura 16 se puede observar el resultado.

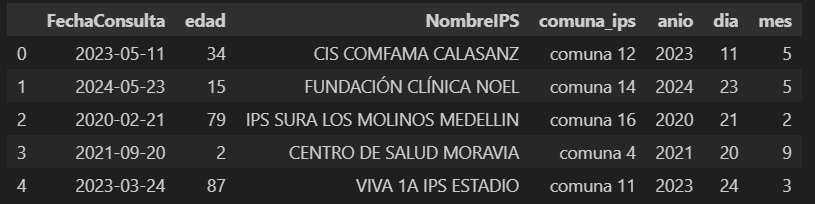


Figura 16 creación columnas año, mes y día

* Se creó una nueva característica categórica llamada “sector\_ips” esto es para saber la IPS en qué sector de Medellín está ubicada a partir de la columna “comuna\_ips” uno de estos valores pueden ser Centroccidental, Suroriental entre otros en la figura 17 se encontrará el resultado.

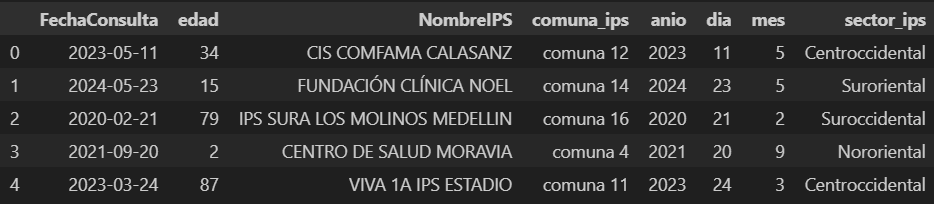


Figura 17 creación columna sector\_ips.

* Se crea la columna “atenciones\_por\_dia” esto es apartir de una agrupación que nos permitiera determinar cuántos atenciones se realizaron en la IPS esta variable es de vital importancia por que va ser la característica que nos ayudará a pronosticar nuestro objetivo en la siguiente imágen podemos detallar este resultado.

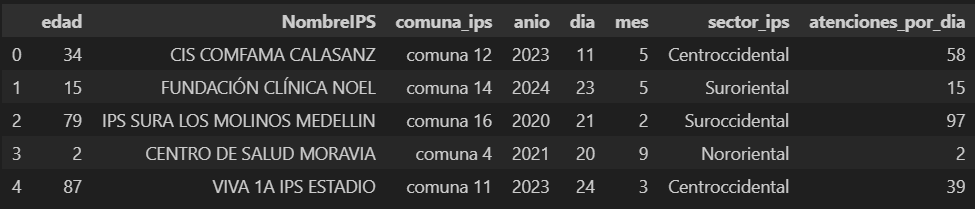


Figura 18 creación columna atenciones\_por\_dia.

Ya después de todo este proceso concatenamos nuestros dos datasets el de calidad aire junto con el de los datos médicos para tener un único dataset en la figura 19 se puede ver parte de este dataframe.

