PROFESORES: Felipe González Casabianca
INTEGRANTES:

Maria Camila Parra Díaz (201819464)
Esteban Emmanuel Ortiz Morales (2018194646)
Sergio Julián Zona Moreno (201914936)

Tabla de contenido

1 Introducción

En el presente documento desarrollaremos el segundo proyecto del curso de Inteligencia de negocios. El propósito central es aplicar la metodología modelado dimensional para dar solución al caso de estudio presentado por “Infraestructura Visible”. Un caso enfocado en el tema de analítica y almacenamiento de defunciones en Colombia. Procederemos a desarrollar todo el proceso ETL requerido, al igual que detallar los motivos de nuestra manera de modelarlo. De manera complementaria, nos enfocaremos en las prioridades de negocio para que nuestra propuesta otorgue valor al mismo.

1.1 Aclaraciones preliminares

- Se recomienda al calificador/lector de este documento visualizarlo con un Zoom de 150% para evitar forzar la vista. Además, este hecho permite ver con nitidez los gráficos y las tablas presentadas.
- Adjunto con este archivo se encuentra el Notebook de operaciones realizado en Jupyter, el tablero de control de estadísticas de los datos, la presentación de resultados en formato .pdf para las organizaciones beneficiadas y el video con los resultados del proyecto. Todo esto puede ser visualizado en la sección de anexos.

2 Justificación de necesidades analíticas

2.1 Justificación de temas analíticos

Histórico de muertes general:

Dentro del contexto de negocio, y dadas las entrevistas presentadas por el cliente, era de su particular interés conocer valores históricos a nivel general para generar aproximaciones primarias al contexto nacional. Posteriormente, se procede a aumentar el nivel de granularidad para conocer detalles más particulares del conjunto de muertes.

Tendencia de muertes general:

Dentro del contexto de negocio, y dadas las entrevistas presentadas por el cliente, era de su particular interés conocer tendencias de muertes a nivel general que permitiesen la obtención de valores para análisis macroeconómicos. Estos valores usualmente se encuentran altamente relacionados con las condiciones demográficas en el territorio nacional.

Histórico de muertes geográfico:

Dentro del contexto de negocio, y dadas las entrevistas presentadas por el cliente, era de su particular interés conocer valores históricos según ubicaciones geográficas. Esto para otorgar mejores análisis a entidades territoriales. También, para conocer las causas de muerte “anormales” en algunos departamentos alejados e intentar explicar los motivos que las suscitan.

Tendencia de muertes geográfico:

Dentro del contexto de negocio, y dadas las entrevistas presentadas por el cliente, era de su particular interés conocer las tendencias poblaciones a nivel geográfico. De igual forma que a nivel general, esto permite obtener valores macroeconómicos que facilita recomendaciones de políticas públicas a gobiernos distritales, municipales y hasta departamentales.

2.2 Descripción y disponibilidad de las fuentes de datos

Manejaremos tres fuentes de datos abiertas (disponibles en todo momento al público general), estas son: fuente de datos oficial de defunciones del DANE (2010-2020), fuente de datos externa poblacional de Terridata (enfocada a municipios y departamentos) y fuente de datos externa de distribución poblacional de DatosMacro para agrupar grupos de edad.

2.3 Factibilidad e impacto

2.3.1 Factibilidad y una aproximación para ser medida

Para medir la factibilidad, optamos por determinarla como “la probabilidad para completar correctamente una tarea/proceso”. Con base en esto, procesos como la recolección de datos (que son abiertos al público) tienen una factibilidad del 100%; mientras que procesos como la predicción de valores en una serie temporal específica tienen una factibilidad del 60%.

Sabemos que la medida puede llegar a ser subjetiva según cada interpretación, pero lo importante de esta métrica es “aterrizar al mundo real” el proceso de negocio que se desea implementar, para ver qué tan viable y eficiente es su desarrollo.

