Universidad de los Andes

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

1. Definición de la problemática y entendimiento del negocio

La Fundación Santa Fe de Bogotá cuenta con el Servicio de Cirugía Hepatobiliar y de Trasplantes, el cual realiza alrededor de 30 trasplantes de hígado al año. El personal del Servicio ha consolidado una base de datos extensa, que recoge distintas variables operativas, clínicas, quirúrgicas y epidemiológicas, permitiendo detallar en la cirugía de cada paciente. Sin embargo, la consulta y actualización de la información puede tomarles a los médicos entre 1 y 2 horas por paciente, lo cual limita el uso que se le dan a los datos.

Por esto, se trabajó en conjunto con la médica Valentina Mejía para plantear el presente proyecto, en el que se busca:

- Desarrollar un panel interactivo en Power BI que permita analizar diferentes variables de la base de datos, facilitando el 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.
- Identificar factores que están relacionados con la presentación de complicaciones postrasplante, y plantear intervenciones relacionadas a estos.
- 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.

A partir del entendimiento de los objetivos institucionales, se identifican las siguientes métricas de éxito (KPIs) que pueden verse fortalecidas mediante la implementación de la solución de datos:

- **Reducción del tiempo de hospitalización**: un mayor tiempo de hospitalización posquirúrgica supone un mayor costo para la Fundación. Se espera que el producto de datos ayude a entender qué factores afectan el tiempo que pasa un paciente hospitalizado (tanto en UCI como en habitación). Asimismo, el producto permitirá evaluar si programas como el Fast Track (paso de un paciente directo a habitación) se refleja en diferentes métricas de éxito.
- **Disminución de la tasa de mortalidad postoperatoria**: la tasa de mortalidad es un KPI para el Servicio de Trasplantes. Se espera que el producto de datos permita identificar variables asociadas a la presentación de complicaciones, y con esto plantear intervenciones que puedan disminuir las complicaciones y por ende la mortalidad. Asimismo, permitirá hacer seguimiento a estas intervenciones.
- **Reducción del tiempo de elaboración de reportes clínicos**: la consulta de la base de datos consolidada puede tomar entre 1 y 2 horas por paciente. El dashboard centraliza la información y automatiza la consulta de datos, disminuyendo

significativamente el tiempo requerido para preparar reportes y aplicar filtros sobre los datos.

- **Aumento de la productividad científica**: si bien al Servicio no se le mide por número de publicaciones, para los médicos esta es una métrica de interés. El acceso a una base de datos estructurada y depurada facilita el análisis de los datos, de forma que se puedan compartir datos clínicos de alta calidad con la comunidad científica.

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 usan esta información para generan reportes operativos y realizar investigación clínica. Adicionalmente, los médicos buscan que esta información sirva para identificar tempranamente factores de riesgo que lleven a complicaciones postrasplante.

Sin embargo, dada la complejidad y extensión de la base de datos, así como problemas de calidad de los registros, el análisis y la identificación de patrones relevantes se dificulta. Para permitir una mejor consulta de los datos se plantea como un tablero de control en Power BI que incluya estas 4 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.
- **Pre trasplante:** muestra información asociada a antecedentes de los pacientes, tiempo en lista de espera (en días, meses y años), causas o etiologías de la enfermedad hepática y gravedad de esta (Child-Pugh y MELD score). Esto facilita la evaluación de los riesgos asociados, con el objetivo de aumentar la probabilidad de éxito de los trasplantes.
- **Intra quirúrgico:** proyecta variables asociadas al procedimiento quirúrgico como qué cirujanos realizaron el rescate y el procedimiento principal, el tiempo de isquemia fría y caliente del órgano a trasplantar, los antibióticos utilizados entre otras. Proporciona comprensión de lo que ocurre durante la cirugía para mejorar los resultados clínicos, con el objetivo de reducir complicaciones y tomar decisiones basadas en datos.
- **Pos trasplante:** visualiza diferentes variables asociadas al desenlace del paciente, como si hubo rechazo agudo o crónico, gráficas de Kaplan-Meier, y si hubo falla cardíaca, arritmias o infarto. Brinda la facilidad de evaluar los riesgos que enfrentan los pacientes tras la intervención, con el objetivo de mejorar los resultados futuros.

Adicional al tablero de control se optimizarán diferentes modelos de clasificación binaria que nos permitan analizar el peso que tienen diferentes variables en la posibilidad de que el paciente presente complicaciones. Adicionalmente, se expondrá un servicio API para predecir a partir de varias variables si el paciente tendrá o no complicaciones.