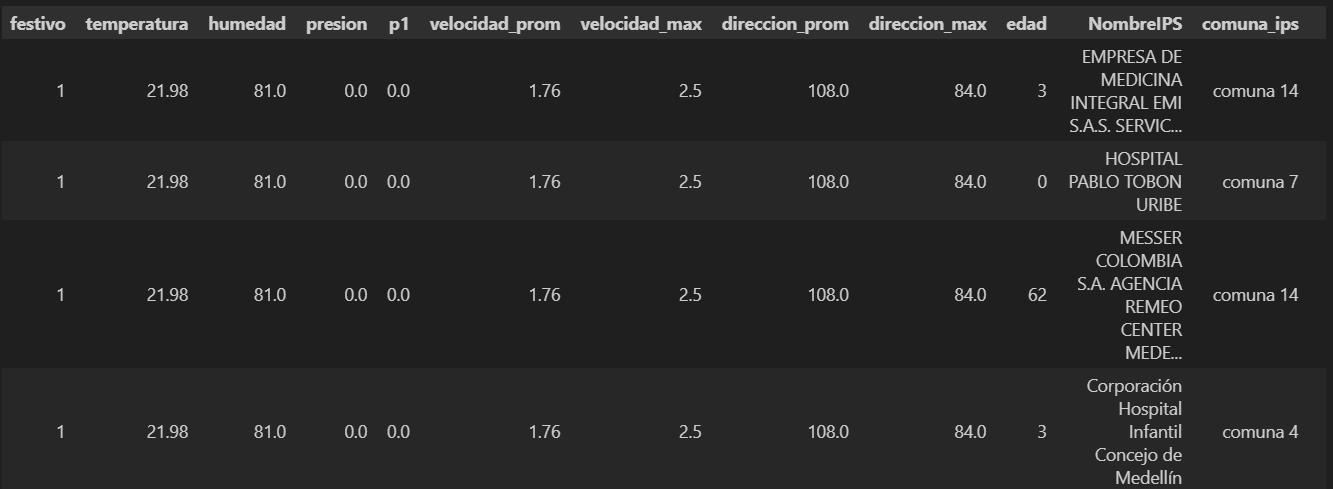


Figura 19 concatenación de los dos datasets.

A partir de esto se realizó la limpieza del dataset realizando lo siguiente:

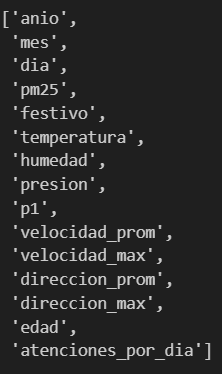
* Se calculó los valores atípicos mediante el rango intercuartil esto mediante nuestra variable de interés “atenciones\_por\_dia”, nos quedamos con todos los registros que estuvieran en la intersección del valor inferior y superior de este proceso.
* Se encontraron valores atípicos mediante la librería “LocalOutlierFactor” que nos permite identificar cuales son los índices que generan ruido a partir de los parámetros establecido que fueron:
  + n\_neighbors : 10
  + contamination : 0.10 (10%)

Los índices que se identificaron por esté método fueron eliminados del DataSet

Para finalizar con este apartado se eliminó la columna “NombreIPS” y se hizo one hot encoding de nuestras variables categóricas para que nuestro DataSet estuviera todas sus variables de tipo numérico. A su vez se estandarizaron los datos para tener más consistencia y que sea más óptimos los modelos desarrollados.

## Analítica descriptiva

En este apartado queremos detalla la prueba de normalidad de nuestros datos, en este caso las evaluaremos con las siguientes características:

  
Figura 20 Características prueba de normalidad.

Para ello iniciamos con los gráficos Q-Q plot que nos ayudan a identificar si nuestros datos siguen una distribución normal en la siguiente imágen veremos el resultado:

