Prototipo de diagnóstico ambiental sobre partículas PM2.5 en el Valle de Aburrá, a través de IoT y analítica avanzada de datos

Universidad Pontificia Bolivariana

GIDATI
Ferney Amaya (ferney.amaya@upb.edu.co)
Ana Oviedo Carrascal
Sebastian Parra Sanchez
Sebastian Ruiz Gomez

MINTIC
vive digital para la gente
CENTRO DE INNOVACIÓN PÚBLICA DIGITAL
TODOS POR UN NUEVO PAÍS PAZ EQUIDAD EDUCACIÓN
PNUD
UN-DESA

Contenido

GENERALIDADES DE LA INVESTIGACIÓN

1 Contexto de la investigación

1.1 Información de contacto del grupo de investigación

GIDATI: Grupo de Investigación Desarrollo y Aplicación en Telecomunicaciones e Informática

Nombre del líder de la propuesta de investigación	**FERNEY ORLANDO AMAYA FERNÁNDEZ**		
Celular	**300 468 2264**	**Correo**	ferney.amaya@upb.edu.co

1.2 Entidad Beneficiaria

IDEAM

1.3 Necesidad de la Entidad Beneficiaria

Analítica para diagnóstico ambiental; Medición y aprovechamiento de la información de calidad del aire en el territorio nacional, a través de IoT y analítica avanzada de datos.

1.4 Reto

Convocatoria para Grupos y Centros de Investigación 2018.

Alcance de los servicios y resultados esperados del proyecto: Generar una investigación en un periodo máximo de tres meses y medio, que demuestre de forma teórica y práctica (a través de un prototipo), las potencialidades y los beneficios que puede generar la aplicación de tecnologías emergentes, en la resolución de las problemáticas de entidades del sector público y la sociedad relacionados con ODS.

RETO 1. Entidad: IDEAM / subdirección de estudios ambientales
Nombre del Reto: Analítica para diagnóstico ambiental; Medición y aprovechamiento de la información de calidad del aire en el territorio nacional, a través de IoT y analítica avanzada de datos.

1.5 Resumen de los Resultados

Se desarrolló un prototipo que integra varias herramientas de procesamiento y analítica de datos avanzada para realizar una primera aproximación a la estimación de la población humana expuesta a la contaminación atmosférica que se presenta en el área de influencia de dos estaciones seleccionadas en el Área Metropolitana del Valle de Aburrá de Medellín (AMVA).

2 Alcance de la investigación

2.1 Enfoque del proyecto de investigación

Se desarrolló un prototipo analítico para determinar la población expuesta clasificada por grupos sensibles al material particulado PM2.5 alrededor de dos estaciones de monitoreo seleccionadas del Valle de Aburrá. Se seleccionará una estación de monitoreo de representatividad poblacional y otra de tráfico o industriales, ambas disponibles en el Área Metropolitana.

2.2 Cambios esperados en la entidad beneficiaria

El prototipo desarrollado puede emplearse como punto de partida para determinar la población expuesta a la contaminación ambiental. El prototipo servirá como insumo para el desarrollo de posteriores investigaciones que permitan determinar la influencia de la contaminación ambiental.

RESUMEN DE INVESTIGACIÓN

1. Resumen de la investigación

Un estudio de la Organización de las Naciones Unidas (ONU) señala que por primera vez en la historia más de la mitad del planeta vive en ciudades lo que genera diferentes problemáticas en materia de desigualdad social, movilidad, inseguridad y contaminación del aire (Bouskela, 2016).

La contaminación del aire incluye la presencia de partículas que pueden tener efecto en la salud ya que aumentan el riesgo de enfermedades cardiovasculares y respiratorias que causan muertes prematuras (Desarkar, 2018). En esta propuesta se aborda específicamente la problemática asociada al material particulado PM 2.5 en el Valle de Aburrá.

En esta propuesta se analizó información compilada por el IDEAM de calidad del aire de las estaciones de las autoridades ambientales que manejan información climática.

Se desarrolló un prototipo que integra varias herramientas de procesamiento y analítica de datos avanzada. El prototipo permite identificar los días tipo para una estación de monitoreo de calidad del aire y calcular el área de dispersión de contaminante PM2.5 para cada día tipo. Posteriormente, empleando los datos de una consulta al geoportal del DANE, es posible estimar la población humana expuesta a la contaminación atmosférica que se presenta en el área de influencia de la estación de monitoreo. El prototipo se empleó para analizar dos estaciones en el Área Metropolitana del Valle de Aburrá de Medellín (AMVA).