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.

4. Enfoque analítico

El análisis se guía por la siguiente pregunta de negocio: *¿Qué factores clínicos y demográficos se asocian con un mayor riesgo de complicación postoperatoria?*

A partir de esta pregunta, se plantean hipótesis que sugieren que variables como edad, estadio del tumor, antecedentes de consumo (tabaquismo o alcohol), marcadores bioquímicos como la alfa-feto proteína, obesidad, entre otras, pueden influir significativamente en la probabilidad de que se presenten complicaciones.

Se propone un enfoque analítico mixto que combina técnicas estadísticas y de visualización para caracterizar el panorama clínico, junto con modelos de machine learning supervisado para predecir complicaciones, experimentando con cuatro algoritmos: *Regresión Logística, Random Forest, Árboles de Decisión y AdaBoost*.

La métrica principal para seleccionar el mejor modelo será el F1-score ponderado, dado que permite equilibrar precisión y recall en un contexto con desbalance de clases, maximizando la correcta identificación de pacientes con complicación sin incrementar los falsos positivos. Adicionalmente, se reportarán métricas complementarias como accuracy, precisión, recall de cada modelo, y la matriz de confusión del mejor modelo.

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:

1. Selección inicial de variables, basadas en el análisis de interés
2. Evaluación de completitud de los datos
3. Medición de la variabilidad y consistencia de los datos
4. Validación con el stakeholder.

Para dar inicio con el proceso de limpieza y análisis exploratorio (EDA) se realizó una selección de variables de interés. Esta selección tuvo en cuenta el número de datos nulos y la variabilidad de los datos válidos. Finalmente, se contactó con el stakeholder del Servicio de Trasplantes para validar la pertinencia de las variables seleccionadas, y añadir otras que pudieran haberse dejado de lado. Como resultado, se obtuvo una base depurada de 121 variables seleccionadas para los siguientes pasos del proyecto.

6. Entendimiento de los datos

Se realizó un proceso de limpieza de las variables seleccionadas, que se enfocó en una limpieza de valores no válidos. Esta limpieza puede verse en detalle en el Notebook Limpieza inicial de datos consolidado.

Posterior al proceso de limpieza se realizó un análisis exploratorio de datos (EDA). Se encontró que la mayoría de los pacientes que requieren trasplante hepático tienen entre 60 y 74 años. Asimismo, la mayoría de los pacientes que presentan complicaciones se encuentran en ese rango de edad. En cuanto al sexo de los receptores, la muestra está balanceada (masculino = 50,8 % vs. femenino = 49,2 %). La mayoría de los beneficiarios del procedimiento pertenecen a la EPS Sanitas.

El año con mayor número de trasplantes registrados fue 2017 y se evidencia que en 2020 se redujeron notablemente, probablemente debido a la emergencia sanitaria por COVID-19. La mayor parte de los trasplantes se realizan en menos de 1 año (mediana del tiempo de espera = 0 años), lo que indica que el procedimiento suele realizarse con prontitud; sin embargo, existen casos en los que los pacientes esperan 4 años o más para acceder al trasplante.

La principal etiología que motiva el trasplante es la NASH (esteatohepatitis no alcohólica), con un 21,1%. Por otro lado, se identificó que aplicar vasopresina y noradrenalina durante la cirugía tiene una relación estadísticamente significativa con la presencia de complicaciones.

7. Preparación de los datos

Se realizó una limpieza para identificar datos no válidos, de manera que se imputaran a valores válidos cuando se pudiera (por ejemplo, llevando métricas a rangos válidos). Además, se homogenizaron variables clínicas binarias (llevarlas a valores de 0 y 1).

Con este dataset depurado, se realizó una preparación adicional para entrenar los clasificadores binarios:

1. Se imputó la variable de interés (presencia de complicaciones posquirúrgicas) a partir de otras variables relacionadas. Por ejemplo, si un paciente tuvo una falla cardiaca, un infarto, arritmias o infecciones, se asume que presentó complicaciones.
2. Selección inicial de variables según su tipo (numéricas, booleanas y ordinales), excluyendo aquellas que estuvieran en lenguaje natural o que tuvieran información operativa (por ejemplo EPS).

3. Procesamiento de variables categóricas para convertirlas en numéricas, aplicando una estrategia de *One Hot Encoding*.
4. Filtrado de variables por número de entradas nulas, excluyendo aquellas categorías con más de 16% valores faltantes.
5. Eliminación de registros con valores nulos, ya que los modelos a usar no permiten valores nulos al realizar el entrenamiento.

