
 <p>Universidad de los Andes Colombia</p>	<p>Ingeniería de Sistemas y Computación Pregrado ISIS-3301 – Inteligencia de Negocios Primer Proyecto Semestre: 2023-20</p>	 <p>Engineering Accreditation Commission</p>
--	---	---

Proyecto 2

Integrantes:

Juan Sebastián Sánchez Delgado - js.sanchezd1

Miguel Cárdenas Cárdenas - ma.cardenas

Santiago Paerez Gonzales – s.paeres

Institución:

Universidad De Los Andes

Curso:

Inteligencia de negocios – ISIS3301 (Sección 1)

Profesores:

Fabian Peña

María del Pilar Villamil

Haydemar Nuñez

1. Identificación de Necesidades Analíticas.....	2
2. Modelamiento de Data Marts	5
2.1 Modelo Multidimensional	5
2.2 Justificación del Modelo	5
3. Entendimiento de los Datos, Creación del Data Mart y Proceso ETL.....	11
3.1 Entendimiento de las Fuentes de Datos.....	11
3.1.1 Completitud	11
3.1.2 Unicidad	11
3.1.3 Consistencia	12
3.1.4 Validez.....	12
3.1.5 Entendimiento de los datos.....	13
3.2 Diseño e Implementación el Proceso de ETL.....	16
4. Arquitectura de Solución y Tableros de Control.....	16
4.1 Arquitectura de Solución Propuesta.....	16
4.2 Implementación de Tableros de Control	16
5. Evaluación de Trabajo en Equipo.....	17
5.1 Autoevaluación en las competencias	17
5.2 Autoevaluación de la calidad y aporte al proyecto entregado	18
5.3 Evaluación entre miembros del equipo.....	18
6. Bibliografías	18

1. Identificación de Necesidades Analíticas

A continuación, se presentan los 4 requerimientos analíticos que se plantearon en colaboración con el grupo de medicina. Los 2 primero que se presentan son aquellos que se pretenden

desarrollar durante la presente entrega del proyecto, mientras que los restantes son sugerencias para posible iteraciones o entregas posteriores que abarquen el mismo tema de interés.

Tema analítico	Análisis requeridos o inferidos	Categoría del análisis - <u>Tablero de control,</u> análisis OLAP, Minería de datos	Procesos de negocio	Fuentes de datos y datos
¿Cuál es la incidencia de las características físicas de la vivienda en la prevalencia del asma en la población de Bogotá?	Cuantificar el nivel de correlación entre la estructura de la vivienda (techo, paredes y piso), así como también de sus materiales en la prevalencia del asma	Análisis OLAP	Implementación de recursos tecnológicos	Material Pared (NVCBP12) Material Piso (NVCBP13) Humedad Techo/Paredes (NVCBP8A) Goteras Techo (NVCBP8B) Grietas (NVCBP8C) Grietas Piso (NVCBP8E) Asma (NPCFP14I)
	Evaluar cual es la relación existente entre el sistema de drenaje del agua con respecto al padecimiento de asma	Tablero de control	Evaluación de correlaciones y resultados finales	Falla tuberías (NVCBP8D) Acueducto (NVCBP11B) Asma (NPCFP14I)
¿Cómo afectan los diferentes tipos de contaminación en la probabilidad de contraer enfermedades respiratorias como el asma?	Estudiar la relación existente entre la contaminación de aire y el desarrollo de asma	Análisis OLAP	Implementación de recursos tecnológicos	Casa con industrias (NVCBP9) Fábricas en barrio (NVCBP14A) Contaminación del aire (NVCBP15D) Escasa ventilación (NVCBP8G) Asma (NPCFP14I)
	Impacto del estado de la contaminación del agua en la prevalencia del asma	Tablero de control	Evaluación de correlaciones y resultados finales	Caños (NVCBP14I) Contaminación agua (NVCBP15I) Asma (NPCFP14I)

Impacto de Factores Socioeconómicos en la Prevalencia del Asma en Bogotá	Evaluar la relación entre variables socioeconómicas (ingreso familiar, nivel educativo) y la prevalencia del asma en la población de Bogotá.	Análisis OLAP	Implementación de recursos tecnológicos	Afiliado (NPCFP1)
	Analizar cómo el acceso a servicios de salud, como atención médica regular y medicamentos, influye en la incidencia del asma.			Régimen de salud (NPCFP2)
Evaluación del Impacto de Estilos de Vida en la Prevalencia del Asma en Bogotá	Analizar cómo los hábitos alimenticios (dieta, consumo de ciertos alimentos) se relacionan con la prevalencia del asma.	Análisis OLAP	Implementación de recursos tecnológicos	Medicina Prepagada (NPCFP10B)
	Evaluar cómo la cantidad y tipo de ejercicio y actividad física impactan en la incidencia y gravedad del asma.			EPS complementaria (NPCFP10C)
				Nivel educativo (NPCHP4)
				Asma (NPCFP14I)
				Bebidas alcohólicas, cigarrillos y tabaco (NHCMP5A)
				Consumo Cigarrillos (NPCFP38)
				Asma (NPCFP14I)
				Actividad Física (NPCHP31GA)
				Alimentación saludable (NHCLP18)
				Nutrición (NHCLP8F)

Tabla 1. Requerimientos analíticos

La selección de los requerimientos analíticos presentados anteriormente se basó principalmente en los intereses y necesidades transmitidas por el grupo de medicina. En concreto, durante la entrevista oral llevada a cabo en la semana 14, los estudiantes de medicina hicieron especial hincapié que querían explorar la correlación de la prevalencia del asma con respecto a otras variables menos convencionales en este tipo de estudios. Esto debido a que ya existe evidencia tangible de la relación existente entre la contaminación del aire o el padecimiento de otras enfermedades respiratorias con respecto al asma.

Tras la entrevista, se llegó a la resolución de centrarse en el estado de la vivienda y el acueducto, así como también de diversos tipos de contaminación en nuestro estudio. Con respecto al estado de arte, la mayoría de los investigadores concuerdan que la contaminación atmosférica exacerba los síntomas del asma, lo cual también sucede si el paciente sufre de otras enfermedades respiratorias [2,3].

Por otro lado, si bien no existen respuestas concluyentes en cuanto a la incidencia de todos los tipos de materiales de interés en el desarrollo y empeoramiento del asma, existen algunos

estudios relevantes acerca de “los materiales asmáticos”, los cuales pueden provocar asma ocupacional en ciertos grupos poblacionales [4]. Algunos de estos materiales son: el látex, los pesticidas, el aserrín, algunos perfumes, el amoníaco y sus derivados, entre otros [5].

