Proyecto Ciencia de Datos Aplicada

Informe Entrega 1

TRASPLANTES HEPÁTICOS FUNDACIÓN SANTA FÉ

Integrantes:
Daniel Esteban Aguilera Figueroa - d.aguilera@uniandes.edu.co
Diego Felipe Carvajal Lombo - df.carvajal@uniandes.edu.co
Jesús Manuel Ospino Bernal - jm.ospino@uniandes.edu.co
María Alejandra Pérez Petro - ma.perezp@uniandes.edu.co

1. Definición de la problemática y entendimiento del negocio

1.1. Contexto y relevancia de la problemática

La Fundación Santa Fe de Bogotá es una institución privada que busca aportar de manera positiva al sistema de salud para mejorar el bienestar de las personas y sus comunidades (Fundación Santa Fe de Bogotá, s.f). Entre sus servicios médicos se encuentra la atención a pacientes con enfermedades hepáticas, incluyendo, cuando es necesario, la realización de trasplantes de hígado. Actualmente, la organización busca comprender la incidencia y los factores de riesgo asociados a las infecciones postrasplante hepático, una complicación de alto impacto clínico que afecta hasta el 71% de los pacientes en los seis meses posteriores al procedimiento (Tezcan et al., 2023). Estas cifras evidencian la necesidad de profundizar en el estudio de los factores de riesgo que predisponen a los pacientes trasplantados de hígado a desarrollar infecciones, con el objetivo de identificar patrones, diseñar estrategias preventivas y proponer intervenciones clínicas diferenciadas según los perfiles de riesgo identificados.

1.2. Objetivo del Proyecto

Aprovechar las técnicas de ciencia de datos para analizar, estructurar y optimizar la información clínica de los pacientes sometidos a trasplante hepático en la Fundación Santa Fe de Bogotá, con el fin de identificar factores de riesgo asociados a infecciones post-trasplante y fortalecer la toma de decisiones clínicas y administrativas a partir de evidencia basada en datos.

1.3. Indicadores clave (KPIs)

- **Reducción de la tasa de infecciones postrasplante:** Disminuir la proporción de pacientes con infecciones dentro de los seis meses posteriores al trasplante hepático, pasando de 14% a 10%.
- **Disminución de infecciones severas o con reingreso a UCI:** Reducir los casos de infecciones que requieren reingreso hospitalario del 62% al 55%.
- **Incremento de la supervivencia a 12 meses postrasplante:** Aumentar la tasa de pacientes vivos al año del trasplante del 67% al 70%.
- **Optimización de la estancia en UCI postoperatoria:** Reducir el promedio de días en UCI de 4 a 2 días, fortaleciendo los protocolos de recuperación temprana.
- **Reducción del impacto de comorbilidades en el riesgo infeccioso:** Disminuir la brecha de incidencia de infecciones entre pacientes con y sin comorbilidades (diabetes, tabaquismo, alcoholismo).

2. Ideación

2.1. Producto de datos a desarrollar

Producto 1: Proceso de recolección de datos utilizando herramientas de PowerApps (mockup)

Descripción	Usuarios Potenciales	Requerimientos
Aplicación digital para la captura estandarizada de información de pacientes trasplantados. Permite registrar, validar y actualizar los datos en tiempo real, asegurando la consistencia y trazabilidad de la información clínica.	Personal médico y administrativo encargado del registro de pacientes.	- Formularios estandarizados con validaciones de campos y reglas de consistencia. - Integración con bases de datos existentes (SQL Server / Excel / SharePoint). - Control de acceso y permisos según rol del usuario. - Interfaz intuitiva y de fácil uso.

Producto 2: Modelo de Machine Learning

Descripción	Usuarios Potenciales	Requerimientos
Algoritmo predictivo diseñado para identificar grupos de pacientes según características clínicas y demográficas, y clasificar el riesgo de infección postrasplante hepático. Permite generar conocimiento a partir de los datos históricos y apoyar la toma de decisiones clínicas basadas en evidencia.	Equipo médico e investigadores del área de hepatología.	- Entrenamiento y validación del modelo con una base de datos limpia y anonimizada. - Clasificación de nuevos pacientes en niveles de riesgo. - Documentación técnica con métricas de desempeño y guías de interpretación. Este último es bastante importante debido a que los usuarios potenciales pueden no tener conocimiento técnico en el tema.

Producto 3: Tablero Power BI (mockup)

Descripción	Usuarios Potenciales	Requerimientos
Herramienta interactiva de visualización que integra indicadores clave (KPIs) permitiendo monitorear la evolución clínica y apoyar decisiones estratégicas del hospital.	- Jefe de departamento. - Equipo médico tratante - Investigadores del área de hepatología.	- Conexión en tiempo real o programada con la base de datos central. - Visualización de KPIs: tasa de infección, promedio de días en UCI, comorbilidades, estado vital, etc. - Filtros dinámicos y segmentaciones por periodo, tipo de infección o perfil de paciente. - Exportación automática de reportes en PDF o Excel.