Con este prototipo se espera aportar a la labor del IDEAM, que entre sus funciones provee de información ambiental y de los efectos del cambio climático a nivel nacional, ya que de esta forma es posible estimar la población afectada alrededor de una estación de monitoreo de calidad del aire.

M. Bouskela, M. Casseb, S. Bassi, C. De Luca y M. Facchina (2016). «La ruta hacia las Smart Cities: Migrando de una gestión tradicional a la ciudad inteligente.,»

Desarkar y A. Das (2018). «A smart air pollution analytics framework,» Advances in Intelligent Systems and Computing, vol. 625, pp. 197-205.

2. Palabras claves

Contaminación ambiental, población expuesta, analítica de datos.

3. Impacto

El prototipo analítico permitirá determinar, de acuerdo el nivel de alerta, el número de personas que podrían ser afectadas por esos niveles de contaminación.

4. Objetivo de Investigación

Objetivo general

Realizar un prototipo analítico de diagnóstico ambiental sobre partículas PM2.5 en el Valle de Aburrá, a través de IoT y analítica avanzada de datos.

Objetivos específicos

- Integrar los datos relacionados con material particulado PM 2.5, meteorología y demografía de al menos dos estaciones de calidad del aire del Valle de Aburrá.
- Preparar los datos para el proceso de minería de datos mediante la evaluación de calidad y limpieza.
- Crear los modelos analíticos con los datos seleccionados.
- Evaluar los resultados obtenidos del proceso de modelamiento.
- Definir recomendaciones y conclusiones con los resultados del proceso de modelado.

5. Tipo de Investigación

Transferencia tecnológica y asesoría.

6. Resumen de las etapas de la investigación y metodología

Para desarrollar el proceso de analítica se propone la aplicación de la metodología CRISP-DM, resaltando que se trata de una metodología flexible y se puede personalizar fácilmente según los objetivos del proyecto a implementar (IBM, 2012).

FASE 1 - COMPRENSIÓN DEL PROBLEMA:

- Identificación del problema.
- Definición de los supuestos del modelo analítico.
- Selección de las estaciones.

FASE 2 - COMPRENSIÓN DE LOS DATOS:

- Análisis de los datos suministrados por el IDEAM.
- Análisis de los datos demográficos disponibles en la página del DANE.

FASE 3 - PREPARACIÓN DE LOS DATOS:

- Evaluación de la calidad y limpieza de los datos suministrados por el IDEAM.
- Preparación de los datos para el modelo analítico.
- Preparación de los datos para el sistema de modelamiento AERMOD.

FASE 4 - MODELADO:

- Definición de escenarios: Con los datos preparados se realiza un análisis de clustering para identificar los escenarios o días típicos para cada una de las estaciones seleccionadas.
- Análisis de dispersión por escenario: Por cada escenario identificado en el análisis de clustering, con la herramienta AERMOD se calcula la dispersión del PM2.5 lo que indica el área de influencia, que es el área cercana a la estación donde el nivel de concentración del contaminante es alto.

FASE 5 – EVALUACIÓN:

- Estimación de la calidad y grado de asertividad de los modelos creados.

FASE 6 – DESPLIEGE:

- Análisis de cada escenario o cluster identificado.
- Análisis de demografía por escenario: Según la dispersión calculada por AERMOD o área de influencia, se define a través del geoportal del DANE, la población afectada, con los resultados extrapolados al año 2018.

7. Elementos conceptuales de la investigación

A continuación, se presenta el marco conceptual referente a las tecnologías emergentes y el contexto nacional para la solución del reto.

Calidad del aire en el territorio nacional

Según la resolución 2254 de 2017, de acuerdo con la Organización Mundial de la Salud, se considera que el aire limpio es un requisito básico de la salud y el bienestar humano. Sin embargo, su contaminación sigue representando una amenaza importante para la salud en todo el mundo; en este contexto, se requiere definir una nueva norma de calidad del aire que incorpore un ajuste progresivo de los niveles máximos permisibles de contaminantes, incluir nuevos contaminantes y definir elementos técnicos integrales para mejorar la gestión de la calidad del aire.