2. Modelamiento de Data Marts

2.1 Modelo Multidimensional

A continuación, se presenta el modelo multidimensional diseñado para dar respuesta a los requerimientos analíticos planteados anteriormente:

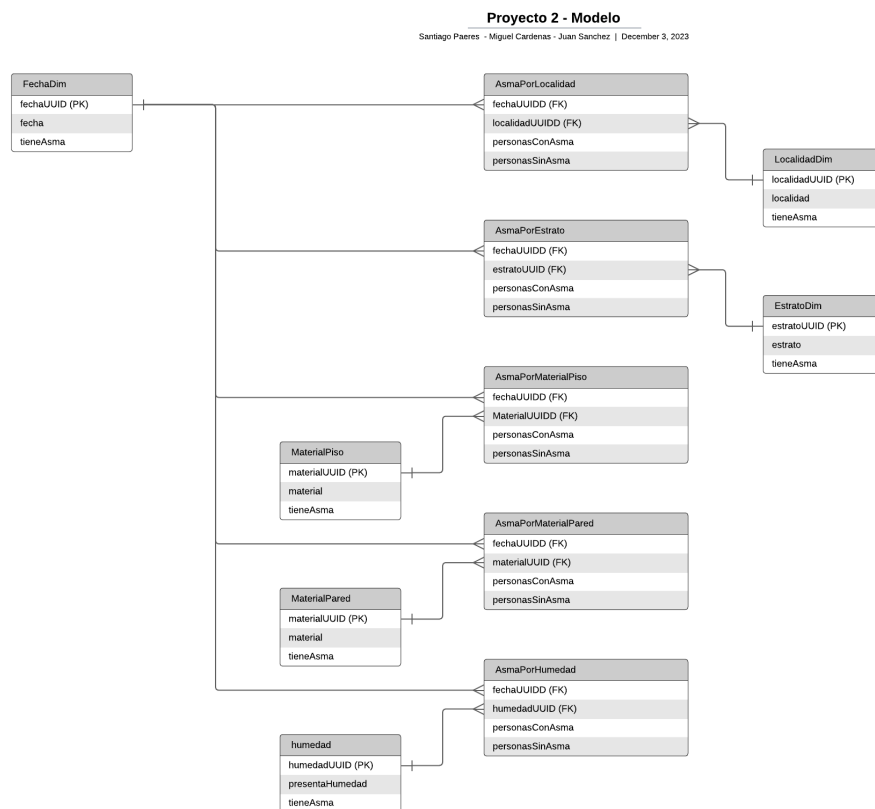


Figura 1. Modelo Multidimensional

El modelo presentado anteriormente cuenta con 5 tablas de hecho, que están directamente relacionadas con los requerimientos analíticos a solucionar. Además, se crearon 6 dimensiones diferentes, cada una relacionada directamente con las diferentes tablas de hecho pertenecientes al modelo. Se decidió crear múltiples tablas de hecho con el objetivo de manejar correctamente las variables definidas como las más importantes por el grupo de medicina, y así evitar tener una única tabla de hecho que aumente en gran medida la complejidad del modelo. A continuación, se realizará una explicación completa de cada tabla de hecho y dimensión asociada.

2.2 Justificación del Modelo

Tablas de hecho:

- **Asma Por Localidad**

Esta tabla permite desarrollar el requerimiento analítico específico para las variables socioeconómicas, dado que permite cuantificar la cantidad de personas que tiene asma respecto una localidad específica, y así encontrar una correlación aparente entre las enfermedades respiratorias (Asma) y la ubicación geográfica de los encuestados. Como se mencionó anteriormente, la granularidad de las tablas es muy detallada, pues se toman solo los datos necesarios para dar una respuesta coherente al requerimiento analítico. En este caso, lo que se espera es encontrar la correlación entre la localidad y el número de personas que padecen Asma, por ende, la tabla únicamente tiene relación con la dimensión localidad, fecha y encuestado. Así mismo, mantiene únicamente 2 medidas necesarias para el requerimiento, descritas a continuación:

Medidas/ Hechos:

Personas con Asma: esta medida permite calcular el número de personas que presentan asma, discriminado por localidad y fecha, considerando que los datos de las encuestas son de dos años diferentes. Es una medida aditiva, pues bien, puede ser agregada a cada una de las dimensiones de la tabla de hecho (Fecha, Localidad).

Personas sin Asma: esta medida calcula el número de personas que no presentan asma de acuerdo a la localidad y fecha (año) específico. Es una medida aditiva, pues bien, puede ser agregada a cada una de las dimensiones de la tabla de hecho (Fecha, Localidad).

- **Asma por Estrato**

Esta tabla permite desarrollar el requerimiento analítico específico para las variables socioeconómicas, dado que permite cuantificar la cantidad de personas que tiene asma respecto un estrato específico, y así encontrar una correlación aparente entre las enfermedades respiratorias (Asma) y una de las condiciones socioeconómicas más importantes de los encuestados. Como se mencionó anteriormente, la granularidad de las tablas es muy detallada, pues se toman solo los datos necesarios para dar una respuesta coherente al requerimiento analítico. En este caso, lo que se espera es encontrar la correlación entre el estrato y el número de personas que padecen Asma, por ende, la tabla únicamente tiene relación con la dimensión estrato, fecha y encuestado. Así mismo, mantiene únicamente 2 medidas necesarias para el requerimiento, descritas a continuación:

Medidas/ Hechos:

Personas con Asma: esta medida permite calcular el número de personas que presentan asma, discriminado por estrato y fecha, considerando que los datos de las encuestas son de dos años diferentes. Es una medida aditiva, pues bien, puede ser agregada a cada una de las dimensiones de la tabla de hecho (Fecha, Estrato).

Personas sin Asma: esta medida calcula el número de personas que no presentan asma de acuerdo a un estrato y fecha (año) específico. Es una medida aditiva, pues bien, puede ser agregada a cada una de las dimensiones de la tabla de hecho (Fecha, Estrato).

- **Asma por Material de Piso**

La tabla permite desarrollar el requerimiento analítico específico para las variables o características físicas de una vivienda que podrían tener prevalencia en el asma y en las

enfermedades respiratorias. Ahora bien, esta tabla de hecho permite cuantificar la cantidad de personas que presentan asma respecto a un material específico, para posteriormente encontrar la correlación entre uno o varios tipos de materiales, y las enfermedades respiratorias (Asma). Como se mencionó en el anterior apartado, la granularidad de las tablas es muy detallada, pues se toman solo los datos necesarios para dar una respuesta coherente al requerimiento analítico. En este caso, lo que se espera es encontrar la correlación entre el tipo de material y el número de personas que padecen Asma, por ende, la tabla únicamente tiene relación con la dimensión MaterialPiso, fecha y encuestado. Además, presenta únicamente 2 medidas específicas para la solución del requerimiento analítico descritas a continuación:

Medidas/ Hechos:

Personas con Asma: esta medida permite calcular el número de personas que presentan asma, discriminado por Material de piso y fecha, considerando que los datos de las encuestas son de dos años diferentes. Es una medida aditiva, pues bien, puede ser agregada a cada una de las dimensiones de la tabla de hecho (Fecha, Material).

