Proyecto #2 – Bodegas de datos y tableros de Control

Universidad de los Andes Departamento de Ingeniería de Sistemas y Computación
ISIS 3301 Inteligencia de Negocios
202320

Andrés Francisco Borda 201729184
Juan Pablo Lora Hernández 202012524
Gabriela Vargas Rojas 202013830

Contenido

1.1 Identificar necesidades analíticas

Luego de la charla realizada con nuestro compañero de medicina y la discusión sobre el enfoque podíamos tener en este proyecto, el tema analítico seleccionado fue el Impacto de las basuras en la presencia de enfermedades respiratorias (EPOC, Asma y enfisema) y su influencia en el estado de salud general de las personas. De esta manera, los análisis inferidos fueron con datos demográficos, el manejo de las basuras para los diferentes casos y la presencia de la enfermedad en la población. También se tuvo en cuenta el aspecto de las localidades para enfocar cuales de estas en Bogotá tienen un peor manejo y por ende se presenta una alteración en el estado de salud de las personas.

Se llegó a la consolidación de este análisis tras una revisión del estado del arte fundamentado en:
El estudio del impacto de las basuras en la salud respiratoria ha ganado relevancia en la literatura científica y médica en los últimos años. Diversas investigaciones han establecido una relación directa entre la gestión inadecuada de residuos y el aumento de enfermedades respiratorias crónicas, como EPOC, asma y enfisema. Esta elección analítica se fundamenta en las siguientes razones respaldadas por estudios previos:

Contaminación del Aire:

Numerosos estudios han demostrado que la acumulación de basuras contribuye significativamente a la contaminación del aire, liberando partículas finas y sustancias químicas perjudiciales. La exposición prolongada a estos contaminantes se ha asociado directamente con el desarrollo y la exacerbación de enfermedades respiratorias.

Demografía y Vulnerabilidad:

La susceptibilidad a las enfermedades respiratorias puede variar según factores demográficos como la edad, el género y las condiciones socioeconómicas. Analizar datos demográficos en conjunto con la gestión de residuos proporciona una comprensión más completa de cómo estas variables pueden interactuar y afectar la salud respiratoria de la población.

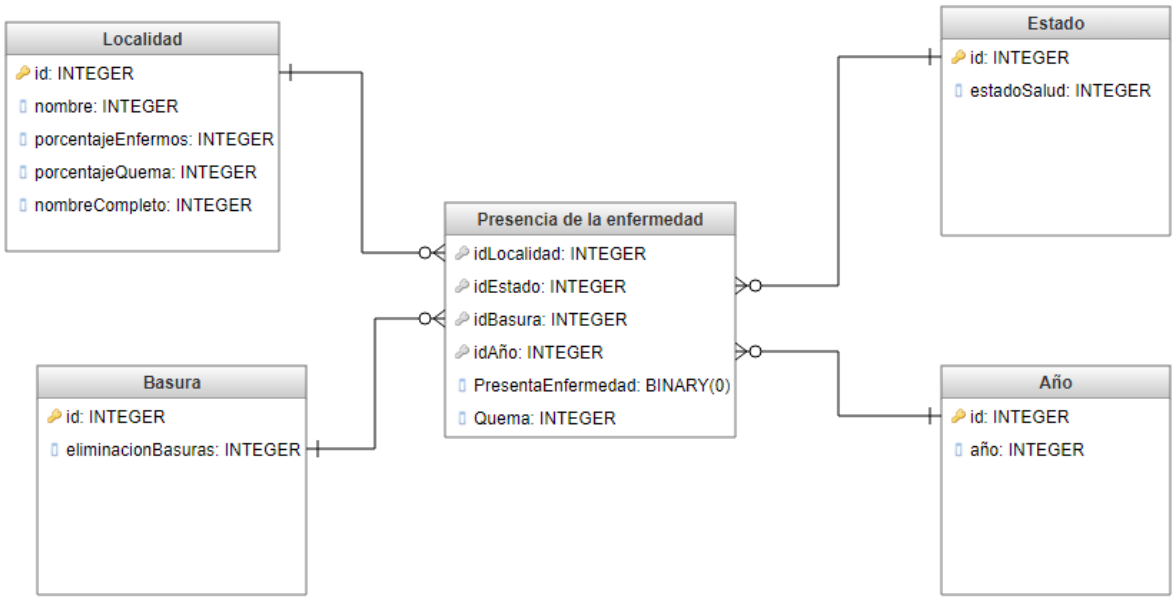
Localidades y Disparidades Ambientales:

Existen disparidades significativas en la gestión de residuos entre diferentes localidades urbanas. Al enfocarse en las áreas específicas de Bogotá con un manejo deficiente de basuras, se puede identificar cómo estas disparidades contribuyen a diferencias en la salud respiratoria de la población, brindando información valiosa para intervenciones y políticas específicas.