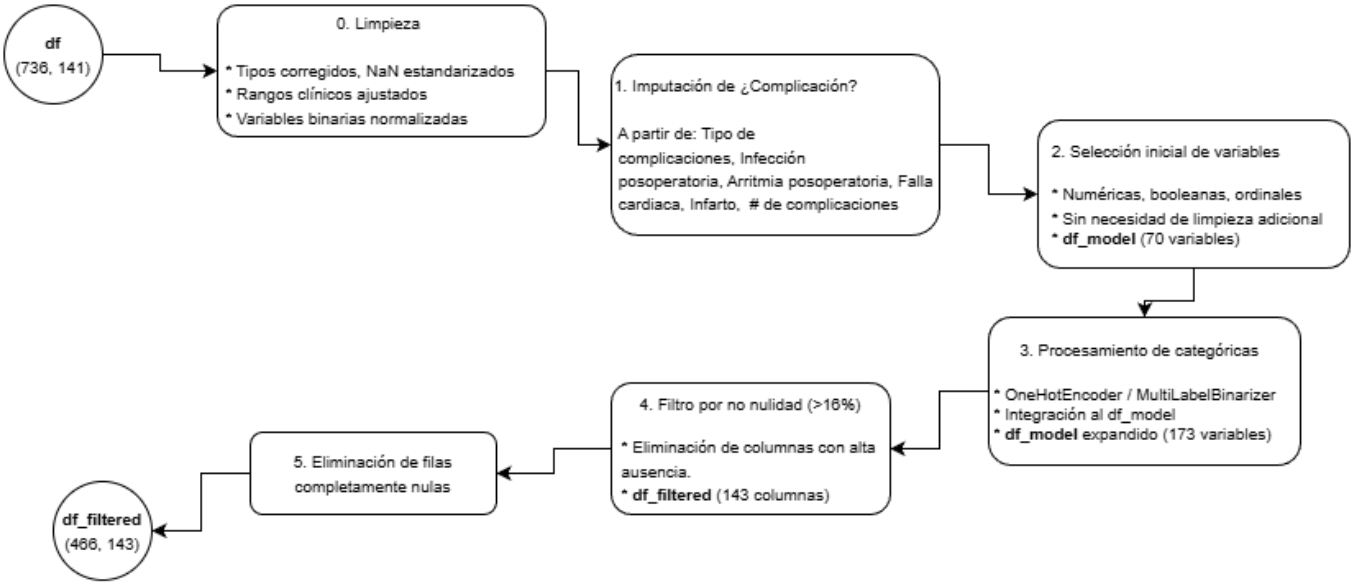

Figura 1. Diagrama de flujo preparación de datos

El resultado final fue un conjunto de datos con 466 entradas y 143 variables. Es importante notar que luego del paso 1 la variable objetivo se volvió desbalanceada, de manera que 60% de los pacientes presentaron complicaciones.

Este conjunto es la base final para el modelado y fue dividido en dos subconjuntos: un 70% para entrenamiento y 30% para validación. Se garantizó la conservación de la distribución original de la variable objetivo en ambos conjuntos.

Dataset	¿Complicación? = 1		¿Complicación? = 0		Total
Conjunto inicial (Con imputación de consistencia)	410	60%	268	40%	678
Conjunto de datos limpio y filtrado	268	57%	198	43%	466
Train (70%)	187	57%	139	43%	326
Validation (30%)	81	57%	59	43%	140

8. Estrategia de validación y selección de modelo

Se evaluó el desempeño de 4 modelos usados para clasificación binaria: regresión logística, Random Forest, árboles de decisión y AdaBoost.

Tomando el 70% de los datos para entrenamiento, se experimentó con diferentes hiperparámetros para los diferentes modelos a analizar. En cada etapa del entrenamiento se dividieron los datos siguiendo una estrategia de validación cruzada, de forma que en cada paso el conjunto de datos de entrenamiento se dividió en 5 particiones. Esto permite tener un desempeño más cercano al que tendrá con datos nuevos.

Los diferentes modelos se compararon usando el puntaje F1 ponderado. Este se seleccionó ya que es un puntaje que tiene en cuenta tanto la precisión como el recall. Además, aplica un peso a las métricas globales usando la proporción de cada categoría, para contrarrestar el desbalanceo.

9. Construcción y evaluación del modelo

Luego de seleccionar los mejores hiperparámetros para cada modelo, se evaluó el desempeño de estos con el conjunto de datos de validación, los cuáles no se habían usado durante el entrenamiento.