Personas sin Asma: esta medida calcula el número de personas que no presentan asma de acuerdo a un Material y fecha (año) específico. Es una medida aditiva, pues bien, puede ser agregada a cada una de las dimensiones de la tabla de hecho (Fecha, MaterialPiso).

- **Asma por Material de Pared**

La tabla permite desarrollar el requerimiento analítico específico para las variables o características físicas de una vivienda que podrían tener prevalencia en el asma y en las enfermedades respiratorias. Ahora bien, esta tabla de hecho permite cuantificar la cantidad de personas que presentan asma respecto a un material de pared específico, para posteriormente encontrar la correlación entre uno o varios tipos de materiales y las enfermedades respiratorias (Asma). Como se mencionó en el anterior apartado, la granularidad de las tablas es muy detallada, pues se toman solo los datos necesarios para dar una respuesta coherente al requerimiento analítico. En este caso, lo que se espera es encontrar la correlación entre el tipo de material y el número de personas que padecen Asma, por ende, la tabla únicamente tiene relación con la dimensión MaterialPared, fecha y encuestado. Además, presenta únicamente 2 medidas específicas para la solución del requerimiento analítico descritas a continuación:

Medidas/ Hechos:

Personas con Asma: esta medida permite calcular el número de personas que presentan asma, discriminado por Material de pared y fecha, considerando que los datos de las encuestas son de dos años diferentes. Es una medida aditiva, pues bien, puede ser agregada a cada una de las dimensiones de la tabla de hecho (Fecha, MaterialPared).

Personas sin Asma: esta medida calcula el número de personas que no presentan asma de acuerdo a un Material de pared y fecha (año) específico. Es una medida aditiva, pues bien, puede ser agregada a cada una de las dimensiones de la tabla de hecho (Fecha, MaterialPared).

- **Asma Humedad**

La tabla permite desarrollar el requerimiento analítico específico para las variables o características físicas de una vivienda que podrían tener prevalencia en el asma y en las enfermedades respiratorias. Ahora bien, esta tabla de hecho permite cuantificar la cantidad de personas que presentan asma respecto a si presentan o no humedad en sus viviendas, para posteriormente encontrar la correlación entre la presencia de humedad y las enfermedades respiratorias (Asma). Como se mencionó en el anterior apartado, la granularidad de las tablas es muy detallada, pues se toman solo los datos necesarios para dar una respuesta coherente al requerimiento analítico. En este caso, lo que se espera es encontrar la correlación entre la humedad y el número de personas que padecen Asma, por ende, la tabla únicamente tiene relación con la dimensión Humedad, Fecha y Encuestado. Además, presenta únicamente 2 medidas específicas para la solución del requerimiento analítico descritas a continuación:

Medidas/ Hechos:

Personas con Asma: esta medida permite calcular el número de personas que presentan asma, discriminado por presencia de humedad y fecha, considerando que los datos de las encuestas son de dos años diferentes. Es una medida aditiva, pues bien, puede ser agregada a cada una de las dimensiones de la tabla de hecho (humedad, MaterialPared).

Personas sin Asma: esta medida calcula el número de personas que no presentan asma de acuerdo a la presencia de humedad y fecha (año) específico. Es una medida aditiva, pues bien, puede ser agregada a cada una de las dimensiones de la tabla de hecho (Fecha, Humedad).

Porcentaje de personas con Asma: esta medida permite hacer una comparativa entre el número de personas que presentan asma, vs el número de personas que no. Es una medida aditiva, pues bien, puede ser agregada a cada una de las dimensiones de la tabla de hecho (Fecha, Humedad).

Dimensiones:

- **FechaDim**

Esta es una dimensión que permite registrar la fecha y año de la encuesta, y de las tablas de hechos asociadas. Es importante discriminar por año las medidas registradas en la tabla, especialmente porque los datos tienen encuestas de 2017 y 2021. Ahora bien, se presentan a continuación los atributos de la dimensión:

FechaUUID: representa la PK correspondiente de la dimensión.

fecha: representa la fecha de la encuesta o los datos asociados. Se decidió que este atributo sea un string, con el siguiente formato: “diaMESaño” (02SEP2017). Ahora bien, debido a que este dato es la fecha en la que se realizó una encuesta, es un valor que necesario para nuevos análisis dado que representa un registro único, por ende, el tipo de manejo de historia de variación lenta es de tipo 2 y así evitar perder valores de fechas anteriores.

tieneAsma: representa un “booleano” que indica si la persona encuestada tiene o no asma, toma el valor 1 como si, y el valor 2 como no. Ahora bien, se decidió mantener el tipo de manejo de historia de variación lenta como tipo 0, debido a que es un valor que posiblemente no deba ser cambiado, o no se tiene cierta certeza de su cambio ya que el asma es una condición crónica.

- **LocalidadDim**

Esta es una dimensión que permite registrar la localidad o ubicación de un encuestado, dado que se requiere discriminar por localidad los valores y las medidas en el ETL. Ahora bien, se presentan a continuación los atributos de la dimensión:

localidadUUID: representa la PK correspondiente de la dimensión.

localidad: el atributo registra la localidad del encuestado, esto mediante un valor de tipo string. Dado que es un valor que puede cambiar, y considerando que posiblemente sea en una fecha específica necesaria para nuevos análisis, se opta por el tipo de manejo de historia de variación lenta de tipo 2.

tieneAsma: representa un “booleano” que indica si la persona encuestada tiene o no asma, toma el valor 1 como si, y el valor 2 como no. Ahora bien, se decidió mantener el tipo de manejo de historia de variación lenta como tipo 0, debido a que es un valor que posiblemente no deba ser cambiado, o no se tiene cierta certeza de su cambio ya que el asma es una condición crónica.

