Proyecto Ciencia de Datos Aplicada
FSFB Proyecto 2: Servicio de trasplantes
2025-2

Integrantes: Juan Diego Enriquez Ramos, Johana Alejandra Rativa, Juan Manuel Rivera López y Lina Bejarano

El siguiente proyecto de ciencia de datos hace uso de la metodología ASUM-DM, en esta primera entrega se llevan a cabo las fases: Comprensión del negocio, enfoque analítico, requisitos de los datos y comprensión y preparación de los datos. A continuación, se muestra el flujo de ejecución junto con las actividades realizadas en cada una de ellas.

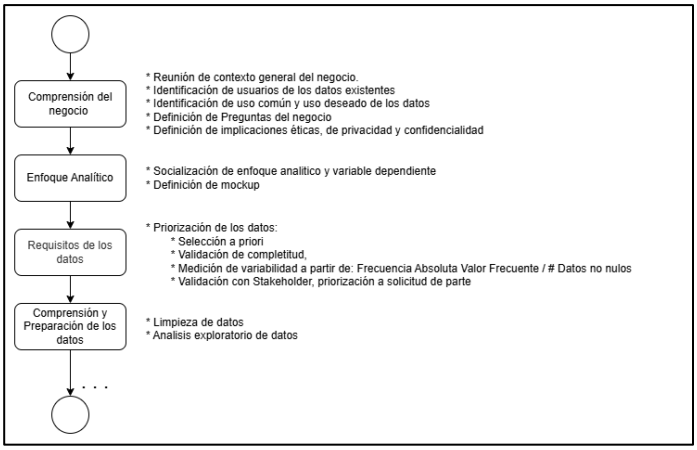

Figura 1. Metodología ASUM-DM

1. Definición de la problemática y entendimiento del negocio (Fase 1 Comprensión del negocio)

La Fundación Santa Fe de Bogotá cuenta con un servicio de cirugía hepatobiliar y de trasplantes que, de la mano de sus médicos rurales y bajo el acompañamiento de la médica Valentina Mejía, ha venido consolidando información de pacientes sometidos a trasplante hepático. Esta información, antes dispersa en diversas fuentes, fue integrada en una base de datos que representa un esfuerzo significativo de consolidación de variables operativas, clínicas, quirúrgicas y epidemiológicas. A partir de ella se busca:

- Desarrollar un panel interactivo o Power BI que permita consolidar esta base de datos, facilitando análisis clínico, operativo y epidemiológico de los pacientes sometidos a trasplante hepático.
- Facilitar la generación de reportes del año rural y obtener un panorama general poblacional de los trasplantes.
- Apoyar la toma de decisiones clínicas mediante una comprensión más clara del estado del paciente y sus posibles complicaciones postrasplante.

- Impulsar la investigación clínica al permitir el análisis de supervivencia y otros indicadores como factores de riesgo por grupos poblacionales o por año de trasplante.

Para evaluar el impacto de este proyecto de ciencia de datos se usarán los siguientes KPIs (Indicador clave de desempeño)

- Porcentaje de datos no nulos: medir el porcentaje de mejora en la completitud y calidad de los datos, incentivando una captura más limpia en futuras iteraciones a partir del uso de diccionarios de datos.
- Tiempo promedio de depuración de los datos: Medir el tiempo pre y pos desarrollo del dashboard, respecto a la duración del proceso de limpieza/estandarización de la data.
- Tiempo promedio de entendimiento del estado del paciente: Medir el tiempo promedio en el que se pueden entender los datos del paciente, ya que la herramienta facilita la lectura integral de un caso clínico a través de un perfil dinámico que reúne variables significativas 8pre y pos desarrollo de la herramienta)
- Número de hipótesis o estudios derivados: cuantificar el aporte del análisis exploratorio y del dashboard en la investigación clínica a partir de la participación en eventos de investigación (abstracts enviados, manuscritos sometidos/publicados) y generación de nuevas líneas o proyectos de investigación clínica.
- Frecuencia de uso de los datos: medir el aumento del acceso y aprovechamiento de la información consolidada, antes subutilizada por su complejidad y volumen.

2. Ideación

Dentro del servicio de cirugía hepatobiliar y de trasplantes, los médicos rurales constituyen el principal grupo de usuarios que alimenta y consulta la base de datos consolidada de trasplante hepático. Como se mencionó previamente, estos usuarios realizan dos usos principales de la información: generan reportes operativos y realizan investigación clínica.

Adicionalmente, los médicos buscan que esta información sirva para identificar tempranamente factores de riesgo o vulnerabilidad frente a posibles complicaciones postrasplante. Sin embargo, esta necesidad se ve limitada por la complejidad y extensión de la base de datos, así como por problemas de calidad y limpieza de los registros, lo que dificulta el análisis y la identificación de patrones relevantes. Por eso se plantea como solución el uso de un tablero de control en power BI en donde se proponen 3 vistas principales.