Para el resto de los análisis que se plantearon, estos se encuentran en el Excel subido al repositorio del proyecto llamado “Plantilla-AnalisisRequeridosyETL”

1.2 Modelar Data Marts

A partir del análisis propuesto, para satisfacer la pregunta planteada se propuso el siguiente modelo multidimensional:



Así mismo se propuso el siguiente indicador, relacionado con la relevancia del análisis a nivel de la práctica clínica y la resolución de la práctica:

- ✓ Regresión Lógica Multinomial: Se tomaron los datos completos y se mapearon las diferentes variables para realizar el modelo de regresión multinomial. Esto se hizo con el objetivo de hallar una relación entre el impacto que tiene en el estado de salud de las personas la interacción con

la quema de basura, obteniendo un resultado que muestra una relación significativa del 5% entre estas variables, tomando como base la categoría del estado de salud como “Muy bueno”, “Bueno” y pasando a “Normal” o “Regular”. Toda la prueba estadística fue realizada en el archivo que se encuentra en el repositorio y tiene por nombre “Prueba Estadística.ipynb”

Respecto al modelo multidimensional propuesto, tenemos la tabla de hechos “Presencia de la Enfermedad” donde asignamos una granularidad media.

La elección de una granularidad media en la tabla de hechos se justifica por la necesidad de realizar análisis más detallados sin comprometer la complejidad del modelo multidimensional. En este contexto específico, donde se analiza el impacto de las basuras en la presencia de enfermedades respiratorias, la granularidad media ofrece un equilibrio entre la profundidad de los datos y la capacidad de interpretación. Esto se observa en la variabilidad del estado general de salud de las personas en una localidad determinada a través de los años 2017 y 2021 con las respectivas influencias del impacto de la basura.

En el caso de las medidas encontramos “Presenta Enfermedad” la cual es aditiva ya que permite sumar o contar la presencia de enfermedades respiratorias a lo largo de todas sus dimensiones y por otra parte, la medida “Quema” que también seria de tipo aditiva al sumar en la dimensión y poder establecer un contraste con los que tienen problemas de salud y realizan está práctica.

Por otra parte, respecto a las dimensiones encontramos localidad, basura, estado y año, las cuales se describen a continuación:

- ✓ Localidad: En este caso tiene un id como llave primaria que es su identificador único, el nombre que corresponde a una de las localidades de Bogotá sumado a algunas catalogadas como “Otras localidades rurales” o simplemente “Otras localidades”. Por otra parte, esta dimensión también contiene un atributo para indicar el porcentaje de enfermos, un porcentaje de personas que queman basura y un nombreCompleto que hace referencia a Bogotá “Nombre localidad”, esto se implementó con el fin de poder generar un mapa por localidades en el tablero de control . Estos fueron seleccionados acorde al grado de detalle que se quería tener dentro del análisis a nivel de localidad.
- ✓ Basura: Compuesta de un id que funciona como llave primaria y un atributo de tipo entero “eliminacionBasuras” que indica de qué forma las personas se deshacen de la basura según el catalogo que está definido en la encuesta multipropósito. Esta selección hace parte de uno de los componentes principales para el análisis donde podemos observar cómo es la relación de la basura en Bogotá
- ✓ Año: Esta dimensión fue creada para la visualización de cómo ha evolucionado el análisis de este impacto en Bogotá del año 2017 hacia el año 2021 con los datos proporcionados relacionados a estos. De esta forma

tenemos un id como llave primaria y un atributo “año” que indica si es de 2017 o 2021.

- ✓ Estado: Está dimensión hace referencia al estado de salud general de una persona y lo establecimos para hacer la comparación de la presencia de la enfermedad impactada por la basura y cómo se ha visto peor su estado de salud tras estar en esa situación de enfermo. De esta manera, también posee un id como llave primaria y un atributo “estadoSalud” de tipo entero que nos indica según los datos de la encuesta en que categoría está o que tan bueno es en esa persona.

1.3 Entendimiento de los datos, creación del Data mart y proceso ETL

Para el entendimiento de los datos, tomamos las referencias tanto de los años 2017 como 2021 y los cargamos a modo de DataFrame en Pandas. Cabe aclarar que creamos dos archivos.csv únicamente con las columnas que requerimos para realizar el análisis. De esta manera, se inició el proceso de perfilamiento y entendimiento para el análisis de calidad de estos.

Para el año 2017 tenemos 319952 registros de persona, repartidos en este caso en 6 columnas, 5 numéricas y una categórica que es la localidad. Sin embargo, cabe destacar que las numéricas están representadas así acorde a las posibilidades ofrecidas en la encuesta multipropósito.