2.3.2 Impacto y una aproximación para ser medido

Para medir el impacto, optamos por determinarlo como “la importancia de un proceso particular en el desarrollo de las metas del negocio y de otros procesos”. Con base en esto, procesos como la recolección de datos (que son abiertos al público) tienen un impacto del 100%, puesto que sin ellos ningún otro proceso puede ser desarrollado; mientras que procesos como la depuración de los datos con un 75% tienen menor impacto (ya que los datos en su mayoría ya vienen preprocesados), aunque esto no quiere decir que no sean importantes.

Sabemos que la medida puede llegar a ser subjetiva según cada interpretación, pero lo importante de esta métrica es determinar qué procesos presentan mayor relevancia e impedir fallos en cascada que afecten múltiples procesos de valor en la organización.

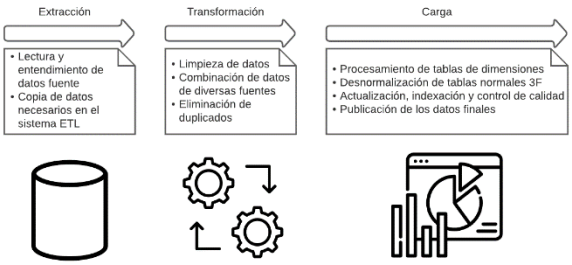
2.4 Selección de primer proceso de negocio a implementar

Dado el análisis de factibilidad y relevancia de los procesos de negocio (que puede ser evidenciado en el Excel adjunto), recomendamos que el primer proceso de negocio a implementar sea el de “Adquisición del conjunto de datos”. Esto se debe a que, sin estos valores, no se puede diseñar un modelo adecuado, tampoco se pueden presentar estadísticas descriptivas de los mismo, y mucho menos generar conclusiones al respecto. Como este proceso es la base para desarrollar todo lo demás, solicitamos a negocio su implementación de la manera más eficiente posible.

3 Modelado

3.1 Diagrama de alto nivel del proceso ETL

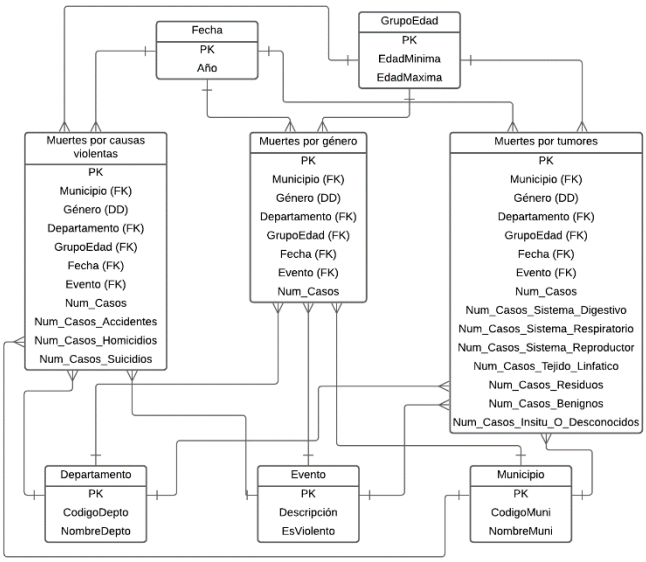
Dentro de este diagrama de alto nivel utilizamos los tres pasos de un proceso ETL. Particularmente, el modelo solamente implementa un manejo de historia Tipo I para el manejo de los atributos de las dimensiones. Durante nuestro desarrollo seguimos el flujo de trabajo del modelo planteado. El diagrama general no tuvo modificaciones entre la entrega 1 y la entrega 2.



3.2 Modelado dimensional

El modelo realizado cuenta con seis dimensiones y tres tablas de hechos, que serán descritas a continuación. Cada dimensión cuenta con al menos dos atributos además de su llave primaria, a excepción de la dimensión degenerada 'Género'. Dentro de las tablas de hechos encontramos entre una y siete medidas aditivas, las cuales podrán proporcionarle a negocio información valiosa relacionada con las muertes en Colombia, teniendo en cuenta las dimensiones del modelo.