Dentro del Plan Operacional para Enfrentar Episodios Críticos de Contaminación Atmosférica del Valle de Aburrá, se contemplan un conjunto de medidas tendientes a reducir los niveles de contaminación en el corto plazo, orientadas a prevenir la exposición de la población a altos índices de contaminación atmosférica.

Los datos necesarios para analizar el potencial meteorológico de contaminación por medio de la minería de datos, pueden ser tomados de la red de monitoreo de la calidad del aire. Esta red consta de varias estaciones distribuidas a lo largo y ancho del Valle de Aburrá, para realizar seguimiento de las concentraciones en puntos representativos de los entornos que conforman el área metropolitana. Estas estaciones cuentan con equipos para la medición de parámetros meteorológicos y concentraciones de algunas especies contaminantes, entre ellas el material particulado 2.5 (PM2.5), ozono y monóxido de carbono.

La red meteorológica está compuesta por sensores multiparamétricos, que proporcionan información minuto a minuto de temperatura, humedad relativa, precipitación, presión atmosférica, velocidad y dirección de vientos. Además de contar con dispositivos para la medición de vario contaminante, entre ellos el PM2.5.

IoT

Un sistema IoT (Internet of Things) integra componentes tecnológicos para la captura, transmisión y análisis de datos para generar sistemas completamente autónomos para monitorear o tomar decisiones sobre sistemas complejos. En el caso de la contaminación ambiental, un sistema IoT emplea los datos provenientes de sensores de variables ambientales y de calidad del aire que son transmitidos a través de redes de transmisión de datos existentes para posteriormente emplear herramientas de análisis y visualización de los datos.

La Red de calidad del aire del Valle de Aburrá, cuenta con 43 puntos de monitoreo entre automáticos y manuales (Siata). Esta red emplea la información de los siguientes instrumentos:

- Ceilómetros
- Radiómetro
- Radar Perfilador de Vientos
- Estaciones Meteorológicas

Otros equipos de sensado con que cuenta la SIATA son:

- Pluviómetros
- Piranómetros
- Cámaras

Analítica avanzada

La Ciencia de Datos es un estudio interdisciplinar de un conjunto de datos que involucra el diseño de bodegas de datos, la aplicación de técnicas estadísticas, el diseño de tableros de visualización de datos, el desarrollo de algoritmos computacionales para minería de datos (estructurados y no estructurados) y el procesamiento en la nube (Oviedo y otros, 2017). En la actualidad, la ciencia de datos ha tenido un especial interés gracias a la explosión generada por Big Data y Data Analytics. En específico, data analytics se define como el proceso de descubrir y extraer información con valor por medio de algoritmos de búsqueda e identificación de patrones, tendencias, desviaciones y otros indicadores que extraen conocimiento de grandes repositorios de datos para brindar a los directivos un panorama detallado de su negocio. En analítica se pueden desarrollar tareas predictivas y tareas descriptivas.

8. Principales logros por etapas de la investigación

Se presentan los principales logros por cada una de las fases del desarrollo del piloto.

FASE 1 - COMPRENSIÓN DEL PROBLEMA:

El modelo analítico planteado tiene como supuesto que la medición de la estación de monitoreo es igual a la emisión de la fuente. Para cumplir este supuesto la estación de monitoreo debe estar en el mismo lugar que la fuente de emisión que se desea supervisar.

Para el caso del área metropolitana, en que las estaciones son fijas, dos estaciones que cumplen con este requisito con las estaciones MUSEO DE ANTIOQUIA y CASA DE JUSTICIA DE ITAGÜÍ del Valle de Aburrá. Estas fueron las estaciones seleccionadas y cumplen con el criterio de estar colocadas cerca de las fuentes principales de contaminación. La estación MUSEO DE ANTIOQUIA está ubicada en el centro de Medellín y fue seleccionada por tener a su alrededor principalmente fuentes de emisión de tráfico vehicular. La estación CASA DE JUSTICIA DE ITAGÜÍ está ubicada en zona urbana-industrial de Itagüí y fue seleccionada por tener a su alrededor principalmente fuentes de emisión industrial.

FASE 2 - COMPRENSIÓN DE LOS DATOS:

Se identificaron las variables asociadas a cada estación que fueron entregadas por el IDEAM, las cuales fueron: Concentración PM 2.5, Velocidad del viento, Temperatura del aire, Precipitación, Humedad relativa, Radiación solar, Año, Mes, Día del mes y datos de la Estación de monitoreo ambiental.