De los cuatro algoritmos evaluados el Random Forest obtuvo el mejor desempeño, alcanzando un puntaje F1 ponderado de 69,3%.

Algoritmo	Puntaje F1 ponderado	Recall ponderado	Precisión precisión	Accuracy
Regresión logística	64,2	64,3	64,1	64,3
Árbol de decisión	62	62,1	62,1	62,1
Random Forest	69,3	69,3	69,4	69,3
AdaBoost	68,3	68,3	68,3	68,6

Al analizar la matriz de confusión (Figura 2), se observa que el modelo logró identificar correctamente 59 de los 81 pacientes que sí presentaron complicaciones, lo que equivale a un recall del 73%, es decir, capturó la mayoría de los casos reales positivos.

Por otro lado, el modelo clasificó como complicados a 21 pacientes que en realidad no presentaron complicaciones, lo cual se refleja en una precisión del 74%, indicando que tres de cada cuatro predicciones positivas fueron correctas.

En conjunto, estos resultados muestran que **Random Forest** ofrece un equilibrio adecuado entre sensibilidad y precisión, siendo el algoritmo con mejor capacidad predictiva para este caso.

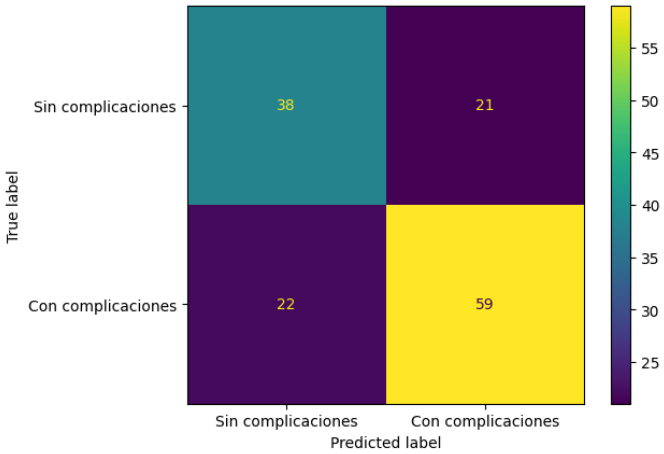

Figura 2. Matriz de confusión modelo ML implementado

Es importante notar que la variable objetivo fue imputada debido a la calidad inicial de los datos. Una mejor depuración de la base de datos inicial llevaría a mejores modelos, que puedan identificar mejor las variables asociadas con las complicaciones.

Por otro lado, el modelo propuesto tiene en cuenta tanto variables pre quirúrgicas, intra quirúrgicas y posquirúrgicas, por lo que el modelo puede estar sesgado por otras variables asociadas al desenlace del trasplante. Se podría plantear a futuro un modelo entrenado únicamente sobre variables prequirúrgicas, o intra quirúrgicas. Esto permitiría predecir complicaciones previo a la cirugía y entender mejor variables específicas de cada etapa.

10. Construcción del producto de datos

- **Dashboard (Power BI)**

Acorde con la retroalimentación recibida por parte de los *stakeholders* se replanteó el Mockup presentado. Este Dashboard se realizó empleando Power BI, enlazado con un script de Python que permitió hacer la limpieza y la codificación de las variables necesarias para realizar el reporte. Este reporte se divide en tres hojas principales: Panorama general, panoramas y supervivencia poblacional.

En la primera hoja, de panorama general se presentan características base de los pacientes como: edad, sexo, aseguradora y grupo sanguíneo.

La segunda hoja, permite visualizar gracias a paneles interactivos, las diferentes variables presentes en los eventos de pre-trasplante, intra-quirúrgico y pos-trasplante de acuerdo con el número de trasplantes. Adicional a una visualización relacionada con los donantes del órgano.

Finalmente, se muestra la hoja de supervivencia poblacional, que, de acuerdo con la edad, permite ver la supervivencia global de los pacientes a 1, 3 y 5 años, post cirugía.

Es importante destacar que al abrir el Power BI es necesario tener instalada una versión de Python, ya que la sección de supervivencia poblacional se realizó con un script empleando este lenguaje de programación. Al abrir el Power BI, es necesario realizar la configuración.

Indicaciones: Al abrir el Power BI, ir archivo > opciones y configuraciones > opciones > creación de scripts > seleccionar ruta y versión de Python > Aceptar.

Adicional a esto, es necesario actualizar las rutas a la carpeta donde está el Power BI, dado que de lo contrario no permitirá actualizar.