- Panorama general de trasplantes: permite la generación de reportes operativos y la descripción poblacional de las cirugías realizadas. Incluye filtros por grupo etario, sexo, año y otras variables demográficas o clínicas de interés.
- Perfil del paciente: muestra información individual asociada a complicaciones (infección, cáncer postrasplante, inmunodeficiencia, entre otras) y a las variables identificadas como relevantes en las etapas de análisis exploratorio y modelamiento. Facilita la comprensión del estado del paciente y de sus riesgos asociados.
- Análisis de sobrevida poblacional: proyecta los resultados del análisis exploratorio, mostrando tasas de supervivencia por grupo etario, sexo o año de trasplante, junto con las dependencias o variables que más inciden en estos resultados.

Cabe aclarar, que antes de cargar la base de datos a power BI se desarrolla la elección de variables clínicas relevantes y un script en python para llevar a cabo un proceso de limpieza de la data.

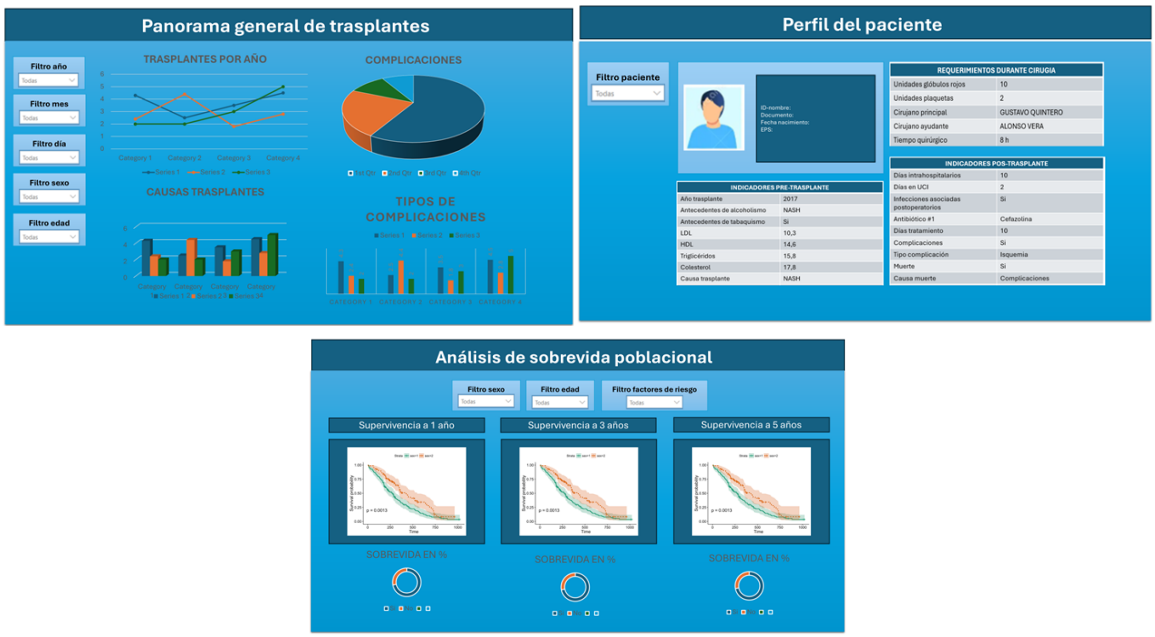

Figura 2. Mockup Power BI

Adicional al tablero de control se propone optimizar un modelo de regresión logística que nos permita calcular la probabilidad de que un paciente presente complicaciones, junto con un API-rest. Este API permitirá consultar datos del paciente y usar el modelo de regresión logística.

3. Responsable

De acuerdo con la Resolución 8430 de 1993 del Ministerio de Salud de Colombia, por la cual se establecen las normas científicas, técnicas y administrativas para la investigación en salud, y específicamente conforme al Capítulo I, Artículo 11, la investigación realizada por la Fundación Santa Fe de Bogotá a partir de los datos de pacientes sometidos a trasplante hepático se clasifica como una investigación sin riesgo, dado que la obtención de la información se efectuó de manera retrospectiva, empleando datos procedentes de historias clínicas.

Para el acceso a la información, el Servicio de Trasplantes suscribió un acuerdo de confidencialidad con el equipo de trabajo, en el cual se establecen cláusulas específicas sobre el uso ético, la reserva y la protección de los datos entregados. Adicionalmente, la base de datos proporcionada fue previamente anonimizada por el equipo clínico responsable, garantizando así que no fuera posible la identificación directa o indirecta de los pacientes.

Nota: Se anexa el acuerdo de confidencialidad firmado por las partes.