Se identificaron las características y requerimientos para solicitar información al Geoportal del IDEAM que son: latitud y longitud de la ubicación y radio del área a solicitar la población.

FASE 3 - PREPARACIÓN DE LOS DATOS:

En la base de datos entregada por el IDEAM se encontraron algunos problemas de calidad de datos en cuanto a granularidad de los registros, completitud y consistencia.
La base de datos suministrada por el IDEAM luego de realizar el proceso de calidad sugerido, es necesario prepararlo y para esto se desarrolló el script en R:

1_PreparacionModeloAnalitico.R

Este script toma las variables entregadas por el IDAM y calcula otras variables que serán empleadas en el modelo para cada una de las dos estaciones seleccionadas.

Para determinar el área de influencia se calcula la dispersión de contaminantes empleando un sistema de modelado que tiene el programa principal AERMOD y dos pre-procesadores AERMET (AERMOD Meteorological Preprocessor) y AERMAP. Se identificaron las variables y fuentes para realizar el procesamiento en AERMOD.

FASE 4 - MODELADO:

El modelo analítico parte de una identificación de escenarios. Con los datos preparados se realizó un análisis de clustering (agrupamiento) empleando el algoritmo K-means para identificar los tipos de comportamiento del PM2.5 según la meteorología, la fecha y el tipo de estación. De cada cluster o grupo se identificaron los registros más representativos (más cercanos al centroide del cluster). Para estos registros representativos se realizó un análisis de dispersión empleando la herramienta AERMOD que entrega el área de influencia alrededor de cada una de las estaciones de monitoreo seleccionadas.

Los scripts en R desarrollados para esta fase fueron:

2_PreparacionClusterización.R

3_SeleccionFactores.R

4_Clusterización.R

FASE 5 – EVALUACIÓN:

Para estimar la calidad y grado de asertividad del proceso de *clustering* empleamos el Índice de Silueta. El índice de silueta es un valor entre 0 y 1 y para obtener una buena agrupación el índice de la silueta debe ser al menos de 0.5. En la Tabla 13 se presenta el índice de silueta para cada una de las estaciones, obteniendo 0.50 y 0.48, indicando que la agrupación obtenida es adecuada.

FASE 6 – DESPLIEGE:

Se realizó una descripción de los clusters para cada una de las estaciones.
Se calculó la población afectada (por grupo poblacional) de cada uno de los clusters de cada estación. Esto se realizó con la ayuda del geoportal del DANE.

9. Principales aprendizajes

Se articuló el conocimiento teórico y práctico de un grupo de investigación académico con una situación tangible de una entidad pública. En este caso, se emplearon técnicas de analítica avanzada de datos para el diagnóstico ambiental, aprovechando la información de calidad del aire en el territorio nacional que tiene disponible el IDEAM.

De esta forma se demuestra que puede darse respuesta a un problema público identificado, como es la contaminación del aire a través del desarrollo de un prototipo de analítica de datos que permite estimar la población afectada alrededor de una estación de monitoreo ambiental.

10. Lo que sigue en el futuro

Como trabajo futuro (fuera del alcance de este proyecto) se propone:

- Realizar un modelo de predicción de PM2.5: sistema donde se ingrese la meteorología esperada para una fecha y modelo arroje la predicción sobre el PM2.5 esperado.
- Realizar un modelo de predicción de población afectada: sistema donde se ingrese la meteorología esperada en una fecha y el PM2.5 esperado; el sistema arroja la población afectada según los escenarios definidos en este proyecto.

11. Conclusiones

Se desarrolló una metodología que permite calcular escenarios o días típicos a partir de un conjunto histórico de variables meteorológicas. Para el caso de la estación Museo de Antioquia se identificaron cinco días tipo que tienen las siguientes características:

	Descripción	avg_pm25
1	Precipitación alta, radiación media, humedad media.	32.137
2	Precipitación media, radiación media, máxima dirección del viento, velocidad del viento baja.Se presenta pocas veces en marzo y octubre.	31.27
3	Precipitación baja, radiación alta, humedad baja, presión baja, dirección del viento baja, máxima velocidad del viento. Se presenta en octubre y diciembre, los días viernes, sábado y domingo.	28
4	Precipitación media, máxima radiación. Se presenta pocas veces en marzo y mayo.	29.481
5	Precipitación media, mínima radiacion, máxima humedad, máxima presión. Se presenta en marzo, abril y octubre.	34