La primordial diferencia con respecto a la primera entrega se encuentra en que la tabla de hechos "Muertes por tumores", cuenta con una nueva métrica "Num_Casos_Insitu_O_Desconocidos". Producto de agregar el nuevo conjunto de datos del Dane 2018-2020.



3.3 Análisis en tablas de hechos

Como interés particular para negocio, optamos por modelar tres hechos fundamentales relacionados con el número de condiciones: muertes por género, muertes por desnutrición y muertes por causas violentas.

El primer hecho busca enfocarse principalmente en conocer las diferencias entre los tipos de muertes que abarcan los géneros en Colombia. Determinar cuáles son las más comunes e intentar explicar la causalidad detrás de las mismas.

El segundo hecho, busca enfocarse en conocer cuáles zonas del país se ven arduamente afectadas por problemas de desnutrición, generalmente comprendidas dentro de un marco de pobreza extrema o condiciones de difícil acceso.

Por último, el tercer hecho, busca analizar las muertes violentas dentro de un contexto polifacético, se comprende principalmente de estadísticas con muertes por homicidios, suicidios y accidentes de diversa índole. Buscamos encontrar una distribución porcentual

demográfica que indique cuáles zonas del país se ven más embebidas en dichos tipos de eventos (como es lógico, normalizando los datos para efectuar un análisis adecuado entre ciudades y municipios).

3.3.1 Granularidad y justificación

Muertes por género:
Hecho con baja granularidad, si bien realiza una pequeña subdivisión según el género del individuo fallecido, toma en cuenta todos los tipos de posibles eventos de muertes, lo que abarca una gran cantidad de datos. Aunque parezca un análisis poco útil, realmente se busca encontrar cuáles son los tipos de muertes que diferencia a los grupos y determinar si existe una influencia fenotípica o genotípica en los mismos; o, por el contrario, si existe una influencia determinada por condiciones geográficas y/o demográficas.

Muertes por tumores:
Hecho con alta granularidad, especifica particularmente eventos muy puntuales, los cuáles son las muertes por tumores. Aquí se divide en 5 grupos: muertes por tumores en sistema digestivo, en sistema respiratorio, en sistema reproductor, en tejido linfático, residuos de tumores y tumores benignos. Como nos enfocamos particularmente en este nicho, se pueden obtener análisis muy interesantes por determinados grupos de edades y ubicaciones geográficas.

Muertes por causas violentas:
Hecho con medio granularidad, toma diversos tipos de causas de muerte y no se centra con claridad en uno solo. Sin embargo, este análisis facilitar la búsqueda de epicentros puntuales donde pueden ocurrir con más frecuencia estos tipos de muertes, con el fin de prevenirlas a futuro.

3.3.2 Medidas y hechos

Muertes por género (Num_Casos):
Esta variable mide la cantidad de muertes que ocurren dependiendo del género (hombre – mujer). Esta medida es de tipo aditiva, ya que se puede sumar a través de todas las dimensiones del modelo.
Muertes por tumores (Num_Casos):
Esta variable mide la cantidad de muertes que ocurren debido a la desnutrición. Esta medida es de tipo aditiva, ya que se puede sumar a través de todas las dimensiones del modelo.
Muertes por tumores - sistema digestivo (Num_Casos_Sistema_Digestivo):
Esta variable mide la cantidad de muertes de personas por tumores en el sistema digestivo. Esta medida es de tipo aditiva, ya que se puede sumar a través de todas las dimensiones del modelo.
Muertes por tumores - sistema resp. (Num_Casos_Sistema_Respiratorio):
Esta variable mide la cantidad de muertes de personas por tumores en el sistema respiratorio. Esta medida es de tipo aditiva, ya que se puede sumar a través de todas las dimensiones del modelo.
Muertes por tumores - sistema repr. (Num_Casos_Sistema_Reproductor):
Esta variable mide la cantidad de muertes de personas por tumores en el sistema reproductor. Esta medida es de tipo aditiva, ya que se puede sumar a través de todas las dimensiones del modelo.
Muertes por tumores - tejido linfático (Num_Casos_Tejido_Linfático):
Esta variable mide la cantidad de muertes de personas por tumores en el tejido linfático. Esta medida es de tipo aditiva, ya que se puede sumar a través de todas las dimensiones del modelo.

