Mayo 29 , 2023 Universidad de los Andes

Andres Molina – 201923434

Josue Rivera –201914138

Jaime Alfonso- 202116525

Proyecto 2 Inteligencia de Negocios

1. [10%] Identificar necesidades analíticas

Tema analítico	Análisis requeridos o inferidos	Categoría del análisis - Tablero de control, análisis OLAP, Minería de datos	Procesos de negocio	Fuentes de datos y datos
Prevalencia de asma en relación con la presencia de mascotas en el hogar en Bogotá	Comparar las tasas de prevalencia de asma en hogares con y sin mascotas (gatos o perros) en todas las localidades de Bogotá	Tablero de control.	Análisis epidemiológico	Datos demográficos y médicos de los estudios citados y de la base de datos del DANE.
Prevalencia de asma en diferentes rangos de edad en relación con la presencia de mascotas en el hogar	Analizar la prevalencia de asma en diferentes grupos de edad y comparar los resultados entre hogares con y sin mascotas	Análisis OLAP	Investigación Seguimiento médico	Datos demográficos y médicos de los estudios citados y de la base de datos del DANE.
Cambio en la prevalencia de asma en los últimos años en relación con la presencia de mascotas en el hogar	Rastrear los cambios en las tasas de prevalencia de asma a lo largo del tiempo y analizar cualquier correlación con la presencia de mascotas en el hogar	Minería de datos	Vigilancia de la salud pública	Datos históricos de prevalencia de asma y datos demográficos y médicos de los estudios citados y de la base de datos del DANE.
Prevalencia de asma en relación con el estrato y la localidad de la persona	Comparar las tasas de prevalencia de asma en hogares dependiendo de su estrato socioeconomico y la localidad de bogota en el que se encuentra ademas de saber que tanto influye el entorno en relacion con la cantidad de afectados	Tablero de control	Evalucion de encuesta multiproposito en condiciones de vida. Seguimiento medico	Datos históricos de prevalencia de asma y datos demográficos y médicos de los estudios citados y de la base de datos del DANE.

Con respecto a la tabla, explicaremos a continuación cada uno de los Temas Analíticos y por qué los elegimos.

1. Prevalencia de asma en relación con la presencia de mascotas en el hogar en Bogotá: Este análisis busca entender la relación entre la presencia de mascotas en los hogares (específicamente gatos y perros) y la prevalencia de asma en Bogotá. Con los datos proporcionados, se construirá un tablero de control que permitirá a los expertos visualizar esta relación de manera clara y concisa, contribuyendo a un mejor entendimiento de cómo la interacción con estos animales puede influir en el desarrollo del asma.

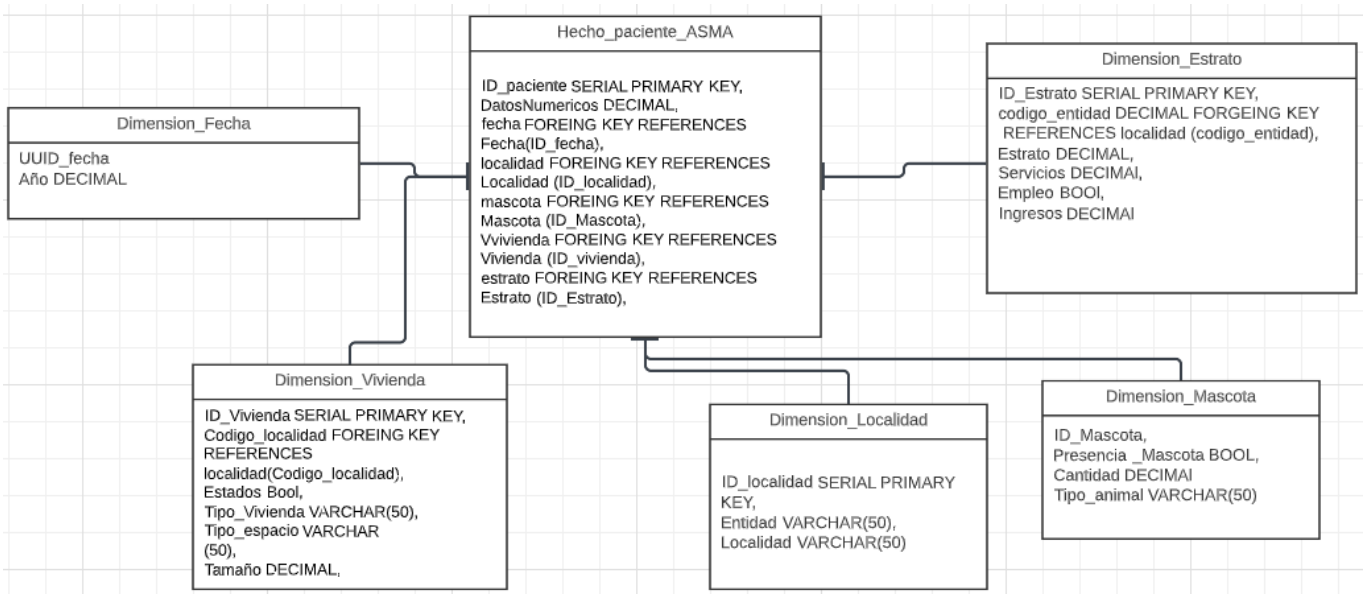
2. Prevalencia de asma en diferentes rangos de edad en relación con la presencia de mascotas en el hogar: Este análisis buscará segmentar la población en diferentes grupos de edad para evaluar si la presencia de mascotas en el hogar tiene un impacto diferente según la edad de los individuos. Este análisis se realizará a través de operaciones OLAP, lo que permitirá a los investigadores profundizar en la información de manera rápida y flexible.