Para el caso de la estación Casa de Justicia de Itagüí se identificaron cinco días tipo que tienen las siguientes características:

Cluster	Descripción	avg_pm25
1	Alta radiación, humedad baja, baja dirección del viento, alta velocidad de viento. Se presenta pocas veces en enero, junio, julio y diciembre.	26.64
2	Radiación media, presión alta, poca lluvia, máxima dirección del viento	26.53
3	Radiación alta, humedad baja, presión baja, poca lluvia, velocidad del viento alta. No se presenta en enero	24.71
4	Radiación muy baja, humedad alta, precipitación muy alta, velocidad del viento baja. No se presenta en febrero ni septiembre. Se presenta muy poco en diciembre	27.59
5	Radiación media, humedad baja, dirección del viento alta.	24.73

A partir de los resultados es posible calcular la población afectada alrededor de una estación de monitoreo de la calidad del aire. Para el caso de la estación Museo de Antioquia, el peor caso ocurre para el día tipo número 5 en el cual se afectan 540 personas alrededor de la estación. Para el caso de la estación Casa de Justicia de Itagüí, el peor caso ocurre para el día tipo número 5 en el cual se afectan 1576 personas alrededor de la estación.

PROCESO DE INVESTIGACIÓN

1. En qué consiste la Investigación?

La investigación consiste en la realización de un prototipo analítico para el análisis ambiental sobre partículas PM2.5 en el Valle de Aburrá, a través de IoT y analítica avanzada de datos.

2. Proceso de Investigación

El desarrollo de la investigación se realizó en 6 fases:

FASE 1 - COMPRENSIÓN DEL PROBLEMA

FASE 2 - COMPRENSIÓN DE LOS DATOS

FASE 3 - PREPARACIÓN DE LOS DATOS

FASE 4 - MODELADO

FASE 5 - EVALUACIÓN

FASE 6 - DESPLIEGE

3. Requerimientos funcionales

El modelo analítico planteado tiene como supuesto que la medición de la estación de monitoreo es igual a la emisión de la fuente.

Para cumplir este supuesto la estación de monitoreo debe estar en el mismo lugar que la fuente de emisión que se desea supervisar, para lo cual se tienen las siguientes opciones:

- Usar estaciones de monitoreo móviles para situarlas lo más cerca posible de la fuente de emisión.
- Si la estación de monitoreo es fija debe seleccionarse la estación monitoreo que cumpla con este requisito.

4. Requerimientos no funcionales

- No aplica

5. Casos de uso (describa detalladamente cada caso de uso)

Se obtuvo la población afectada para las Estaciones Casa de justicia y Museo de Antioquia (mayor detalle en el informe técnico) como se presenta en las siguientes tablas y gráficas.

	Área total	*CLUSTER* 1	*CLUSTER* 2	*CLUSTER* 3	*CLUSTER* 4	*CLUSTER* 5
Total personas	11134	505	487	226	191	540
0-4 años	546	25	24	11	9	26
5-9 años	469	21	21	10	8	23
10-14 años	733	33	32	15	13	36
15-20 años	938	43	41	19	16	45
21-24 años	865	39	38	18	15	42
25-29 años	873	40	38	18	15	42
30-34 años	754	34	33	15	13	37
35-39 años	889	40	39	18	15	43
40-44 años	947	43	41	19	16	46
45-49 años	809	37	35	16	14	39
50-54 años	825	37	36	17	14	40
55-59 años	629	29	28	13	11	30
60-64 años	472	21	21	10	8	23
65-69 años	390	18	17	8	7	19
70-74 años	572	26	25	12	10	28
75-79 años	229	10	10	5	4	11

80 o más años	194	9	8	4	3	9
Hombres	6199	281	271	126	107	300
Mujeres	4935	224	216	100	85	239

Tabla. Población afectada por *cluster* para la estación Museo de Antioquia.