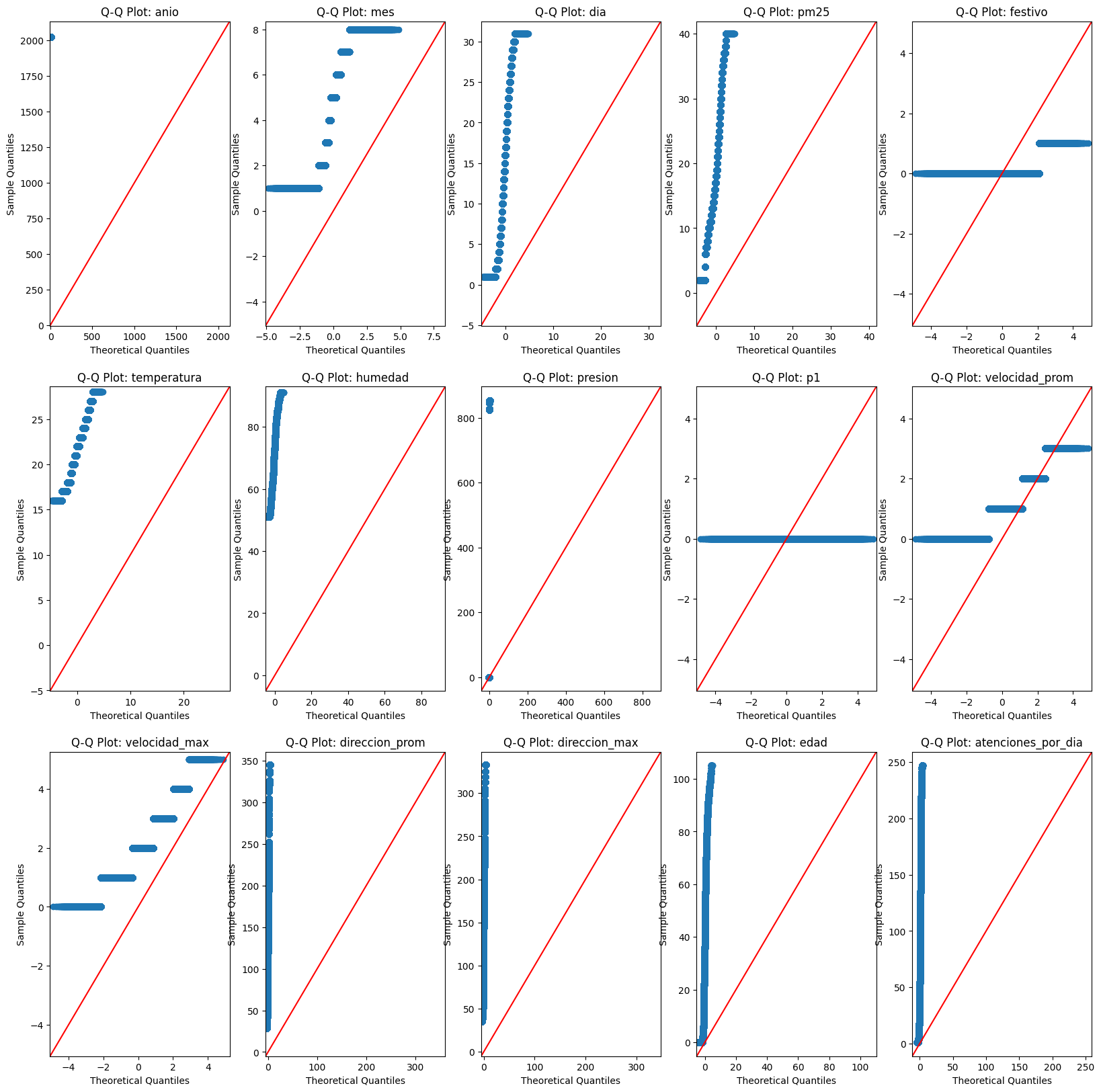


Figura 21 Q-Q plots de variables.

Gracias a la figura 21 podemos concluir que la distribución de los datos no siguen una distribución normal por ende, para confirmar lo anterior se hizo la normalidad de los datos con KDE (Estimación de densidad de kernel) esto suaviza la distribución para crear una curva continua que representa la densidad de probabilidad de los datos el resultado se puede ver en la siguiente figura:

# 

# Figura 22 distribución de los datos con KDE.

Dado el comportamiento asimétrico negativo de las distribuciones de los datos, estos no se comportan bajo una distribución normal debido al sesgo de las colas inferiores en la mayoría de las columnas y principalmente en nuestra variable de estudio "atenciones\_por\_dia". Esto indica que los cuantiles observados de los datos son más grandes que los cuantiles esperados bajo una distribución normal. Esto sugiere que los datos tienen una cola izquierda menos pesada que una distribución normal.

Para rectificar lo anterior se realizaron pruebas de shapiro para cada una de nuestras variables dando el siguiente resultado.