Muertes por tumores - residuos (Num_Casos_Residuos):
Esta variable mide la cantidad de muertes de personas por residuos de tumores. Esta medida es de tipo aditiva, ya que se puede sumar a través de todas las dimensiones del modelo.
Muertes por tumores - benignos (Num_Casos_Benignos):
Esta variable mide la cantidad de muertes de personas por tumores benignos. Esta medida es de tipo aditiva, ya que se puede sumar a través de todas las dimensiones del modelo.
Muertes por tumores – in situ o desconocidos (Num_Casos_Insitu_O_Desconocidos):
Esta variable mide la cantidad de muertes de personas por tumores in situ o desconocidos. Esta medida es de tipo aditiva, ya que se puede sumar a través de todas las dimensiones del modelo.
Muertes por causas violentas (Num_Casos):
Esta variable mide la cantidad de muertes que ocurren por causas violentas. Por causas violentas se entienden: accidentes en transporte terrestre, caídas, accidentes causados por máquinas, exposición al humo, fuego y llamas, etc. Esta medida es de tipo aditiva, ya que se puede sumar a través de todas las dimensiones del modelo.
Muertes por causas violentas - accidentes (Num_Casos_Accidentes):
Esta variable mide la cantidad de muertes que ocurren por causas violentas relacionadas con accidentes. Por ejemplo: accidentes en transporte terrestre, caídas, accidentes causados por máquinas, ahogamiento y sumersión accidentales, etc. Esta medida es de tipo aditiva, ya que se puede sumar a través de todas las dimensiones del modelo.
Muertes por causas violentas - homicidios (Num_Casos_Homicidios):
Esta variable mide la cantidad de muertes que ocurren por causas violentas relacionadas con homicidios (agresiones, inclusive secuelas). Esta medida es de tipo aditiva, ya que se puede sumar a través de todas las dimensiones del modelo.
Muertes por causas violentas - suicidios (Num_Casos_Suicidios):
Esta variable mide la cantidad de muertes que ocurren por causas violentas relacionadas con suicidios (lesiones autoinfligidas intencionales, inclusive secuelas). Esta medida es de tipo aditiva, ya que se puede sumar a través de todas las dimensiones del modelo.

3.4 Análisis de dimensiones y atributos

Dimensión fecha:
La tabla de fechas es importante porque mediante esta se podrá hacer referencia a información puntual dentro del modelo. Gracias a esto se podrán realizar búsquedas sobre fechas especificas o rangos de fechas y hacer análisis más profundos sobre los resultados de búsqueda.

La fuente de datos principal cuenta con una columna “año” de la cual se nutrirá la tabla de fechas, por lo que el único atributo con el que contará dicha tabla será el año.

- Manejo de historia: Para esta dimensión se utilizará el manejo de historia Tipo I, es decir, se sobrescribirán los datos actualizados. Esto se debe a que los cambios que se realicen son con el propósito de corregir información incorrecta o incompleta, por lo que no es necesario mantener versiones anteriores.

Dimensión departamento:

Dentro de las necesidades de análisis en el proceso ETL, el hecho de poder hacer un análisis sobre los datos de muertes basados en la ubicación geográfica es trascendental dentro de los objetivos del negocio. En este caso la diferenciación geográfica basada en departamento ayuda a establecer correlaciones sólidas entre el departamento y los tipos de muerte.