3. Cambio en la prevalencia de asma en los últimos años en relación con la presencia de mascotas en el hogar: Este análisis utilizará técnicas de minería de datos para investigar si las tasas de prevalencia de asma han cambiado a lo largo del tiempo y si estos cambios están correlacionados de alguna manera con la presencia de mascotas en los hogares. Este análisis ayudará a la vigilancia de la salud pública a entender mejor las tendencias y patrones relacionados con el asma. En este caso se hará la comparación directa entre el año 2017 y el año 2021, que fueron las fuentes de datos que se nos suministraron para hacer los análisis.

4. Prevalencia de asma en relación con el estrato y la localidad de la persona: Este análisis buscará entender si existe una relación entre el estrato socioeconómico de los hogares, la localidad en Bogotá y la prevalencia de asma. Este análisis se presentará en un tablero de control y se utilizará para evaluar las condiciones de vida y cómo podrían estar relacionadas con la prevalencia del asma.

En general, estos análisis buscan proporcionar una visión más profunda de la prevalencia del asma y cómo podría estar relacionada con varios factores como la presencia de mascotas en el hogar, la edad de los individuos, el estrato socioeconómico y la ubicación en Bogotá. Esta información será de gran valor para los expertos médicos y los responsables de la toma de decisiones en el ámbito de la salud pública.

2. **(15%) Modelar Data Marts:**
   a. (7,5%) Elaborar los modelos dimensionales propuestos. Entregue una representación gráfica de los modelos multidimensionales. En el modelo se deben representar nombres de atributos, llaves primarias, llaves foráneas y roles. Evite diagramas ilegibles



   b. (7,5%) Justificación del modelo Para cada tabla de hechos
      i. (2%) Especificar y justificar la granularidad

La granularidad de esta tabla de hechos se establece en el nivel de detalle de los datos relacionados con los pacientes asmáticos. Según la información disponible, la granularidad es a nivel de paciente asmático por cada combinación de fecha, localidad, mascota, vivienda y estrato, junto con el UUID para identificar de manera única cada registro. Esta elección permite un análisis detallado y la capacidad de responder preguntas específicas a nivel individual y contextual.

      ii. (3%) Definir los hechos/medidas que contiene. Para cada medida indicar y justificar el tipo de medida (aditiva, semi-aditiva y no aditiva).

DatosNumericos: Esta medida podría representar diferentes datos numéricos relacionados con los pacientes asmáticos. Dado que la medida puede acumularse y agregarse en función de las dimensiones, se considera una medida aditiva.

      iii. (2%) Para cada atributo, si se requiere, especificar el tipo de manejo de historia (1,2, 3, …) de variación lenta (Slowly Changing Dimension), y justificar la elección.

fecha: Normalmente, la dimensión de fecha no requiere manejo de historia, ya que representa una serie de puntos en el tiempo y no suele cambiar. No se justifica un manejo de historia en este caso.

localidad: Los atributos de localidad pueden cambiar con el tiempo (por ejemplo, cambios en los nombres de las localidades o cambios en los códigos), sería necesario aplicar un manejo de historia. En el caso de un modelo multidimensional, el tipo de manejo de historia puede variar dependiendo de los requisitos específicos del negocio y el análisis. Esto podría implicar la implementación de Slowly Changing Dimensions (SCD) de tipo 1, 2 o 3, según los cambios necesarios y la importancia de mantener el historial.