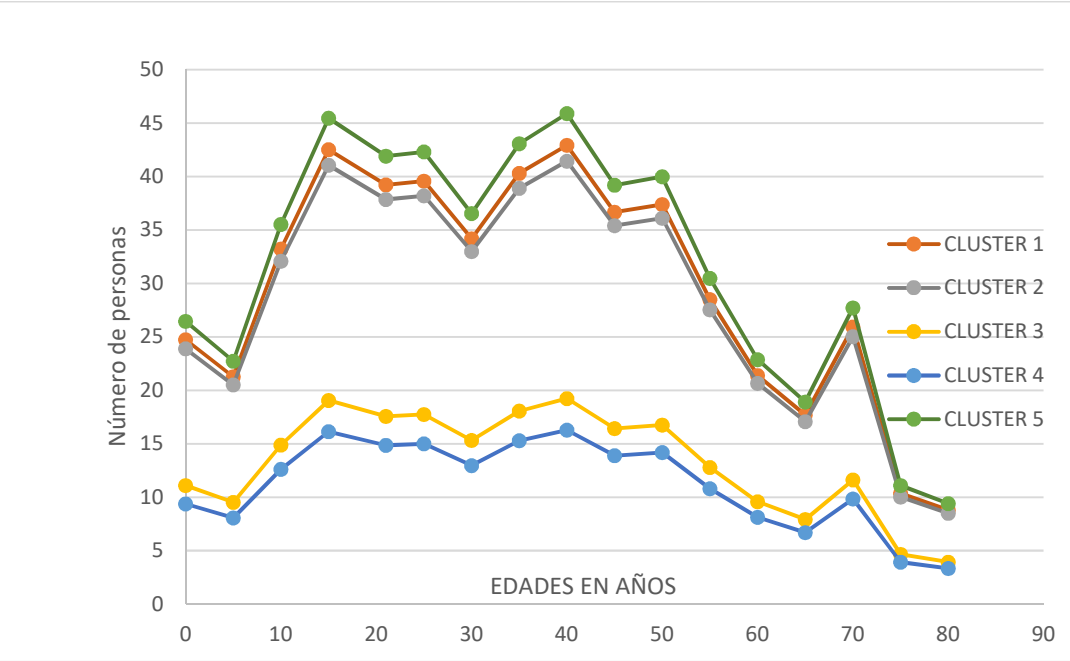

Figura. Población afectada por grupo poblacional para cada uno de los *clusters* para la estación Museo de Antioquia.

	Consulta	*CLUSTER* 1	*CLUSTER* 2	*CLUSTER* 3	*CLUSTER* 4	*CLUSTER* 5
Total personas	14225	1293	1656	929	1818	1576
0-4 años	818	74	95	53	105	91
5-9 años	1037	94	121	68	133	115
10-14 años	1136	103	132	74	145	126
15-20 años	1222	111	142	80	156	135
21-24 años	1291	117	150	84	165	143
25-29 años	1136	103	132	74	145	126
30-34 años	1056	96	123	69	135	117
35-39 años	1218	111	142	80	156	135
40-44 años	1199	109	140	78	153	133
45-49 años	995	90	116	65	127	110
50-54 años	729	66	85	48	93	81
55-59 años	682	62	79	45	87	76
60-64 años	507	46	59	33	65	56
65-69 años	446	41	52	29	57	49
70-74 años	299	27	35	20	38	33
75-79 años	240	22	28	16	31	27

80 o más	206	19	24	13	26	23
Hombres	6463	587	753	422	826	716
Mujeres	7754	705	903	507	991	859

Tabla. Población afectada por *cluster* para la estación Casa de la Justicia.