Dentro del CSV utilizado para el análisis existe una columna llamada departamento que contiene el nombre del departamento asociado a un registro especifico dentro de los datos, de esta columna justamente se extraerá el atributo nombre que constituye a el único atributo dentro de la dimensión además de su llave privada.

- Manejo de historia: Para esta dimensión se utilizará el manejo de historia Tipo I, es decir, se sobrescribirán los datos actualizados. Esto se debe a que los cambios que se realicen son con el propósito de corregir información incorrecta o incompleta, por lo que no es necesario mantener versiones anteriores.

Dimensión municipio:

Como se aclaró previamente en la dimensión departamento, poder establecer correlaciones entre hechos y ubicaciones geográficas da como resultado información muy valiosa para el negocio. Puesto que dicha información y los respectivos análisis que se pueden hacer con base en ella son muy atractivos para instituciones estatales.

Los datos de muertes que fueron otorgados para este análisis cuentan con una columna llamada municipio y dentro de ella se especifica el nombre el municipio. A partir de esta columna se estableció el atributo único de la tabla de dimensión municipio, su nombre.

- Manejo de historia: Para esta dimensión se utilizará el manejo de historia Tipo I, es decir, se sobrescribirán los datos actualizados. Esto se debe a que los cambios que se realicen son con el propósito de corregir información incorrecta o incompleta, por lo que no es necesario mantener versiones anteriores.

Dimensión evento:

Probablemente la dimensión más importante dentro de los requerimientos analíticos del negocio. Puesto que el análisis de los diferentes tipos de muertes en el país es la base sobre la que se funda el estudio que se quiere proponer. A partir de la diferenciación de eventos de muerte es que toman sentido las demás dimensiones y se pueden hacer búsquedas más complejas, completas y provechosas que resolverán los problemas planteados por el negocio.

En función de la tabla de datos suministrada, se extrajo el contenido de la columna evento y se convirtió en el atributo descripción de su dimensión. Adicionalmente se consideró importante agregar otro atributo, uno de tipo booleano llamado EsViolento que permita hacer una distinción dentro de los eventos y posibilite hacer análisis sobre las muertes violentas, uno de los objetivos principales planteados en el negocio. Dicho atributo se asignará arbitrariamente a todo evento que tenga una descripción que no se asocie a una muerte relacionada a causas naturales.

- Manejo de historia: Para esta dimensión se utilizará el manejo de historia Tipo I, es decir, se sobrescribirán los datos actualizados. Esto se debe a que los cambios que se realicen son con el propósito de corregir información incorrecta o incompleta, por lo que no es necesario mantener versiones anteriores.

Dimensión género:

Esta dimensión es de gran importancia para negocio para poder realizar análisis relacionados con el género. Esta información es bastante relevante para el contexto actual del país, puesto que, por ejemplo, se cuenta con un exceso de muertes por violencia de género en mujeres. Tener la posibilidad de realizar análisis acerca de estas temáticas permiten tener un mejor acercamiento a estas problemáticas para entenderlas de una forma más efectiva y clara.

Dentro de los atributos encontramos únicamente su llave primaria y el tipo, donde se indica si el género es masculino o femenino.

- Manejo de historia: Para esta dimensión se utilizará el manejo de historia Tipo I, es decir, se sobrescribirán los datos actualizados. Esto se debe a que los cambios que se realicen son con el propósito de corregir información incorrecta o incompleta, por lo que no es necesario mantener versiones anteriores.

Dimensión grupo edad:

Con esta dimensión negocio puede encontrar información relevante de acuerdo con la edad de las víctimas. Se podrán encontrar medidas que brinden un panorama más claro de cómo se ven afectadas las diferentes poblaciones de Colombia de acuerdo con su edad. De esta forma se podrá saber cuál es el tipo de incidente que afecta en mayor medida a las personas de la tercera edad, o cuál es la cantidad de menores de edad que mueren por aflicciones como la desnutrición.