mascota: Si los atributos de mascota pueden cambiar con el tiempo (por ejemplo, cambios en la presencia o cantidad de mascotas), se podría aplicar un manejo de historia. Al igual que en el caso anterior, el tipo de manejo de historia puede variar según los requisitos y la importancia de mantener el historial.

vivienda: Si los atributos de vivienda pueden cambiar con el tiempo (por ejemplo, cambios en los estados, tipo de vivienda, tamaño, etc.), se podría requerir un manejo de historia. Al igual que en los casos anteriores, el tipo de manejo de historia puede variar según los requisitos y la importancia de mantener el historial.

estrato: Si los atributos de estrato pueden cambiar con el tiempo (por ejemplo, cambios en los servicios, empleo, ingresos, etc.), se podría requerir un manejo de historia. Del mismo modo, el tipo de manejo de historia puede variar según los requisitos y la importancia de mantener el historial.

3. **(30%) Entendimiento de los datos, creación del Datamart y proceso ETL.**
   a. (15%) Entender las fuentes de datos recibidas y presentar el resultado del análisis (estadísticos de los datos y análisis de calidad de datos).

Para el caso de las fuentes de datos fue algo complejo, ya que eran bastantes datos repartidos. Para empezar, teníamos los datos de la encuesta multipropósito de la gente que tiene asma de los años 2017 y 2021, pero también teníamos los archivos de la encuesta multipropósito de la gente no asmática de los mismos años. Comprender los datos fue algo complicado ya que al tener tantas columnas sin nombres significativos y solo códigos se dedico demasiado tiempo al entendimiento del diccionario de datos de cada uno de los archivos. Además, es claro que no todos los datos nos sirven.

Para poder limpiar los datos utilizamos un jupyter notebook, esto para poder ver que venia en cada uno de los archivos y así poder ver que tenemos que cambiar. Cabe recordar que para cada uno de los archivos se tenía que cumplir con los requerimientos planeados al inicio del curso con respecto a la preparación de datos. Verificamos, completitud, unicidad, duplicidad, consistencia y validación de los datos.

Los datos que recibimos no tenían mayores problemas, venían bien organizados y en su gran mayoría muy completos. Solo encontramos unos detalles como los valores “NA” que encontramos en algunas celdas. Para esto solo las eliminamos ya

que al tener gran cantidad de datos útiles y muy pocos datos nulos, lo mas sencillo era eliminarlos. Por otro lado, como las columnas no tenían nombres significativos, a todas las columnas que se consideraban importantes se les cambio el nombre a un nombre significativo como “Estrato” o “Tiene_perro”.

Una vez implementado esto se podía trabajar exitosamente con los datos ya que no contábamos con mayores problemas de preprocesamiento. Por último, consideramos que la muestra de datos era muy amplia, lo que permite tener resultados precisos ante los análisis que estamos empleando.

b. (15%) Diseñar e implementar el proceso de ETL. En este punto debe entregar el diseño del ETL y su implementación. A nivel del diseño utilice la plantilla compartida en Excel.

El diseño del ETL se encuentra en nuestro repositorio de GitHub que es el que se entrega por bloque neón como la entrega del proyecto donde encontraran todos los archivos solicitados en la rúbrica del proyecto.

4. **(30%) Proponer la arquitectura de solución**
   a. (8%) Proponer la arquitectura de solución de BI para resolver los análisis realizados hasta este momento en el proyecto.

Para poder cumplir con los requerimientos de negocio que pidió el cliente, como grupo planteamos una arquitectura que le permite al cliente un acceso sencillo a los datos. En este caso, se propone crear un ETL que tiene como función principal preparar y organizar los datos para que el acceso a los mismos sea sencillo. Por otro lado, la base de datos que se utilizo es la , que nos permite ventajas de desempeño, accesibilidad y seguridad para el cliente. Con respecto al ETL, utilizamos AWS, que es líder en el mercado y su confiabilidad es supremamente alta. No solo esto, sino que AWS tiene múltiples ventajas para el cliente, como su soporte y su facilidad de uso con respecto a los datos.

Una vez implementado todo lo mencionado anteriormente, se conectan los datos que se trabajaron con el tablero de control. Utilizamos los .csv dados en el curso en el ETL, ya que por el momento es la manera mas sencilla de poner los datos ya que en caso de que el cliente requiera algún cambio seria sencillo mediante el notebook que generamos para el procesamiento de los datos, ya que adquirir un servidor sin haber consultado antes al cliente sería un grave error.