- **EstratoDim**

Esta es una dimensión que permite registrar el estrato del encuestado, dado que se requiere discriminar por estrato los valores y las medidas en el ETL, con el objetivo de cumplir con el requerimiento de las condiciones socioeconómicas. Ahora bien, se presentan a continuación los atributos de la dimensión:

estratoUUID: representa la PK correspondiente de la dimensión.

estrato: el atributo registra el estrato del encuestado, esto mediante un valor de tipo int. Este atributo puede tomar los valores: 1, 2, 3, 4, 5, 6. Dado que es un valor que puede cambiar, y considerando que posiblemente sea en una fecha específica necesaria para nuevos análisis, se opta por el tipo de manejo de historia de variación lenta de tipo 2 y así mantener nuevos registros con fechas y valores actualizados del encuestado, para no modificar valores anteriores importantes para los actuales análisis.

tieneAsma: representa un “booleano” que indica si la persona encuestada tiene o no asma, toma el valor 1 como si, y el valor 2 como no. Ahora bien, se decidió mantener el tipo de manejo de historia de variación lenta como tipo 0, debido a que es un valor que posiblemente no deba ser cambiado, o no se tiene cierta certeza de su cambio ya que el asma es una condición crónica.

- **MaterialPisoDim**

Esta es una dimensión que permite registrar el tipo de material del piso de la vivienda del encuestado, dado que se requiere discriminar por material los valores y las medidas en el ETL. Ahora bien, se presentan a continuación los atributos de la dimensión:

materialUUID: representa la PK correspondiente de la dimensión.

material: el atributo registra el tipo de material de piso del encuestado, esto mediante un valor de tipo int. Este atributo puede tomar los valores: 1, 2, 3, 4, 5, 6, 7, 8, 9. Dado que es un valor que puede cambiar, y considerando que posiblemente sea en una fecha específica necesaria para nuevos análisis, se opta por el tipo de manejo de historia de variación lenta de tipo 2 y así mantener nuevos registros con fechas y valores

actualizados del encuestado, y no modificar valores anteriores importantes para los actuales análisis.

tieneAsma: representa un “booleano” que indica si la persona encuestada tiene o no asma, toma el valor 1 como si, y el valor 2 como no. Ahora bien, se decidió mantener el tipo de manejo de historia de variación lenta como tipo 0, debido a que es un valor que posiblemente no deba ser cambiado, o no se tiene cierta certeza de su cambio ya que el asma es una condición crónica.

- **MaterialParedDim**

Esta es una dimensión que permite registrar el tipo de material de pared de la vivienda del encuestado, dado que se requiere discriminar por material los valores y las medidas en el ETL. Ahora bien, se presentan a continuación los atributos de la dimensión:

materialUUID: representa la PK correspondiente de la dimensión.

material: el atributo registra el tipo de material de pared del encuestado, esto mediante un valor de tipo int. Este atributo puede tomar los valores: 1, 2, 3, 4, 5, 6, 7, 8, 9. Dado que es un valor que puede cambiar, y considerando que posiblemente sea en una fecha específica necesaria para nuevos análisis, se opta por el tipo de manejo de historia de variación lenta de tipo 2 y así mantener nuevos registros con fechas y valores actualizados del encuestado, y no modificar valores anteriores importantes para los actuales análisis.

tieneAsma: representa un “booleano” que indica si la persona encuestada tiene o no asma, toma el valor 1 como si, y el valor 2 como no. Ahora bien, se decidió mantener el tipo de manejo de historia de variación lenta como tipo 0, debido a que es un valor que posiblemente no deba ser cambiado, o no se tiene cierta certeza de su cambio ya que el asma es una condición crónica.

- **HumedadDim**

Esta es una dimensión que permite registrar si la vivienda de un encuestado presenta o no humedad, dado que se requiere discriminar por la presencia de humedad los valores y las medidas en el ETL. Ahora bien, se presentan a continuación los atributos de la dimensión:

humedadUUID: representa la PK correspondiente de la dimensión.

material: el atributo registra la presencia de humedad en la vivienda del encuestado, esto mediante un valor de tipo int. Este atributo puede tomar los valores: 1: presenta humedad, 2: no presenta humedad. Dado que es un valor que puede cambiar, y considerando que posiblemente sea en una fecha específica necesaria para nuevos análisis, se opta por el tipo de manejo de historia de variación lenta de tipo 2 y así mantener nuevos registros con fechas y valores actualizados del encuestado, y no modificar valores anteriores importantes para los actuales análisis.

tieneAsma: representa un “booleano” que indica si la persona encuestada tiene o no asma, toma el valor 1 como si, y el valor 2 como no. Ahora bien, se decidió mantener el tipo de manejo de historia de variación lenta como tipo 0, debido a que es un valor que posiblemente no deba ser cambiado, o no se tiene cierta certeza de su cambio ya que el asma es una condición crónica.

3. Entendimiento de los Datos, Creación del Data Mart y Proceso ETL

3.1 Entendimiento de las Fuentes de Datos

En esta etapa se analizan las características principales del conjunto de datos, así como también las dimensiones de calidad de estos (completitud, unicidad, consistencia, validez). Posteriormente, se limpian los datos y se transforman para la futura etapa de modelamiento. Como datos de entrada se leen los CSVs correspondientes a las respuestas de las encuestas de los años 2017 y 2021.

Como existen algunos casos puntuales en donde la nomenclatura para los códigos de las preguntas difiere entre versión, se corregirá manualmente utilizando la librería de Pandas de ser necesario. Adicionalmente, únicamente se dejarán aquellas variables que fueron designados como relevantes durante el establecimiento de los requerimientos analíticos. A continuación, se muestra el análisis de las dimensiones de calidad para cada una de las fuentes de datos seleccionadas:

3.1.1 Completitud

- **Encuestas 2017:** Se encontraron aproximadamente 30.7% de los valores nulos para las localidades. Del mismo modo, se registran algunos porcentajes de valores nulos en algunas preguntas significativas. Por ejemplo, en NPCFP36, es aproximadamente del 12.1%. Debido a que la proporción de nulas no es tan significativa y a que aquellas variables que más presentan nulos no son tan relevantes para el presente estudio, se optó por eliminar estos valores.
- **Encuestas 2021:** Algunos valores referentes a la ubicación del encuestado como NOMBRE_LOCALIDAD tienen porcentajes de valores nulos de aproximadamente 19.6%. Del mismo modo, y de forma similar la pregunta de NPCFP36 tienen aproximadamente un total de 10.9% de valores nulos. Al igual que con la encuesta de 2017 se decidió a borrar estos registros nulos.

3.1.2 Unicidad

Los porcentajes de publicados para los datos de las encuestas de 2017 y 2021 fueron de 88.94% y 86.11%, respectivamente. Aunque existe un porcentaje relativamente importante de filas repetidas en los Data Frames, es importante notar que esto puede ser un indicio que la población a la cual se le hizo la encuesta dio respuestas muy similares entre sí. Sin embargo, también podría indicar errores en la recopilación de datos, pues se podrían haber incluido datos repetidos accidentalmente o haber existido problemas en el proceso de carga de datos.