4. Enfoque analítico

El análisis se orienta a explorar y modelar los factores asociados a la presencia de complicaciones postoperatorias, con el objetivo de identificar las variables más relevantes, optimizar su visualización en el tablero de control y generar insights que contribuyan a mejorar la respuesta clínica de los pacientes. Inicialmente, se realizará una priorización de datos con base en su completitud y relevancia para los stakeholders.

Posteriormente, se aplicará un análisis exploratorio univariado y bivariado centrado en la variable dependiente identificada —"¿Complicación?"— para detectar patrones, distribuciones y relaciones entre variables clínicas. Dada la naturaleza y complejidad de los datos clínicos, se implementará una Regresión Logística con el fin de evaluar el impacto multivariado de las variables predictoras sobre la probabilidad de complicación.

Las hipótesis plantean que factores como la edad, el estadio del tumor, los antecedentes de consumo (como tabaquismo) y marcadores bioquímicos como la alfafetoproteína pueden influir significativamente en el riesgo de complicación. El modelo será evaluado mediante métricas de desempeño como exactitud (accuracy), precisión, recall y AUC-ROC, para determinar su capacidad predictiva y robustez estadística.

5. Recolección de datos

La base de datos de trasplantes hepáticos fue recibida el 25 de septiembre de 2025, con un total de 736 registros y 285 variables clínicas. A partir de esta información se diseñó una estrategia de priorización de datos estructurada en cuatro etapas principales: Selección inicial de variables, basadas en el análisis de interés, evaluación de completitud de los datos, medición de la variabilidad y consistencia de los datos y validación con el stakeholder.

En primera instancia, para dar inicio con el proceso de limpieza y EDA, se eligieron 114 variables en total. Posterior a ello se evaluó la presentación de datos nulos en las variables y se calculó la frecuencia relativa del valor más común sobre los datos no nulos. En este caso, variables donde el valor más frecuente estaba presente en más del 40% de los datos, se plantean omitir de los análisis. Al final, se hizo contactó con el stakeholder de la FSFB para validar la pertinencia de las variables. Como resultado, se obtuvo una base depurada de 121 variables seleccionadas para los siguientes pasos del proyecto.

6. Entendimiento de los datos

Para esta sección, en el repositorio github MINE4101-202520-Proyecto-Transplantes-Hepaticos se encuentra el Notebook Limpieza inicial de datos consolidado.ipynb en el que es evidente el proceso en el que se evalúa la calidad de los datos y se realiza la limpieza de las variables seleccionadas para el proyecto. Adicional a ello, se diseña un reporte en power point (adjunto) del EDA inicial de los datos que incluye técnicas de análisis univariadas/multivariadas/gráficas/no gráficas.

7. Conclusiones iniciales

La base de datos entregada presenta varios retos en términos calidad de los datos y nulidad, en algunos casos incluso consistencia entre variables. Se requiere una segunda revisión de la limpieza llevada a cabo hasta el momento.

De acuerdo con el análisis, se encuentra que la mayoría de los pacientes que requieren trasplante hepático se encuentran en el grupo etario de 60 a 74 años, respecto al sexo de los receptores, no predomina ninguno, se presentan porcentajes similares (Masculino=50.8% y Femenino=49.2%).

De acuerdo con los registros el año en que más se realizaron trasplantes fue el año 2017 y en el 2020 se redujeron, probablemente por el problema de salud pública presente para el momento (COVID-19). La mayoría de quienes se benefician de este procedimiento quirúrgico provienen de la EPS sanitas. En el caso del tiempo en el que puede durar un paciente en enlistamiento de trasplante es de 0 años, lo que quiere decir que es un procedimiento que se lleva a cabo con prontitud, sin

embargo, hay casos en los que los pacientes requieren hasta 4 o más años para poder acceder al procedimiento.

La principal causa por la que los pacientes son candidatos a trasplante es por NASH (esteatohepatitis no alcohólica)-21.1%-. Con frecuencia los pacientes antes de un trasplante no presentan hepatocarcinoma.

Al revisar si algunas de estas variables de interés presentan diferencias de acuerdo con si hubo o no complicaciones postoperatorias. Ninguna de estas variables resalta en los análisis, no obstante, los pacientes donde más se presentan complicaciones es en el grupo con edad de 60 a 74 años.

En el caso de las variables “Soporte_Vasopresor_PeriTx”, “Noradrenalina_PeriTx” y “Número de Complicaciones”, se encontró una relación estadística con la variable objetivo “¿Complicación?”. Además, con el gráfico de barras de la variable número de complicaciones se encontró mayor cantidad de casos de complicación cuando se presentó 2 complicaciones.

En consecuencia, para los próximos análisis especialmente, para la realización del modelo, se requiere identificar más variables que puedan relacionarse con las complicaciones de los pacientes a partir del uso de métodos multivariados que permitan abarcar la complejidad de los datos clínicos.