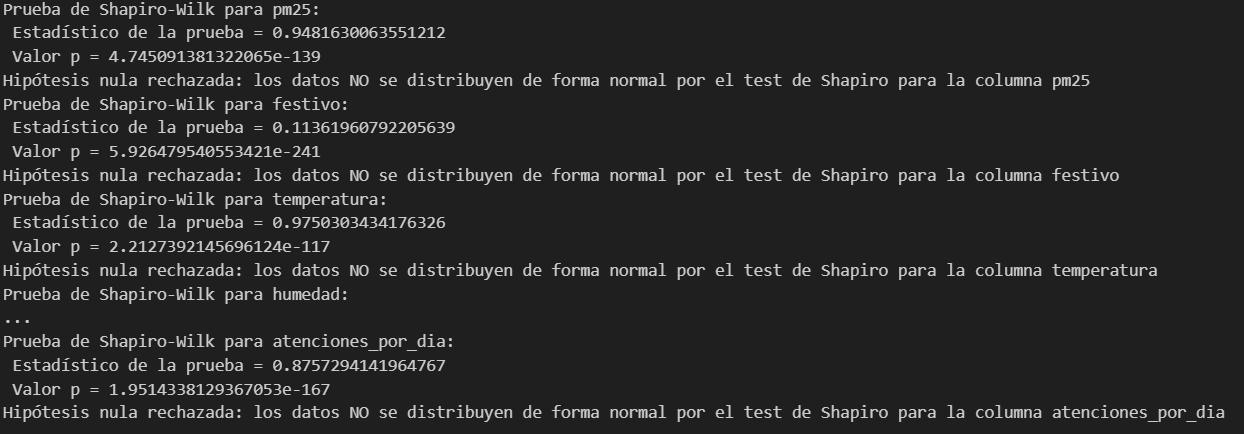


Figura 23 pruebas de Shapiro-Wilk.

# 

Como podemos observar en la figura 23 rectificamos que nuestras variables no siguen una distribución normal para ello para rectificar lo anterior se revisa el comportamiento de las columnas mediante box plots como se detalla en la figura 24.

# 

Figura 24 Box Plots de variables para determinar distribución.

Por los gráficos de box plot véase en la figura 24 podemos confirmar que los datos en la mayoría de variables tienden a establecer una cola inferior en las diferentes distribuciones de datos sesgando estas a los primeros cuantiles.

**Referencias**

Área Metropolitana Valle de Aburrá. (n.d.). *¿Qué es SIATA?* Área Metropolitana SIATA. Retrieved November 20, 2024, from https://www.metropol.gov.co/ambiental/siata/paginas/que-es.aspx

Hospital internacional de Colombia. (2024, Agosto 26). *Contaminación del aire y enfermedades respiratorias: ¿cómo afecta a tu salud?* Neumología y Salud Respiratoria. Retrieved Noviembre 16, 2024, from https://hic.fcv.org/co/blog/neumologia-y-salud-respiratoria/contaminacion-del-aire-y-enfermedades-respiratorias-como-afecta-a-tu-salud#:~:text=Adem%C3%A1s%2C%20la%20contaminaci%C3%B3n%20del%20aire,aumentar%20el%20riesgo%20de%20infecciones

Ministerio de Salud y Protección Social. (2022, November 17). *Muertes por enfermedades respiratorias crónicas han disminuido*. Ministerio de Salud y Protección Social. Retrieved November 16, 2024, from https://www.minsalud.gov.co/Paginas/Muertes-por-enfermedades-respiratorias-cronicas-han-disminuido.aspx

Ministerio de Salud y Protección Social - República de Colombia. (n.d.). *Sistema de Información de Prestaciones de Salud - RIPS*. Ministerio de Salud y Protección Social. Retrieved November 20, 2024, from https://www.minsalud.gov.co/proteccionsocial/paginas/rips.aspx

Agencia de Protección Ambiental de Estados Unidos. (2024, June 25). Conceptos básicos sobre el material particulado (PM, por sus siglas en inglés) | US EPA. EPA en español. Retrieved September 5, 2024, from https://espanol.epa.gov/espanol/conceptos-basicos-sobre-el-material-particulado-pm-por-sus-siglas-en-ingles

Área Metropolitana Valle del Aburrá. (2019). Calidad del aire. Área Metropolitana. Retrieved September 7, 2024, from https://www.metropol.gov.co/ambiental/calidad-del-aire