A pesar de esto, la encuesta no tiene un margen de error tan alto como para justificar que estas duplicaciones se deben al azar o al error estadístico o de contexto. De acuerdo a la documentación de la encuesta en la sección de **Procedimiento de muestreo**:

"Para el cálculo de los tamaños de muestra se establecieron los siguientes parámetros para cada UPZ en Bogotá y para cada municipio de Cundinamarca: precisión esperada medida en términos del error estándar relativo igual a 7%, con un nivel de confiabilidad del 95%, para la prevalencia de alrededor del 10% y un efecto de diseño de 1,2."

Así pues, más bien esta situación se puede deber a que las respuestas dadas fueron idénticas entre la población para el subconjunto de preguntas elegidas y que esta población puede compartir situaciones similares que condicionaron sus respuestas a al conjunto de preguntas de interés seleccionadas. Por tanto, no es prudente, ni necesario, eliminar los registros duplicados,

sino utilizarlos para hacer un análisis más exhaustivo más adelante para definir las causas de esta situación.

3.1.3 Consistencia

La consistencia de los datos es definida en términos de la integridad de los datos entre diferentes filas o columnas de una fuente o varias fuentes. En el caso de la encuesta multipropósito para el subconjunto de preguntas seleccionadas se espera que todas las respuestas sean de tipo numérico (ya sea int o float) pues las respuestas a estas preguntas son de selección múltiple en la encuesta y las respuestas se representan con valores numéricos.

```
# Verificar si todas las filas contienen valores numéricos e ignorar columnas que si pueden tener valores no nume
todas_filas_numericas_2017 = df_preguntas_interes_2017.drop(columns=['LOCALIDAD_TEX']).applymap(lambda x: isinstance(x, (int, float)))
todas_filas_numericas_2021 = df_preguntas_interes_2021.drop(columns=['NOMBRE_LOCALIDAD', 'NOMBRE_ESTRATO']).applymap(lambda x: isinstance(x, (int, float)))

# Verificar si hay algún False en la Serie resultante
hay_falso_2017 = any(~todas_filas_numericas_2017)
hay_falso_2021 = any(~todas_filas_numericas_2021)

# Imprimir el resultado
print("¿Hay algún valor no numerico en las filas para el 2017?", hay_falso_2017)
print("¿Hay algún valor no numerico en las filas para el 2021?", hay_falso_2021)

¿Hay algún valor no numerico en las filas para el 2017? False
¿Hay algún valor no numerico en las filas para el 2021? False
```

Figura 2. Revisión de la presencia de valores de tipo numérico como respuestas a las encuestas

Como se puede observar en la figura 2 todas las filas que se suponen tener valores numéricos, los tienen. Por tanto, las respuestas a las preguntas de selección múltiple son consistentes con las opciones de respuesta esperadas.

3.1.4 Validez

Como parte del proceso de validación, se comprobó que todos los valores fueran coherentes con las encuestas, para ello se verificó en primer lugar que todos los valores fueran mayores a 1 (Figura 3). Posteriormente, según el rango de posibles valores indicados por la página de la DIAN para cada pregunta, se comprobó que todos los valores de cada columna estuvieran dentro del rango dado. Tras haber realizado este procedimiento, se llegó a la conclusión que todos los datos cumplen con la dimensión de validez a cabalidad.

```
# Verificar si hay valores no positivos en cada columna (excluyendo 'LOCALIDAD_TEX')
valores_no_positivos = df_preguntas_interes_2017.drop(columns=['LOCALIDAD_TEX']).applymap(lambda x: x < 0)

# Imprimir los resultados por columna
for columna in valores_no_positivos.columns:
    valores_no_positivos_en_columna = valores_no_positivos[columna]
    if len(valores_no_positivos_en_columna[valores_no_positivos_en_columna.index.tolist()]) > 0:
        print(f"Valores no positivos en la columna '{columna}':")

Como se puede ver todos los valores son positivos y, por tanto, son válidos y se ajustan a las respuestas esperadas de la encuesta.

# Supongamos que df_preguntas_interes_2017 es tu DataFrame

# Verificar si hay valores no positivos en cada columna (excluyendo 'LOCALIDAD_TEX')
valores_no_positivos = df_preguntas_interes_2021.drop(columns=['NOMBRE_LOCALIDAD', 'NOMBRE_ESTRATO']).applymap(lambda x: x < 0)

# Imprimir los resultados por columna
for columna in valores_no_positivos.columns:
    valores_no_positivos_en_columna = valores_no_positivos[columna]
    if len(valores_no_positivos_en_columna[valores_no_positivos_en_columna.index.tolist()]) > 0:
        print(f"Valores no positivos en la columna '{columna}':")
```

Figura 3. Revisión del cumplimiento del rango de las variables

3.1.5 Entendimiento de los datos

En la figura 4 se muestran el total de personas encuestadas en los años 2017 y 2021, así como también el número de personas que afirmaron sufrir de asma al momento en que se realizó cada encuesta. El número total de personas encuestadas fue de 406295, de las cuales solo 9327 afirmaron sufrir de algún tipo de asma, es decir que solo aproximadamente el 0.023% de las personas encuestadas padecían asma.

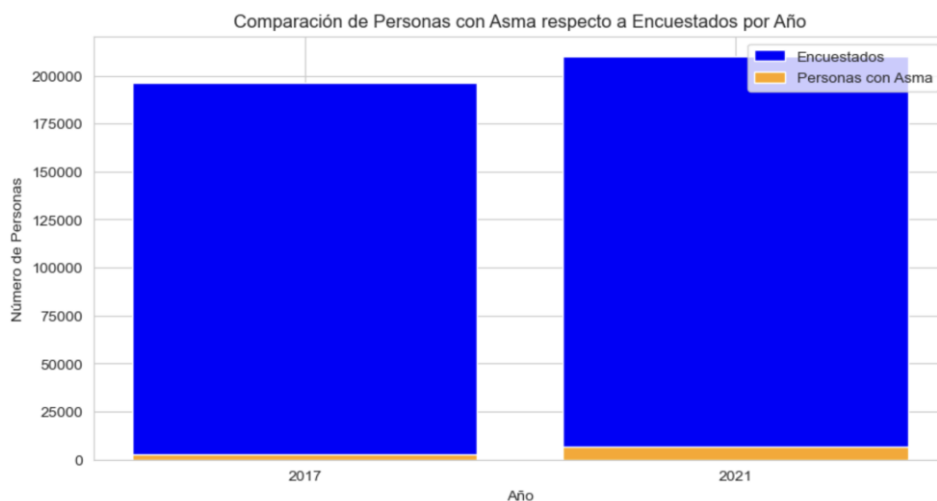


Figura 4. Comparacion entre personas encuestadas y personas con asma por año

Con respecto a los encuestados por localidad, en la figura 5 se muestra de manera grafica el número de personas encuestadas por localidad para el 2017 y el 2021. Las magnitudes para cada localidad son similares entre los diferentes años, lo cual es un buen indicativo de una adecuada distribución de la muestra para representar a la población de Bogotá. También, la frecuencia de los encuestados resulta coherente con las poblaciones estimadas de cada localidad. Por ejemplo, Kennedy, la cual es la localidad con mayor índice población, es aquella con el mayor número de encuestados en ambos casos.