Para esta dimensión encontramos los atributos de edad mínima y edad máxima, que nos permitirán delimitar los rangos para los grupos de edades para posterior procesamiento de los datos y análisis de los mismos.

- Manejo de historia: Para esta dimensión se utilizará el manejo de historia Tipo I, es decir, se sobrescribirán los datos actualizados. Esto se debe a que los cambios que se realicen son con el propósito de corregir información incorrecta o incompleta, por lo que no es necesario mantener versiones anteriores.

4 Perfilamiento de datos, creación de la base de datos y procesos ETL

Para nuestra implementación utilizamos dos herramientas: Python para preprocesamiento y organización de las dimensiones. Y BigQuery para subir la información a una base de datos y poder almacenar nuestro modelo final. A continuación, se detallan los pasos:

4.1 Perfilamiento de datos

Se presentaron las siguientes modificaciones en los datos (tanto de la primera entrega como de la segunda):

- Dado que nuestras tablas de hechos no comprenden análisis que requieran de los totales (si no que se pueden obtener por de medio de Querys simples), se opta por eliminar los registros que contienen dicha información.
- Se ajustan los códigos en la columna de departamentos para unificar ciertos valores que se encontraban con distintos formatos.
- Ajuste de los rangos de edades para la dimensión de GrupoEdad.
- Creación del atributo EsViolento para la dimensión de eventos.
- Creación de métricas aditivas en las columnas de hechos.

Al finalizar el proceso, generamos dimensiones no normalizadas, con métricas redundantes pero adecuadas para el proceso ETL y, tablas de hechos que agrupan ciertos perfiles de muertes en Colombia que son de interés para el negocio. Esto puede ser visualizado en el cuadernillo adjunto en el repositorio de anexos.

4.1.1 Diferencia de los datos entre primera y segunda entrega

Se presentaron las siguientes diferencias entre las dimensiones generadas y tablas de hechos entre la primera y la segunda entrega:

- Las dimensiones evento, municipio y fecha aumentaron la cantidad de datos y registros.
- Todas las tablas de hechos aumentaron la cantidad de datos y registros.
- En la dimensión evento, se presentan dos registros adicionales: uno correspondiente a “Tumores in situ, benignos, y los de comportamiento de origen incierto o desconocido” y el otro “Signos síntomas y afecciones mal definidas”. El primer evento afecta nuestra tabla de hechos de “Muertes por tumores, por lo que tuvimos que adicionar una métrica a la tabla.
- En la dimensión fecha, se agregaron los valores correspondientes a 2018, 2019 y 2020. Esto no afecta las métricas y las tablas de hechos, solamente son registros adicionales.
- En la dimensión municipio, se agregaron 10 registros de municipios que no afectan a la redundancia del Dataware.

4.1.2 Resultado del perfilamiento y figuras

Dentro nuestro repositorio, en la Entrega 2, en la carpeta “Carga de Datos BigQuery”, puede evidenciarse todo el proceso de perfilamiento y carga de datos. De manera paralela, se pueden visualizar las figuras de las tablas de hechos generadas y las tablas maestras. Estas figuras representan Joins entre las dimensiones y tablas de hechos que construyen el Dataware House correspondiente. Contienen datos redundantes para facilitar el procesamiento de Queries complejas, y se limitan a almacenar toda la información en un solo lugar con muchos registros.

4.2 Creación de la base de datos y carga de datos

Después de exportar el modelo perfilado del cuadernillo en Jupyter Lab, subimos nuestra información la plataforma de Google BigQuery, que cuenta con una base de datos integrada. Allí importamos las tablas las cuales crearon los Constraints necesarios entre llaves foráneas y generaron una primera instancia del modelo. En el repositorio adjunto, en la carpeta ‘Proceso ETL’ se encuentran capturas del desarrollo en BigQuery.