Baena, D., Jiménez, J., Zapata, C., & Ramirez, Á. (2019, Marzo 25). Red neuronal artificial aplicado para el pronóstico de eventos críticos de PM2.5 en el Valle de Aburrá. SciELO Colombia. Retrieved September 10, 2024, from http://www.scielo.org.co/scielo.php?script=sci\_arttext&pid=S0012-73532019000200347#t1

BBC News Mundo. (2017, November 15). Qué es EPOC, la enfermedad pulmonar silenciosa que mata a 3 millones de personas al año y no tiene cura. BBC. Retrieved September 7, 2024, from https://www.bbc.com/mundo/noticias-41997332

Ministerio de Salud y Protección Social. (2022, November 17). Muertes por enfermedades respiratorias crónicas han disminuido. Ministerio de Salud y Protección Social. Retrieved September 6, 2024, from https://www.minsalud.gov.co/Paginas/Muertes-por-enfermedades-respiratorias-cronicas-han-disminuido.aspx

Minsalud. (2021, August 23). Minsalud comprometido con la calidad del aire. Ministerio de Salud y Protección Social. Retrieved September 9, 2024, from https://www.minsalud.gov.co/Paginas/Minsalud-comprometido-con-la-calidad-del-aire-.aspx

Parra, J., Oviedo, A., & Omaya, F. (2020, Octubre 25). Analítica de datos: incidencia de la contaminación ambiental en la salud pública en Medellín (Colombia). Rev. Salud Pública, 22(6), 609-617. https://doi.org/10.15446/rsap.V22n6.78985

Pérez, J., Montoya, P., Sanchez, J. M., Hernández, S., & Ramírez, M. (2023, Diciembre 14). Forecasting 24 h averaged PM2.5 concentration in the Aburrá Valley using tree-based machine learning models, global forecasts, and satellite information. ASCMO - Volumes. Retrieved September 10, 2024, from https://ascmo.copernicus.org/articles/9/121/2023/ascmo-9-121-2023.html

Torres Jaramillo, J. A. (2021, September 6). El aire nuestro de cada día. Instituto de Ciencias de la Atmósfera y Cambio Climático. Retrieved September 5, 2024, from https://www.atmosfera.unam.mx/el-aire-nuestro-de-cada-dia/