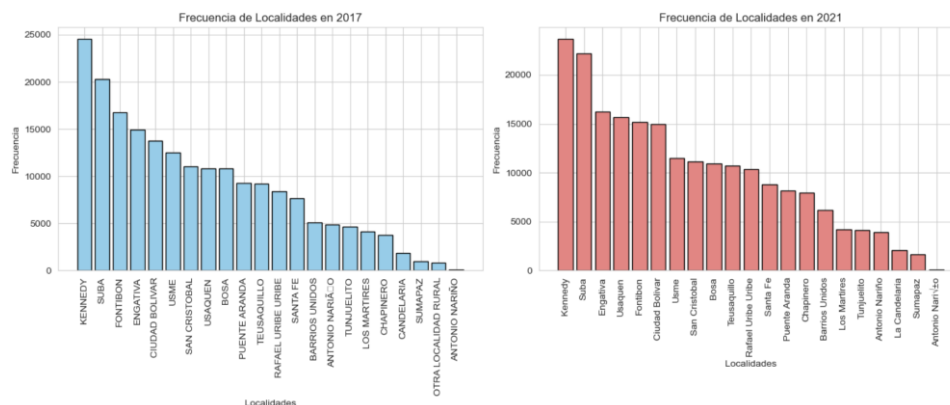


Figura 5. Numero de encuestados según la localidad y el año

En la figura 6 se presenta el número de personas encuestadas según el estrato socioeconómico y el año en que se realizó la encuesta. Al igual que en el caso anterior, la distribución de los

encuestados sigue una tendencia similar entre año y año, y también es congruente con los datos oficiales expedidos por la dirección de censos y demografía (DANE).

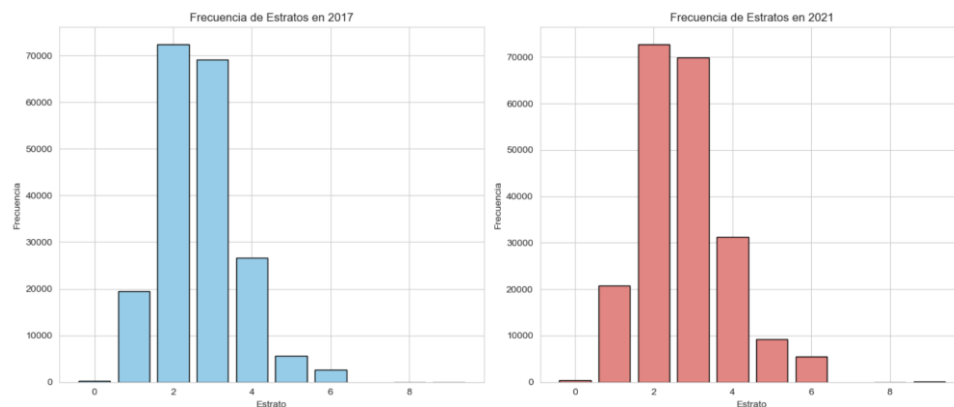


Figura 6. Numero de encuestados según el estrato y el año

Finalmente, para analizar la correlación existente entre el asma y las diferentes variables evaluadas, debido a que las variables son categóricas se optó por utilizar una distribución chi-cuadrado de Pearson, tal como lo establece la literatura. La prueba de chi-cuadrado (χ^2) es una prueba estadística que se utiliza para determinar si hay una asociación significativa entre dos variables categóricas. La prueba evalúa si las observaciones empíricas, que están organizadas en una tabla de contingencia, difieren de las expectativas teóricas bajo la hipótesis nula de independencia entre las variables. A continuación, se muestran los pasos que se siguieron para el cálculo de esta distribución.

1. **Hipótesis Nula (H_0):** La hipótesis nula asume que no hay asociación entre las dos variables categóricas; son independientes.
2. **Valor p:** Se determina la probabilidad de obtener el estadístico de prueba observado (o uno más extremo) bajo la hipótesis nula.
3. **Conclusión:** Si el valor p es menor que el nivel de significancia elegido (de 0.05), se rechaza la hipótesis nula, lo que sugiere que hay una asociación significativa entre las variables

En las figuras 7 y 8 se muestran las significancias obtenidas para cada una de las variables estudiadas en las encuestas del año 2017 y 2021, respectivamente. Como se puede inferir a partir de la matriz de la figura 7, para la encuesta llevada a cabo en 2017 existe una clara correlación entre todas las variables con respecto al asma exceptuando en las siguientes:

- NPCFP14I y DPTOMPIO: Divipola departamento-municipio
- NPCFP14I y NVCBP12: ¿Cuál es el material predominante de las paredes exteriores?
- NPCFP14I y NVCBP4: ¿La edificación está ubicada en un conjunto cerrado

En cuanto a lo referente a las encuestas realizadas en el año 2021, con base a lo obtenido en la matriz de significancias de la Figura 8 existe una clara correlación entre todas las variables con respecto al asma exceptuando en las siguientes:

- NPCFP14F y DPTO: Departamento
- NPCFP14F y MPIO: Municipio
- NPCFP14F y NVCBP12: ¿Cuál es el material predominante de las paredes exteriores?
- NPCFP14F y NVCBP4: ¿La edificación está ubicada en un conjunto residencial?
- NPCFP14F y NVCBP10: Tipo de vivienda