4.2.1 Datos cargados y justificación

Las fuentes de datos escogidas fueron 3: los datos de muerte del DANE, datos de población de Terridata y datos de porcentaje poblacional por edad de DatosMacro. El desarrollo del trabajo se fundamentaba en los datos del DANE, sin embargo, con el objetivo de realizar análisis que pudieran resultar más interesantes para el negocio se decidió relacionar ciertas columnas de los datos originales con información externa para poder presentar resultados normalizados. Específicamente, a cada departamento se le agregaron datos de su población total y a cada grupo de edad se le agregó el porcentaje que representa dentro la población colombiana.

4.2.2 Estadísticas descriptivas de los datos cargados

Posterior a la carga de datos, contamos con 5 dimensiones y 3 tablas de hechos. Dado que las tablas de hechos condensan información relevante de las dimensiones, pero carecen de valores directos (porque se tienen referencias a llaves foráneas), se opta por presentar las estadísticas descriptivas de las tablas maestras (en la sección 4.3 se explican las tablas maestras), que contienen toda la información de nuestro Dataware House.

Las estadísticas se obtuvieron con la herramienta de Python, en conjunto con la librería de Pandas. En el cuadernillo de anexos de la entrega 2 se pueden evidenciar los valores con mayor claridad. Adicionalmente en el repositorio de anexos también se pueden evidenciar dichas capturas en la carpeta “Estadísticas descriptivas tablas maestras”.

Estas estadísticas nos indican a grandes rasgos una distribución normal en el comportamiento de muertes, evidenciamos particularmente que ciertos eventos como las muertes por homicidios, al ser atípicos suelen tener valores que generen poco contraste analítico. Aún con lo anterior, estas estadísticas nos permiten tener una primera aproximación para realizar procedimientos de aprendizaje de máquina que obtengan resultados más acertados.

4.3 Implementación de los procesos ETL

4.3.1 Diseño y transformaciones

La mayoría de las transformaciones de desnormalización fueron efectuadas en la herramienta de Python (véase el cuadernillo). Sin embargo, el manejo de historia Tipo I (sobre-escritura de los datos), fue implementado específicamente en la herramienta de Google BigQuery. Adicionalmente, en Google BigQuery también efectuamos un Join entre cada tabla de hechos y sus dimensiones para crear tres *tablas maestras* finales, que contienen toda la información necesaria para efectuar Queries avanzadas. Estas tablas son particularmente útiles porque contienen toda la información de nuestro Dataware House. En el repositorio adjunto, dentro de la carpeta 'Tablas/Tablas Maestras' se pueden visualizar los .csv generados.

4.4 Comparación modelos entrega 1 y 2.

Adicional a las diferencias de preprocesamiento explicadas en la sección 4.1.1. En general los modelos implementados en ambos procesos son bastante parecidos. La diferencia entre la entrega 1 y la entrega 2 recae en la cantidad de datos que tienen. Durante la entrega 1 se utilizó información de muertes en el periodo de tiempo de 2010-2017, mientras que en la entrega 2 añade datos de 2018-2020. Solamente se agregó la métrica "Num_Casos_Tumores_Insitu_O_Desconocidos" en la tabla de hechos de "Muertes por tumores", pero, más allá de ello, la estructura general de la tabla maestra es la misma. La nueva información permite hacer análisis de tendencias más profundos y completos; además de presentar información más reciente que puede resultar más interesante para quienes quieran consultarla. Ambos modelos pueden ser evidenciados en el repositorio de anexos, especialmente en la carpeta "Carga de Datos BigQuery"; además de las tablas generadas.