# Anexos

**Anexo 1. Archivos utilizados para datos crudos de condiciones meteorológicas y calidad del aire.**

| Nombre archivo | Descripción | peso | Número de filas | Número de columnas |
| --- | --- | --- | --- | --- |
| estacion\_data\_humedad\_252\_\_20240201\_20240229.csv | Datos de humedad estación itagüí. | 13.27 KB | 394561 | 3 |
| estacion\_data\_precipitacion\_252\_\_20240201\_20240229.csv | Datos de precipitación estación itagüí. | 14.86 KB | 394561 | 4 |
| estacion\_data\_presion\_252\_\_20240201\_20240229.csv | Datos de presión estación itagüí. | 12.93 KB | 394561 | 3 |
| estacion\_data\_temperatura\_252\_\_20240201\_20240229.csv | Datos de temperatura estación itagüí. | 13.27 KB | 394561 | 3 |
| estacion\_data\_vientos\_252\_\_20240201\_20240229.csv | Datos de vientos estación itagüí. | 20.65 KB | 394561 | 6 |
| estacion\_data\_calidadaire\_28\_20240301\_20240331.csv | Datos de calidad del aire estación itagüí. | 1.32 KB | 6577 | 36 |
| estacion\_data\_humedad\_105\_\_20240201\_20240229.csv | Datos de humedad estación caldas. | 12.56 KB | 394561 | 3 |
| estacion\_data\_precipitacion\_105\_\_20240201\_20240229.csv | Datos de precipitación estación caldas. | 15.47 KB | 394561 | 4 |
| estacion\_data\_presion\_105\_\_20240201\_20240229.csv | Datos de presión estación caldas. | 12.89 KB | 394561 | 3 |
| estacion\_data\_temperatura\_105\_\_20240201\_20240229.csv | Datos de temperatura estación caldas. | 12.56 KB | 394561 | 3 |
| estacion\_data\_vientos\_105\_\_20240201\_20240229.csv | Datos de vientos estación caldas. | 19.94 KB | 394561 | 6 |
| estacion\_data\_calidadaire\_69\_20240301\_20240331.csv | Datos de calidad del aire estación caldas. | 1.46 KB | 6577 | 36 |
| estacion\_data\_humedad\_198\_\_20240101\_20240131.csv | Datos de humedad estación belen. | 12.51 KB | 394561 | 3 |
| estacion\_data\_precipitacion\_9\_\_20240101\_20240131.csv | Datos de precipitación estación belen. | 12.32 KB | 394561 | 4 |
| estacion\_data\_presion\_198\_\_20240101\_20240131.csv | Datos de presión estación belen. | 12.83 KB | 394561 | 3 |
| estacion\_data\_temperatura\_198\_\_20240101\_20240131.csv | Datos de temperatura estación belen. | 12.51 KB | 394561 | 3 |
| estacion\_data\_vientos\_198\_\_20240101\_20240131.csv | Datos de vientos estación belen. | 19.84 KB | 394561 | 6 |
| estacion\_data\_calidadaire\_83\_20240101\_20240131.csv | Datos de calidad del aire estación belen. | 1.46 KB | 6577 | 36 |
| estacion\_data\_humedad\_68\_\_20240201\_20240229.csv | Datos de humedad estación Aranjuez. | 12.71 KB | 394561 | 3 |
| estacion\_data\_precipitacion\_68\_\_20240201\_20240229.csv | Datos de precipitación estación Aranjuez. | 15.77 KB | 394561 | 4 |
| estacion\_data\_presion\_68\_\_20240201\_20240229.csv | Datos de presión estación Aranjuez. | 13.09 KB | 394561 | 3 |
| estacion\_data\_temperatura\_68\_\_20240201\_20240229.csv | Datos de temperatura estación Aranjuez. | 12.71 KB | 394561 | 3 |
| estacion\_data\_vientos\_68\_\_20240201\_20240229.csv | Datos de vientos estación Aranjuez. | 19.99 KB | 394561 | 6 |
| estacion\_data\_calidadaire\_86\_20240301\_20240331.csv | Datos de calidad del aire estación Aranjuez. | 1.46 KB | 6577 | 36 |
| RIPS\_J000\_J989\_2020-2024P.csv | Datos de ingresos de pacientes en IPS médica con diagnósticos asociados a enfermedades respiratorias | 941 KB | 3436530 | 13 |