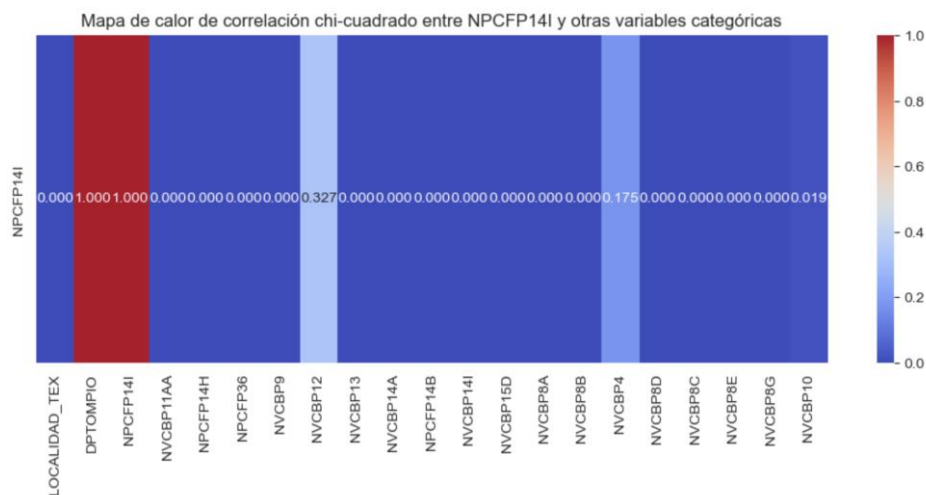


Figura 7. Matriz de las significancias entre el asma y las variables seleccionadas (Año 2017)

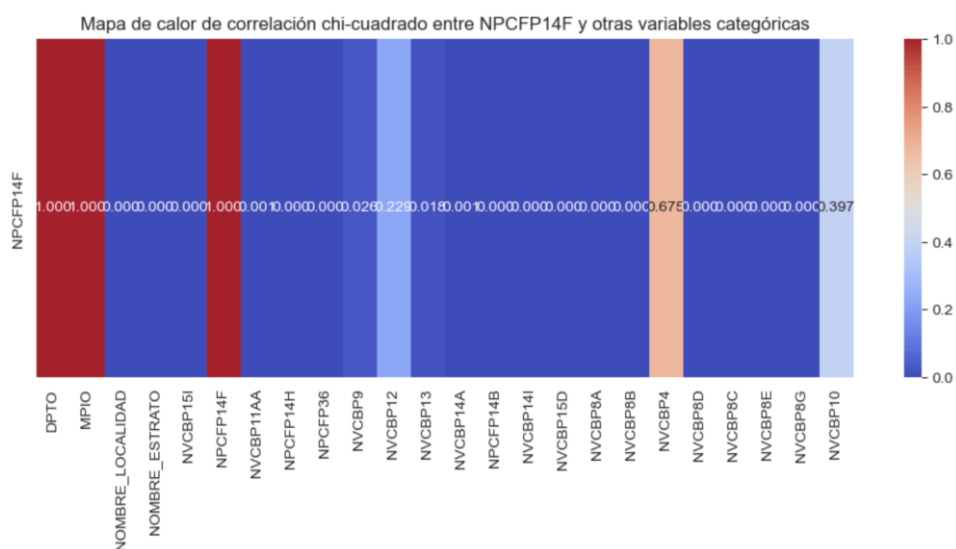


Figura 8. Matriz de las significancias entre el asma y las variables seleccionadas (Año 2021)

3.2 Diseño e Implementación el Proceso de ETL

4. Arquitectura de Solución y Tableros de Control

4.1 Arquitectura de Solución Propuesta

La arquitectura de solución para el proyecto y los requerimientos de negocio puede ser descrita mediante la siguiente imagen:

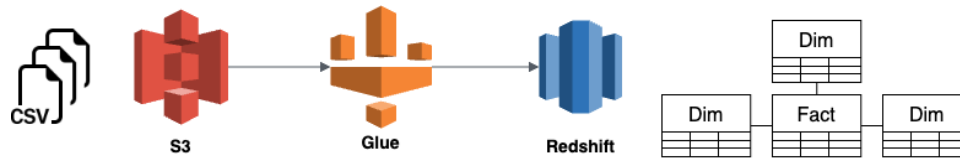


Figura 9. Arquitectura de solución

Fueron proporcionados un conjunto de archivos en formato CSV los cuales contienen la información de las encuestas multipropósito realizadas para el año 2017 y 2021. Estos datos fueron cargados mediante los servicios de Amazon AWS en un data lake (S3). Con este data lake el servicio de integración de datos (Glue) permite la lectura y acceso a estos archivos posteriormente transformarlos y cargarlos, mediante un esquema multidimensional (Figura 1) a la bodega de datos (Redshift). Adicionalmente, se construyó un Dashboard o tablero de control que conectado a la bodega de datos (Redshift) que permite visualizar a dar solución a los requerimientos analíticos seleccionados por los estudiantes de medicina.

4.2 Implementación de Tableros de Control

Para la implementación de los tableros de control (se realizó uno unificado) se tuvieron en cuenta completamente las indicaciones y necesidades de los usuarios finales del proyecto, es decir los estudiantes de medicina. En este caso, se generaron visualizaciones relacionadas a los dos requerimientos analíticos escogidos (factores socioeconómicos y variables o características de la vivienda) siguiendo las pautas acordadas con los usuarios finales. Se generaron graficas que permiten establecer una correlación entre las variables seleccionadas y la prevalencia de las enfermedades respiratorias, especialmente el asma. Además, se intentó construir todas las gráficas que eran de utilidad para los estudiantes de medicina, y especialmente para el análisis que ellos estaban realizando para el proyecto conjunto. Una de las vistas del tablero de control se puede ver a continuación:

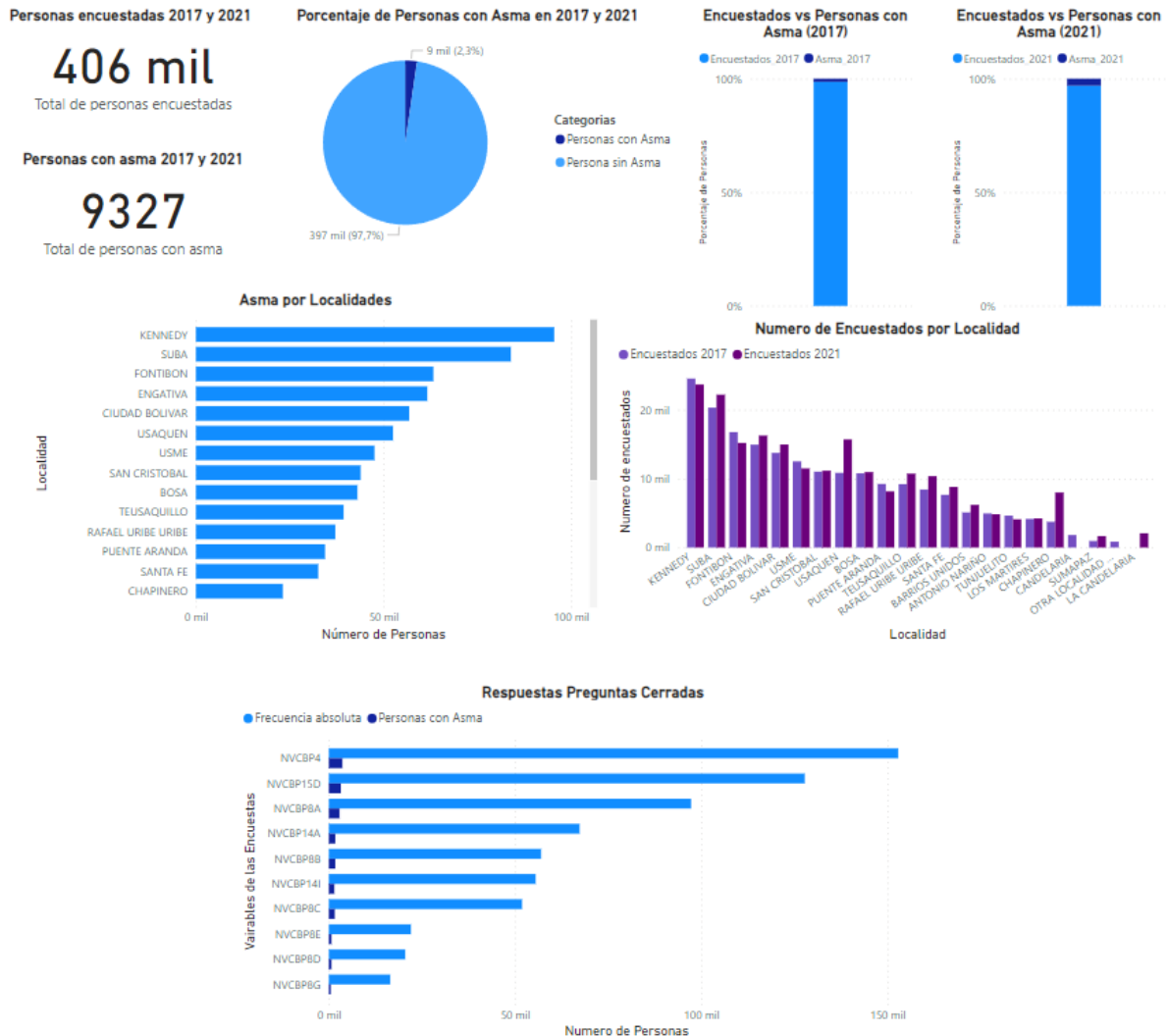


Figura 9. Arquitectura de solución

5. Evaluación de Trabajo en Equipo

5.1 Autoevaluación en las competencias

Como equipo, consideramos que las actividades en clase y el material curricular provisto por el curso representaron herramientas de gran valor para el desarrollo de habilidades como la resolución de problemas. También, el enfoque multidisciplinar del programa de ingeniería de Sistemas y computación de la Universidad de los Andes fue de gran utilidad para conformar un equipo diverso y una relación colaborativa con el grupo de medicina. Consideramos que, a pesar de las dificultades que se atravesaron, existió un interés real y un compromiso latente por todos los miembros del equipo a lo largo del proyecto.

Si bien gran parte de las tareas realizadas se llevaron a cabo de manera conjunta, aquellas que se delegaron en tiempos ajenos a los de clase y las reuniones de equipo se desarrolló con éxito. Esto demuestra que todos los miembros tienen un sentido de autonomía por el equipo.

5.2 Autoevaluación de la calidad y aporte al proyecto entregado

Consideramos que el proyecto desarrollado cumple con todos los estándares de calidad y requisitos transmitidos por el grupo docente. Tras realizar un ejercicio de retroalimentación y retrospectivo con los estudiantes de medicina, ellos nos transmitieron su satisfacción con los resultados obtenidos. En cuanto al aporte de cada miembro, todos colaboraron de manera activa en todas las etapas del proyecto. Para cuantificar de manera más precisa el desempeño y aporte de cada integrante, se replicó una estrategia similar a las anteriores entregas del curso, la cual consiste en indicar las actividades realizadas y los tiempos invertidos en cada una de ellas por parte de los miembros del equipo. Dicha tabla se encuentra en la sección 5.3 del presente documento.

5.3 Evaluación entre miembros del equipo

Miembro	Tareas	Tiempos Aproximados
Juan Sebastián Sánchez	Requerimientos analíticos	120 min
	Configuración del ETL	60 min
	Dashboard – Primera parte	120 min
	Entrevista con grupo de medicina	30 min
	Documentación del trabajo en equipo	40 min
Miguel Cárdenas Cárdenas	Entrevista con grupo de medicina	30 min
	Dashboard – Segunda parte	120 min
	Diseño del modelo multidimensional	70 min
	Documentación modelo multidimensional, construcción de la arquitectura de solución y Tablero de control.	120 min
Santiago Paeres Gonzales	Entendimiento y exploración de datos	150 min
	Diseño e implementación del ETL	120 min
	Entrevista con grupo de medicina	30 min

Tabla x. Distribución de trabajo

Como equipo, consideramos que lo más justo para asignar la calificación respectiva a cada estudiante, consiste en hacer una ponderación con base a la importancia de las tareas realizadas por cada miembro, así como también los tiempos invertidos en cada una de ellas. Este es el criterio más objetivo que se ideó para calificar el desempeño de cada estudiante. Con respecto a la calificación entre miembros, en todos los casos existe una opinión muy favorable de cada miembro respecto al trabajo realizado por los otros, por lo que se optó por no colocar las calificaciones particulares de cada uno con el fin de evitar ser redundante.

6. Bibliografías

[1] EAGAR, Gareth. Data Engineering with AWS. Packt Publishing. 2021.

- [2] Tiotiu, A. I., Novakova, P., Nedeva, D., Chong-Neto, H. J., Novakova, S., Steiropoulos, P., & Kowal, K. (2020). Impact of Air Pollution on Asthma Outcomes. *International journal of environmental research and public health*, 17(17), 6212. <https://doi.org/10.3390/ijerph17176212>
- [3] Guibas, G. V., Mathioudakis, A. G., Tsoumani, M., & Tsabouri, S. (2017). Relationship of Allergy with Asthma: There Are More Than the Allergy "Eggs" in the Asthma "Basket". *Frontiers in pediatrics*, 5, 92. <https://doi.org/10.3389/fped.2017.00092>
- [4] Vallette, Jim & Schettler, Ted & Wolfe, Michael. (2014). Asthmagens in Building Materials: The Problem and Solutions. 10.13140/2.1.1178.4961.
- [5] Currie, G., Ayres, J. Occupational asthmagens. *Prim Care Respir J* **14**, 72–77 (2005). <https://doi.org/10.1016/j.pcrj.2004.11.001>