5 Tableros de control y tarea de aprendizaje de máquina

5.1 Diseño de los tableros de control

Optamos por diseñar dos tableros de control nuevos (y mejorar los tres ya existentes) enfocados a dos necesidades analíticas identificadas: análisis de muertes por componente geográfico y análisis de tendencias. Los hallazgos relevantes se encuentran dentro de la presentación, a la cual se puede acceder a través del enlace que se encuentra en los anexos. Adicionalmente, en el repositorio adjunto, también en los anexos, se pueden evidenciar estos dos nuevos tableros, al igual que los tres realizados para la entrega anterior con sus respectivas correcciones.

5.2 Justificación y diseño de la tarea de aprendizaje de máquina

Se decidió implementar una tarea de predicción con el algoritmo de Random Forest, para predecir el tipo de muerte violenta que podría sufrir una persona teniendo en cuenta su género, grupo de edad y departamento de residencia. Seleccionamos esta tarea en específico debido a la relevancia que podría tener en la prevención de este tipo de muertes.

Para el diseño del modelo se implementó KNN y Random Forest, con el fin de elegir la técnica que obtuviera los mejores resultados. En ambos casos, se llevó a cabo un balanceo

de datos con la técnica de SMOTE para el preprocesamiento de los datos, al igual que K-Folds para hallar los hiperparámetros adecuados. Ambos modelos fueron ejecutados y finalmente se seleccionó Random Forest debido a sus resultados superiores.

El modelo de aprendizaje puede ser visualizado en el repositorio del curso en el cuadernillo en la carpeta “Aprendizaje de máquina”

6 Propuesta plan de proyecto

Ya que uno de los propósitos principales de Infraestructura visible es mostrar de manera transparente estadísticas para uso público, nos parecería interesante que se enfocaran en el nicho de la construcción de infraestructura pública (esto comprende cualquier tipo de infraestructura: parques, vías, edificios, mejoras en edificios, entre otros). Consideramos que puede ser un proyecto que genere impacto a nivel nacional para disminuir el componente de corrupción, y, además, que permita a las personas estar contextualizadas de la ejecución de las obras. Lo ideal sería recolectar los datos directamente con individuos en las poblaciones, que ellos jueguen el papel de “contralores” y actualicen activamente los procesos (que siempre deben estar soportados por un peso contable).

7 Repartición de tareas y puntos asignados por integrante

Repartición de tareas y puntos asignados por integrante

Integrantes	**Tareas realizadas**	**Número de horas**	**Puntos**
María Camila Parra Díaz (201819464)	• Diseño de la presentación. • Modelo de Aprendizaje de Máquina. • Identificación de necesidades analíticas y priorización de procesos de negocio • Colaboración en el documento.	20	100/3
Esteban Emmanuel Ortiz Morales (201913613)	• Diseño tablero de control de Análisis muertes por componente geográfico. • Corrección de tableros de control antiguos. • Integración de las nuevas fuentes de datos en el proceso ETL. • Identificación de necesidades analíticas y priorización de procesos de negocio.	20	100/3
Sergio Julián Zona Moreno (201914936)	• Diseño tablero de control de Análisis de tendencias. • Vídeo y presentación final. • Documento de resultados. • Identificación de necesidades analíticas y priorización de procesos de negocio.	20	100/3

Todos trabajaron de manera adecuada y realizaron aportes significativos e importantes en el proyecto. Por dicho motivo, se dividen los puntos de manera equitativa para todos los integrantes.

8 Bibliografía

• Universidad de los Andes. (2020). *Material del curso: Inteligencia de negocios*. Disponible en BLOQUE NEÓN para estudiantes de la Universidad.

9 Anexos

Enlace al repositorio:
https://github.com/SergioZona/Proyecto2_BI_ee_ortiz_mc_parrad_sj_zona

Enlace de vídeo de presentación: https://youtu.be/sAFrvsbw6nk

Enlace de la presentación:
https://www.canva.com/design/DAExzpAY4Fk/SP-tFdW3Y9l3C0AV-WqNQg/view?utm_content=DAExzpAY4Fk&utm_campaign=designshare&utm_medium=link&utm_source=sharebutton