**Anexo 2. Descripción de campos en archivos de datos**. El acceso a estos datos está libre de restricciones.

| Campo | Descripción | Tipo de Dato |
| --- | --- | --- |
| Temperatura | Variable que indica la temperatura ambiental en °C | Numérico |
| Presión | Variable que indica la presión atmosférica en mmHg | Numérico |
| festivo | Variable que clasifica si una fecha fue un día festivo o no | Binario |
| fecha\_hora | Variable que indica la fecha y hora del registro | Categórico |
| codigoSerial | Variable que indica el código serial de cada estación de monitoreo | Numérico |
| calidad\_pm25 | Variable categórica que indica la calidad del dato tomado en el monitoreo | Categórico |
| calidad\_pm10 | Variable categórica que indica la calidad del dato tomado en el monitoreo | Categórico |
| calidad\_pm1 | Variable categórica que indica la calidad del dato tomado en el monitoreo | Categórico |
| calidad\_no | Variable categórica que indica la calidad del dato tomado en el monitoreo | Categórico |
| calidad\_p\_ssr | Variable categórica que indica la calidad del dato tomado en el monitoreo | Categórico |
| calidad\_pliquida\_ssr | Variable categórica que indica la calidad del dato tomado en el monitoreo | Categórico |
| calidad\_rglobal\_ssr | Variable categórica que indica la calidad del dato tomado en el monitoreo | Categórico |
| calidad\_taire10\_ss | Variable categórica que indica la calidad del dato tomado en el monitoreo | Categórico |
| calidad\_vviento\_ssr | Variable categórica que indica la calidad del dato tomado en el monitoreo | Categórico |
| calidad\_no2 | Variable categórica que indica la calidad del dato tomado en el monitoreo | Categórico |
| calidad\_nox | Variable categórica que indica la calidad del dato tomado en el monitoreo | Categórico |
| calidad\_pst | Variable categórica que indica la calidad del dato tomado en el monitoreo | Categórico |
| calidad\_dviento\_ssr | Variable categórica que indica la calidad del dato tomado en el monitoreo | Categórico |
| calidad\_haire10\_ssr | Variable categórica que indica la calidad del dato tomado en el monitoreo | Categórico |
| calidad\_ozono | Variable categórica que indica la calidad del dato tomado en el monitoreo | Categórico |
| calidad\_so2 | Variable categórica que indica la calidad del dato tomado en el monitoreo | Categórico |
| calidad\_co | Variable categórica que indica la calidad del dato tomado en el monitoreo | Categórico |
| pm25 | Variable que indica la cantidad de material particulado menor de 2.5 micras por m3 | Numérico |
| pm10 | Variable que indica la cantidad de material particulado menor de 10 micras por m3 | Numérico |
| pm1 | Variable que indica la cantidad de material particulado menor de 1.0 micras por m3 | Numérico |
| no | Variable que indica las ppb de óxido de nitrógeno | Numérico |
| no2 | Variable que indica las ppb de dióxido de nitrógeno | Numérico |
| nox | Variable que indica las ppb de óxidos de nitrógeno | Numérico |
| ozono | Variable que indica las ppb de ozono | Numérico |
| co | Variable que indica las ppm de monóxido de carbono | Numérico |
| so2 | Variable que indica las ppb de dióxido de azufre | Numérico |
| pst | Variable que indica la magnitud de la precipitación | Numérico |
| dviento\_ssr | Variable que indica la dirección del viento | Numérico |
| haire10\_ssr | Variable que indica la humedad del aire a 10 metros | Numérico |
| p\_ssr | Variable que indica la presión | Numérico |
| pliquida\_ssr | Variable que indica la precipitación | Numérico |
| rglobal\_ssr | Variable que indica la radiación | Numérico |
| taire10\_ssr | Variable que indica la temperatura del aire a 10 metros | Numérico |
| vviento\_ssr | Variable que indica la magnitud de velocidad del viento | Numérico |
| velocidad\_prom | Variable que indica la velocidad media del viento | Numérico |
| velocidad\_max | Variable que indica la velocidad máxima del viento | Numérico |
| direccion\_prom | Variable que indica la dirección media del viento | Numérico |
| direccion\_max | Variable que indica la dirección del viento máximo | Numérico |
| Encrypt | Variable que indica la los datos encriptados de paciente atendido | Texto |
| FechaConsulta | Variable que indica la fecha de consulta | Texto |
| TipoCupsAtencion | Variable que indica el tipo de consulta | Texto |
| sexo | Variable que indica el sexo del paciente | Texto |
| edad | Variable que indica la edad del paciente | Numérico |
| TipoUsuario | Variable que indica el tipo de usuario | Texto |
| CodigoDiagnosticoPpal | Variable que indica el código del diagnóstico principal asociado en la consulta | Numérico |
| NombreDX | Variable que indica el nombre del diagnóstico científico | Texto |
| Nombre298 | Variable que indica el nombre del diagnóstico | Texto |
| NombreIPS | Variable que indica el nombre del establecimiento de atención | Texto |
| NombreEAPB | Variable que indica la clasificación del tipo de enfermedad | Texto |
| FechaSalidaUrgencias | Variable que indica la fecha de salida del paciente de urgencias | Texto |
| FechaEgresoHospitalizacion | Variable que indica la salida del paciente del establecimiento | Texto |