Ruta: Transformar Datos > Escoger tabla > Source y actualizar el path en el código que aparece (ver Figura 3).

Figura 3. Visualización Power BI

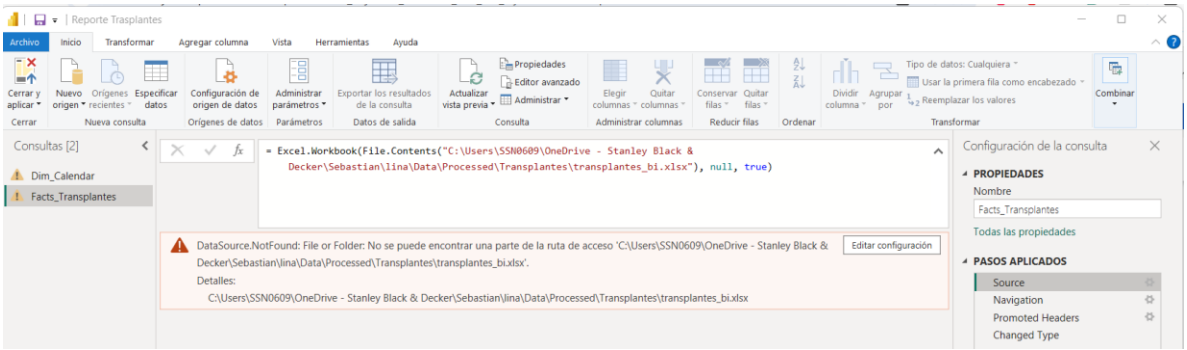

Figura 4. Cambio de ruta carpeta Power BI

- **Despliegue del modelo**

El mejor modelo obtenido de la experimentación se exportó y cargó en un servidor que expone un API. Este servidor permite recibir solicitudes con distintas variables de los pacientes, y calcula la probabilidad de que el paciente presente complicaciones.

Sin embargo, como parte de la retroalimentación dada por el *stakeholder*, se evidencia que tanto el modelo como el API para consultarlo no presentan un valor adicional para el Servicio de Trasplante. Habría que identificar cómo el modelo puede incorporarse dentro del flujo del servicio.

Sin embargo, el análisis del peso de las diferentes variables en la predicción del modelo sí fue de mayor interés, ya que permite identificar oportunidades de mejora en factores que afectan la presentación de complicaciones

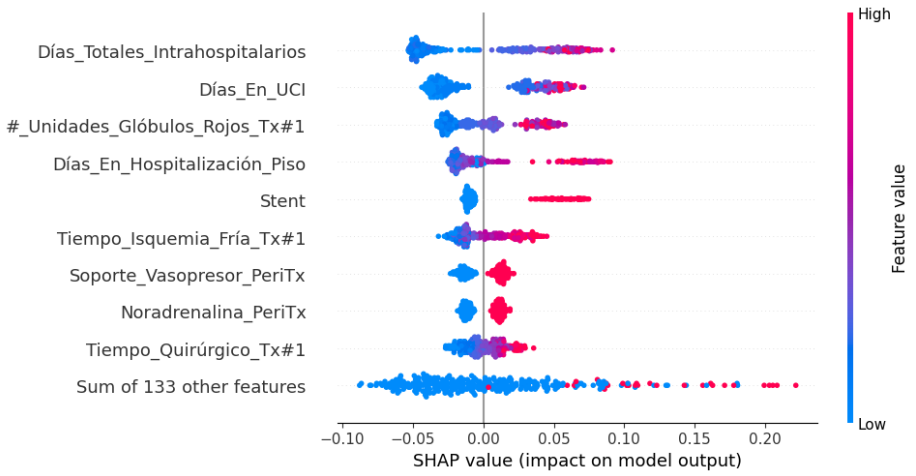

Figura 5. Principales variables que influyen en la predicción de complicaciones

Resulta interesante como variables que tenían una relación estadísticamente significativa con complicaciones, como el uso de noradrenalina o de soporte vasopresor durante la cirugía, también aparecieron en las variables más representativas del modelo.