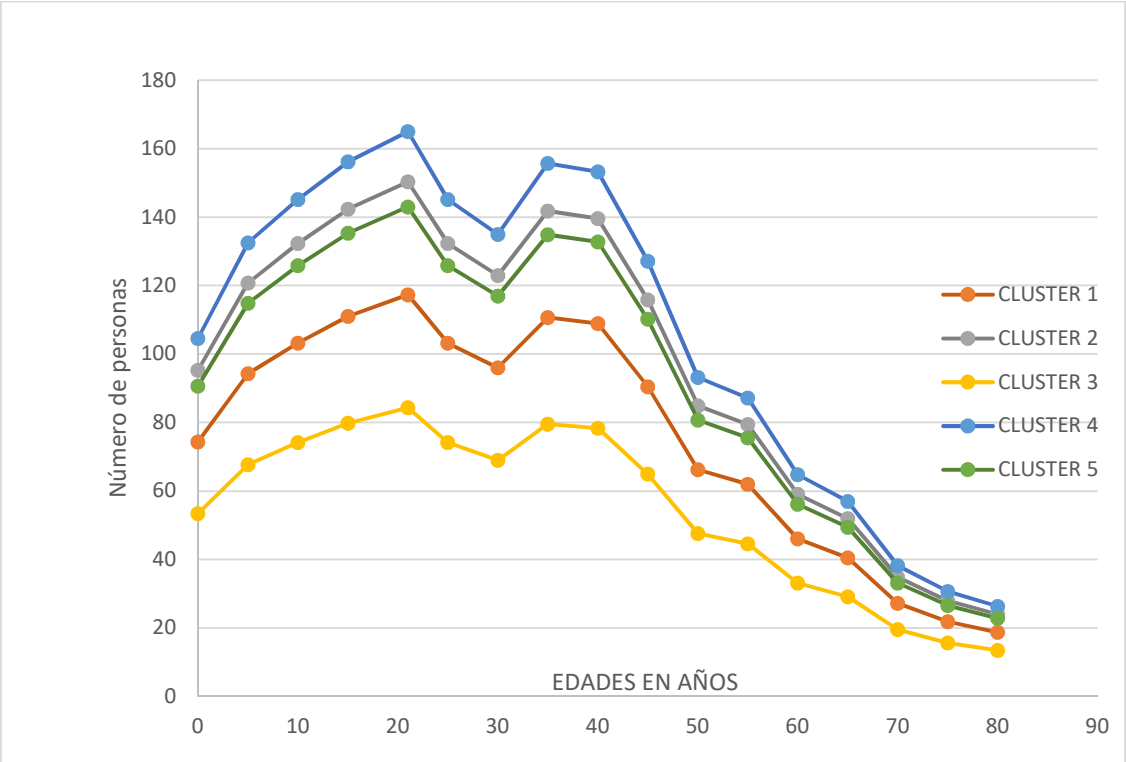

Figura. Población afectada por grupo poblacional para cada uno de los *clusters* para la estación Casa de la Justicia.

6. Actividades de funcionamiento Paso a paso

Las actividades realizadas corresponden a las fases de desarrollo del prototipo.

FASE 1 - COMPRENSIÓN DEL PROBLEMA:

- Definición de los supuestos del modelo analítico: el principal supuesto es que la estación de monitoreo debe estar muy cerca de la principal fuente de contaminación ambiental en el área.
- Selección de las estaciones: se seleccionaron las estaciones MUSEO DE ANTIOQUIA con fuentes de emisión vehicular y CASA DE JUSTICIA DE ITAGÜÍ con fuentes de emisión industrial, ambas estaciones cumplen con el supuesto.

FASE 2 - COMPRENSIÓN DE LOS DATOS:

- Análisis de los datos suministrados por el IDEAM: Concentración PM 2.5, Velocidad del viento, Temperatura del aire, Precipitación, Humedad relativa, Radiación solar, Año, Mes, Día del mes y datos de la Estación de monitoreo ambiental.

FASE 3 - PREPARACIÓN DE LOS DATOS:

- Evaluación de la calidad y limpieza de los datos suministrados por el IDEAM.
- Preparación de los datos para el modelo analítico: Este script toma las variables entregadas por el IDAM y calcula otras consistentes en el valor máximo, mínimo y promedio de las variables meteorológicas, además de la concentración de PM2.5 por franja horaria.
- Preparación de los datos para el sistema de modelamiento AERMOD: se preparan los datos meteorológicos.

FASE 4 - MODELADO:

- Definición de escenarios: Con los datos preparados se realiza un análisis de clustering para identificar los escenarios o días típicos para cada una de las estaciones seleccionadas. Para esto se emplea la técnica k-means.
- Análisis de dispersión por escenario: Por cada escenario identificado en el análisis de clustering, con la herramienta AERMOD se calcula la dispersión del PM2.5 lo que indica el área de influencia, que es el área cercana a la estación donde el nivel de concentración del contaminante es alto.

FASE 5 – EVALUACIÓN:

- Estimación de la calidad y grado de asertividad de los modelos creados. Se emplea el índice de silueta para evaluar la agrupación lograda con el algoritmo K-means.

FASE 6 – DESPLIEGE:

- Análisis de cada escenario o cluster identificado.
- Análisis de demografía por escenario: Según la dispersión calculada por AERMOD o área de influencia, se define a través del geoportal del DANE, la población afectada, con los resultados extrapolados al año 2018.