Para el año 2021 tenemos 253103 registros de personas, repartidos en 4 columnas, donde 3 son numéricas y nuevamente una categórica que es la localidad. En este caso también, las numéricas están representadas así acorde a las características de la encuesta multipropósito.

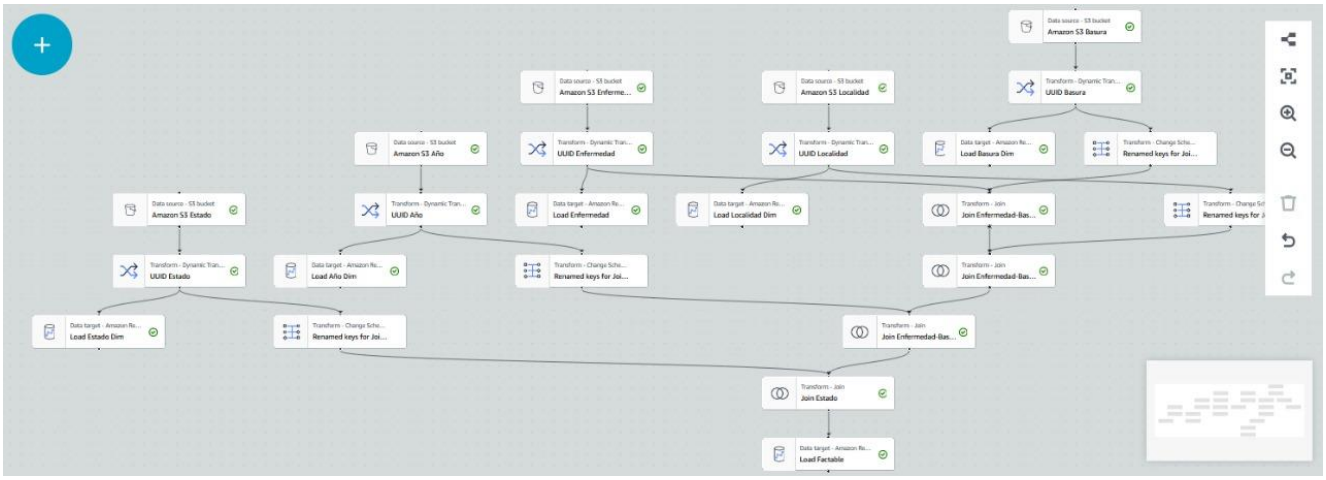
Respecto al análisis de calidad, para ambos conjuntos de datos se les aplicó completitud, unicidad, consistencia y validez obteniendo lo siguiente:

- ✓ Completitud: El porcentaje de valores nulos fue de 0% para cada una de las columnas extraídas del conjunto de datos original y esto aplica para ambos registros de los años 2017 y 2021
- ✓ Unicidad: Para el caso de los registros duplicados, tanto para el archivo del 2017 como el del 2021 no existe ningún caso por lo tanto tiene un valor de 0.
- ✓ Consistencia: Para el caso de consistencia revisamos las variables de mayor importancia para el análisis como Basura, Personas por localidad y Enfermedades Respiratorias. De esta manera, no se encontraron errores de valores similares o de digitación
- ✓ Validez: En el caso de la validez, al ser datos de una fuente del DANE y con la ayuda del diccionario de datos, todo este conjunto se revisó y no se encontraron valores fuera de lo normal para ningún caso. Esto se ve reflejado en los datatypes y en la carga de los datos hacia PowerBi donde no hubo errores.

Para más información sobre este apartado, todo el trabajo realizado se encuentra en el notebook entregado en el repositorio que tiene de nombre “Proyecto2.ipynb”

El diseño del ETL se encuentra en el archivo “Plantilla-AnalisisRequeridosyETL” acorde al modelo propuesto.

Para el proceso del ETL, fue desarrollado en AWS a partir de la creación de los componentes del Data Marts, separados en archivos CSV que fueron cargados en la plataforma.



1.4 Propuesta Arquitectura de solución

Para la solución de nuestro análisis, decidimos implementar un tablero de control en PowerBI a partir de lo solicitado, en un primer lugar, para la presentación de nuestro compañero de medicina, donde separamos este en una primera introducción sobre los datos que se están trabajando tanto de 2017 como de 2021 y porque fueron seleccionados para responder nuestra pregunta.

Estos fueron separados respectivamente para cada uno de los años y a partir de un filtro de localidades podemos distinguir cuántas personas presentan enfermedades respiratorias y de qué manera estas eliminan su basura.