Para desplegar los datos en un tablero de control utilizamos Power BI, ya que fue una de las herramientas que utilizamos en el desarrollo del curso y consideramos que es muy fácil de entender y utilizar, ya que tiene muchas funcionalidades que ante la cantidad de datos que tenemos le será de gran utilidad al cliente. El tablero de control está diseñado para mostrar los datos de l

b. (22%) Implementar los tableros de control, utilizando un software especializado como PowerBI, Tableau, Looker, etc. y conectándose a la base de datos donde tiene los datos que representan el modelo multidimensional propuesto.

La implementación de los tableros de control esta en el repositorio de GitHub que es el que se entrega por bloque neón como la entrega del proyecto donde encontraran todos los archivos solicitados en la rúbrica del proyecto.

5. (10%) Preparar un video

El video esta subido en el padlet tal como lo pide la rúbrica.

6. [5%] Describir las actividades realizadas

En cuanto a la repartición de los 100 puntos, consideramos que estos deben ser repartidos de manera equitativa entre los integrantes del grupo.

- Líder de proyecto: Andres Molina

Andrés Molina es el líder del proyecto y está a cargo de la gestión de este. Ha definido las fechas de reuniones y pre-entregables del grupo, y ha verificado las asignaciones de tareas para que la carga sea equitativa. Andres, estuvo pendiente por medio del grupo de WhatsApp para que se empezara a trabajar el proyecto con antelación. Se ha encargado de subir la entrega del grupo y ha tomado decisiones cuando no ha habido consenso, siempre buscando el beneficio del proyecto y del equipo. Andrés ha sido un líder eficiente y ha fomentado un ambiente colaborativo y productivo entre los miembros del equipo. En cuanto a las actividades, Andrés realizó el entendimiento y procesamiento de los datos, también tuvo participación en la implementación de los tableros de control. Andres, le dedico un total de 8 horas al proyecto, teniendo en cuenta las reuniones previas y la preparación del trabajo antes de arrancar con el total del proyecto.

- Líder de datos: Josue Rivera

El líder de datos en el proyecto se llama Josue Rivera. Su función principal es manejar y organizar los datos que se utilizarán en el proyecto, asegurándose de que estén disponibles para todos los miembros del equipo. Él también es responsable de asignar tareas relacionadas con el manejo y análisis de datos, y de asegurarse de que se completen dentro de los plazos establecidos. Josue trabaja en estrecha colaboración con el líder del proyecto y otros miembros del equipo para garantizar que los datos sean precisos, completos y relevantes para el proyecto. Además, se asegura de que se cumplan los estándares de privacidad y seguridad de los datos en todo momento.

Josué también es el líder de analítica del equipo, su responsabilidad es gestionar las tareas de análisis de datos y asegurarse de que los entregables cumplan con los estándares requeridos. Josué trabaja en conjunto con los otros líderes del equipo para identificar las mejores prácticas de análisis y asegurar que se estén aplicando correctamente en el proyecto. Además, Josué es el encargado de verificar que el equipo esté utilizando las herramientas y metodologías adecuadas para analizar los datos y encontrar la solución más efectiva al problema planteado.

En el caso de las actividades, Josué se encargo de diseñar el ETL del proyecto y también participo activamente en la implementación del tablero de control. Josué le dedico alrededor de 8 horas al proyecto, teniendo en cuenta las reuniones previas y la preparación del trabajo antes de arrancar con el total del proyecto.

- Líder de negocio: Jaime Alfonso

Jaime es el líder de negocio en el proyecto. Se asegura de que el problema o la oportunidad identificada se resuelva de manera efectiva y que esté alineado con la estrategia de negocio para el cual se está llevando a cabo el proyecto. También es responsable de garantizar que el producto se pueda comunicar adecuadamente y que se cumplan los objetivos del proyecto. Jaime tiene el papel crucial de contactar al grupo de expertos en estadística para evaluar el modelo, tanto a nivel cuantitativo como cualitativo, para garantizar la calidad del producto final. Además, trabaja en estrecha colaboración con el equipo técnico y el líder de proyecto para asegurarse de que el proyecto esté avanzando de acuerdo con los objetivos establecidos.

En el caso de las actividades Jaime realizo la implementación el ETL y participo en la implementación del tablero de control. Jaime le dedico alrededor de 8 horas al proyecto, teniendo en cuenta las reuniones previas y la preparación del trabajo antes de arrancar con el total del proyecto.