7. Componentes del funcionamiento

Los principales componentes del prototipo son:

- A partir de los datos meteorológicos suministrados por el IDEAM se identifican los escenarios o días típicos para cada una de las dos estaciones seleccionadas. Para esto se emplean técnicas de analítica de datos.
- Para los escenarios identificados se calcula la dispersión de contaminantes empleando el sistema de modelamiento AERMOD. La dispersión de contaminantes indica el área de influencia que es el área cercana a la estación donde el nivel de concentración de contaminantes es alto.
- Con ayuda del portal de DANE se calcula la población que habita dentro del área de influencia, que correspondería a la población afectada.

8. Validación y Pruebas

Los valores de las variables para los grupos encontrados para la estación Museo de Antioquia se presentan en la Tabla.

	month	day_of_week	avg_WS	max_WD	sum_rain	max_P	avg_H	max_radiation	avg_pm25
1	poco ene, jun, jul, dic	-	0.90	349.36	4.84	637.22	64.30	817.71	26.64
2	-	-	0.85	351.17	3.75	638.43	67.17	477.22	26.53
3	no está enero	-	0.94	349.59	3.60	636.98	63.14	712.83	24.71
4	no está feb y sept. Poco dic	-	0.71	349.56	5.76	637.14	71.80	285.96	27.59
5	-	-	0.95	350.72	4.49	637.16	64.63	599.24	24.73

Tabla. Valores de las variables para los grupos encontrados para la estación Museo de Antioquia.

Los valores de las variables para los grupos encontrados para la estación Casa de Justicia de Itagüí se presentan en la Tabla.

	month	day_of_week	avg_WS	max_WD	sum_rain	max_P	avg_H	max_radiation	avg_pm25
1	poco ene, jun, jul, dic	-	0.90	349.36	4.84	637.22	64.30	817.71	26.64
2	-	-	0.85	351.17	3.75	638.43	67.17	477.22	26.53
3	no está enero	-	0.94	349.59	3.60	636.98	63.14	712.83	24.71
4	no está feb y sept. Poco dic	-	0.71	349.56	5.76	637.14	71.80	285.96	27.59
5	-	-	0.95	350.72	4.49	637.16	64.63	599.24	24.73

Tabla. Valores de las variables para los grupos encontrados para la estación Casa de Justicia de Itagüí.

Para estimar la calidad y grado de asertividad del proceso de *clustering* empleamos el Índice de Silueta. El índice de silueta es un valor entre 0 y 1 y para obtener una buena agrupación el índice de la silueta debe ser al menos de 0.5. En la Tabla se presenta el índice de silueta para cada una de las estaciones, obteniendo 0.50 y 0.48, indicando que la agrupación obtenida es adecuada.

	Número de *clusters* por el método del codo	Índice de Silueta
Estación Museo de Antioquia	5	0.50
Estación Casa de justicia	5	0.48

Tabla. Número de grupos e índice de silueta para cada una de las estaciones.

Se realizó una en uno de los *clusters* para obtener el error en el cálculo de la población al emplear radios de tamaño único en todos los *clusters* y se obtuvo un valor del 17.8 %.

Entrega e implementación

1. Implementación

La implementación del prototipo se realizó de acuerdo a las fases declaradas de acuerdo a la metodología CRISP-DM:

FASE 1 - COMPRENSIÓN DEL PROBLEMA

FASE 2 - COMPRENSIÓN DE LOS DATOS

FASE 3 - PREPARACIÓN DE LOS DATOS

FASE 4 - MODELADO

FASE 5 - EVALUACIÓN

FASE 6 - DESPLIEGE

2. Entrega

Se entrega lo siguiente:

- Documento Técnico con la descripción de la aplicación de la metodología CRISP-DM.
- Descripción de los escenarios de cada estación, que se encuentran en el documento técnico:
  - Resultados del clustering (sección 3.4.1 del informe técnico).
  - Gráfico de la dispersión de cada escenario (AERMOD) (sección 3.4.2 del informe técnico).
  - Descripción de la población afectada en cada escenario (DANE) (sección 3.6.2 del informe técnico).
- Archivos en R para:
  - Preparación de datos.
  - Creación de clustering.
  - Selección de registros más representativos.
- Archivos de entrada y salida a AERMOD.

3 Anexos

Se anexa lo siguiente:

- Documento Técnico con la descripción de la aplicación de la metodología CRISP-DM.
- Scripts en R.
- Archivos en Excel con el análisis demográfico para las dos estaciones.

- Archivos de entrada y salida a AERMOD.