Por otra parte, entrando de lleno en la problemática, se obtuvieron los porcentajes de quema de basura por localidad y el porcentaje de enfermos por localidad. Esto también se ve reflejado a través de un mapa en el que se muestran estas afectaciones en las localidades que más porcentaje de ambas variables presentan.

Otro de los detalles importante dentro de este tablero antes de pasar al análisis final, es la separación de personas enfermas y sanas con una eliminación inadecuada de basuras siguiendo el filtro por localidades para ambos años.

Finalmente, para nuestros hallazgos finales se hizo un recuento por localidad de todas las personas que queman basura y están enfermas y que queman basura y están sanas, filtrados por su estado de salud general teniendo los posibles de “Muy bueno”, “bueno”, “Regular”, “Malo” y “Muy malo”. Esto se contrasta con la prueba estadística al obtener una significancia del 5% cuando hay una quema de basura y que se presenta una desmejora del estado de salud de las personas que lo hacen o comparten un entorno en el que ocurre este evento.

La implementación de este tablero de control, se encuentra en el repositorio bajo el nombre de “MédicoBi.pbix”

1.5 Vídeo Resultados

El vídeo detallado de nuestra solución propuesta se encuentra en el siguiente enlace:

1.6 Evaluación trabajo en equipo

a. Andrés Borda: Mi compromiso para este proyecto fue muy alto y siento que en medio de las dificultades que surgieron con el cronograma y el trabajo interdisciplinar, las resolví de manera adecuada siempre comunicándome con mis compañeros y intentando resolver los problemas de la mejor forma posible
   Juan Pablo Lora: En ningún momento dejé mis responsabilidades con el equipo e intenté desarrollar el mejor producto posible trabajando de una muy buena forma de manera autónoma y siempre apoyando a mis compañeros de equipo, a pesar de las dificultades que surgieron respecto a las fechas de entrega y la comunicación y definición de horarios para trabajar con nuestro compañero de medicina
   Gabriela Vargas Rojas: Siempre mantuve un buen nivel de compromiso con el proyecto, intentando estar pendiente del desarrollo de mis compañeros y facilitando la división de tareas para el trabajo autónomo. De esta forma, considero que tanto yo como mi grupo trabajamos de la mejor forma que pudimos a lo largo del desarrollo de este proyecto.

b. De acuerdo con los comentarios recibidos tanto en la sustentación con el estudiante de medicina como para la presentación realizada el viernes 1 de diciembre para los estudiantes de ingeniería, recibimos muy buenos comentarios de nuestro proyecto por parte del invitado de medicina el cual nos indicó que nuestros hallazgos eran muy valiosos y debían ser debidamente documentados para su futuro uso. De esta manera y sumado al feedback recibido por Nicolas Briceño (nuestro compañero de medicina), consideramos que el proyecto entregado es de muy buena calidad y

realizamos un muy buen trabajo interdisciplinar a pesar de todas las dificultades que surgieron durante el desarrollo de este. Esto también va de la mano con la posibilidad de responder a la pregunta planteada desde un inicio con el análisis que escogimos y su relevancia a nivel clínico con la prueba estadística realizada.

c.

Estudiantes	Horas trabajadas	Tareas hechas	Desafíos	Puntaje
Andrés Borda	40	Establecer el enfoque analítico – Entendimiento de los Datos – Preparación de los datos - Modelo Multidimensional – Generación Tableros de Control – Apoyo Estudiante de Medicina- Creación ETL y Base de Datos -Documento - Vídeo	Acomodación de tiempos debido al cronograma desorganizado – Problemas con el desarrollo del ETL en AWS	33,33
Juan Pablo Lora	40	Establecer el enfoque analítico – Entendimiento de los Datos – Preparación de los datos - Modelo Multidimensional – Generación Tableros de Control – Apoyo Estudiante de Medicina- Creación ETL y Base de Datos -Documento - Vídeo	Acomodación de tiempos debido al cronograma desorganizado - Problemas con el desarrollo del ETL en AWS	33,33
Gabriela Vargas	40	Establecer el enfoque analítico – Entendimiento de los Datos – Preparación de los datos - Modelo Multidimensional – Generación Tableros de Control – Apoyo Estudiante de Medicina- Creación ETL y Base de Datos -Documento - Vídeo	Acomodación de tiempos debido al cronograma desorganizado - Problemas con el desarrollo del ETL en AWS	33,33

1.7 Referencias

*DANE. (*2017). “Encuesta Multipropósito – EM 2017”. Recuperado de: https://microdatos.dane.gov.co/index.php/catalog/565/data-dictionary/F36?file_name=Salud%20(capítulo%20F)

*DANE. (*2021). “Encuesta Multipropósito Bogotá - Cundinamarca - EM 2021”. Recuperado de: https://microdatos.dane.gov.co/index.php/catalog/743/data-dictionary