11. Retroalimentación por parte de la organización

Fecha	Tipo de interacción	Stakeholders	Objetivo de la interacción	Acuerdos / Resultados clave
10 septiembre 2025	Reunión sincrónica	Valentina Mejía Diana Bejarano	Contexto clínico y definición inicial del proyecto	Se identifican usuarios finales (médicos rurales), se define problemática general y se prioriza la base *Trasplante hepático consolidado.xlsx*.
15 octubre 2025	Reunión sincrónica Correo, firma de acuerdo	Valentina Mejía	Validar alcance, variables y	Aprobación formal del proyecto. Definición de variable dependiente (*¿Complicación?*). Alineación de preguntas de

Fecha	Tipo de interacción	Stakeholders	Objetivo de la interacción	Acuerdos / Resultados clave
			preguntas de negocio	negocio (panorama general, complicaciones y sobrevida).
27 noviembre 2025	Reunión sincrónica, correo de retroalimentación	Valentina Mejía, Dr. Vera, Diana Bejarano	Presentación del avance del dashboard	Se valida el diseño general. Se ajusta estructura de pestañas de Power BI (Panorama, Pre, Intra y Post). Solicitud de incluir variables clave (MELD, Child-Pugh, complicaciones_POP, donante, curvas Kaplan–Meier). Se coordina fecha para ajustes finales.

Adicionalmente se contó con el apoyo constante vía WhatsApp con la médica Valentina Mejía asegurando el entendimiento de las variables y validando la toma de decisiones como imputación de la variable complicaciones.

12. Conclusiones

En este proyecto se desarrolló un producto de datos compuesto por un panel en Power BI, un modelo Random Forest para predecir complicaciones y un diccionario de datos para el Servicio de Trasplante Hepático de la Fundación SantaFé de Bogotá.

El tablero de control tiene como propósito apoyar en la generación de reportes, fortalecer la investigación clínica y optimizar la toma de decisiones a partir de los datos, dado que el conjunto original de datos es difícil de trabajar dada su dimensionalidad. El modelo para predecir complicaciones usa un algoritmo de Random Forest, y permitió identificar variables asociadas con la variable complicaciones.

En relación con el cumplimiento de los objetivos, estos fueron alcanzados, logrando el entendimiento de la base, la exploración de los datos, la mejora del diccionario, el desarrollo del Power BI con las secciones Donante, Panorama general de trasplantes, Pretrasplante, Intraquirúrgico y Postrasplante, y la construcción de un modelo aceptable para la predicción de la variable objetivo.

Las principales dificultades encontradas fueron la calidad de los datos y los valores faltantes en la base de datos entregada por la Fundación Santa Fe de Bogotá. En algunos casos también se detectó falta de consistencia entre variables; por ejemplo, nuestra variable objetivo “complicaciones” presentaba bastantes inconsistencias, lo que afectó el desarrollo del Power BI y los 4 modelos evaluados.

Con los productos creados, se espera que el panel desarrollado en Power BI contribuya al cumplimiento y posible disminución de la mortalidad, al permitir un mayor análisis del proceso clínico pre, intra y postrasplante, aportando a la toma de decisiones basada en

datos. Además, se espera reducir el tiempo de hospitalización al menos en un 10 %, gracias a que permite la detección temprana de variables clínicas o quirúrgicas que incrementan el tiempo de hospitalización de los pacientes. Asimismo, facilite la identificación de los parámetros que incrementan la mortalidad en los pacientes trasplantados. Adicionalmente, se espera potenciar el análisis de variables o patrones para la creación de artículos científicos, con el apoyo del Power BI y del modelo de predicción de complicaciones.

Para obtener mejores resultados y aumentar la capacidad predictiva de los modelos, es necesario que la base de datos cuente con información más consistente, con menos valores nulos y menos sesgos. El principal problema radica en su diligenciamiento, ya que requiere mucho tiempo y genera numerosos errores tipográficos; al tratarse de una base con muchas variables, cualquier distracción incrementa las equivocaciones. Esto podría mejorarse significativamente mediante un pipeline superior e, idealmente, con la implementación de un software más amigable y diseñado específicamente para facilitar el llenado de la base de datos.

De los modelos implementados, el mejor fue Random Forest, con un puntaje F1 (ponderado) de 68 %. El modelo mostró buenos resultados en la predicción de la variable objetivo con las pruebas realizadas, aunque se requieren nuevas validaciones con el dataset actualizado que incluya los casos de 2025, ya que actualmente solo se contaba con datos hasta 2024. En caso de que se lleve el modelo a producción, se recomienda implementar estrategias como Canary Deployment, que consiste en un despliegue progresivo con datos reales nuevos, escalando de forma gradual conforme se valida su desempeño. Además, el uso de los resultados del modelo en el ámbito de investigación presenta una oportunidad considerable. Sin embargo, debe tenerse presente la imputación realizada, y habría que realizar una limpieza de los datos.