2.2. Procesos actuales y desafíos

- **Fragmentación de datos:** Actualmente la información se gestiona en hojas de cálculo dispersas, dificultando su integración.
- **Falta de estandarización:** Las variables y formatos no son uniformes, lo que complica el análisis comparativo entre pacientes o periodos.
- **Ausencia de trazabilidad:** No se cuenta con un seguimiento estructurado de la evolución clínica de los pacientes, se cuentan con variables generales que indican el estado del paciente en cierto punto luego del trasplante, pero no se continua un estudio periódico.
- **Baja calidad de datos:** Existen registros incompletos o inconsistentes que afectan la fiabilidad del análisis.

2.3. Componentes tecnológicos

- **Infraestructura y almacenamiento:** Uso de Microsoft Azure, SharePoint o SQL para el alojamiento seguro y centralizado de los datos.
- **ETL y procesamiento de datos:** Implementación de flujos de transformación con Python (Polars, Pandas) o Power Query para limpiar, combinar y estructurar la información.
- **Modelado y análisis:** Uso de bibliotecas como Scikit-learn para construir, entrenar y evaluar modelos predictivos de riesgo de infección.
- **Visualización:** Diseño de tableros interactivos en Power BI que muestren indicadores clave, tendencias y comparaciones entre grupos de pacientes. Adicionalmente, en este componente también se tienen en cuenta la creación de formularios interactivos y posible actualización de flujos usando Power Automate.

3. Responsible

En el contexto colombiano, el artículo 15 de la Constitución Política reconoce el derecho a la intimidad personal y familiar, al buen nombre y a conocer, actualizar o rectificar la información registrada en bases de datos públicas o privadas (República de Colombia, 1991). En cumplimiento de este principio, la base de datos suministrada por la Fundación Santa Fe de Bogotá para el desarrollo del proyecto se encuentra completamente anonimizada, garantizando la protección de los derechos de los pacientes. Además, su uso está amparado por un acuerdo de confidencialidad que asegura el manejo ético y responsable de la información.

De igual forma, la Circular 002 de 2024 de la Superintendencia de Industria y Comercio (SIC, 2024) establece los *lineamientos sobre el tratamiento de datos personales en sistemas de inteligencia artificial*, fundamentados en cuatro criterios esenciales: la idoneidad, entendida como la capacidad del tratamiento para cumplir con el propósito propuesto; la necesidad, que implica la inexistencia de alternativas menos invasivas y con igual eficacia; la razonabilidad, orientada al cumplimiento de fines constitucionales; y la proporcionalidad, que exige que los beneficios obtenidos no superen los posibles riesgos o afectaciones al derecho al *Habeas Data*.

En el ámbito de la salud, este proyecto plantea además un reto ético significativo, dado su potencial impacto en la toma de decisiones clínicas que pueden influir en la recuperación o supervivencia de los pacientes. Por ello, es esencial reflexionar sobre qué métricas priorizar en los algoritmos desarrollados: optar por un modelo altamente preciso, pero poco interpretable, o por uno más transparente, aunque con menor desempeño predictivo. Este equilibrio entre precisión y explicabilidad constituye un aspecto central de la responsabilidad ética en el uso de inteligencia artificial aplicada a la salud.

4. Enfoque analítico

Hipótesis: Se plantea que mayor tiempo quirúrgico y tiempos de isquemia prolongados se asocian con mayor riesgo de infección postoperatoria; que la infección activa al momento del trasplante y el uso terapéutico de antibióticos pre-transplante predicen eventos infecciosos tempranos; y que características del receptor (edad, IMC, comorbilidades) junto con el tipo de reconstrucción biliar modulan el riesgo de complicaciones e infecciones en los primeros 6 meses.

Técnicas estadísticas: Con las variables de caracterización de la población se hará distinción entre las variables categóricas y numéricas. Para las variables categóricas: se determinará la distribución por frecuencias y porcentajes, mientras que para las variables numéricas: se obtendrán medidas de tendencia central (media, mediana, moda) y frecuencias relativas. Por otro lado, se realizará un análisis bivariado en el que se evaluarán asociaciones mediante la prueba Chi cuadrado de Pearson o test exacto de Fisher. Con respecto a la comparación de variables cuantitativas: se aplicará la prueba U-Mann-Whitney o la prueba T de Student, de acuerdo con la normalidad evaluada por Shapiro-Wilk (muestra < 50) o Kolmogorov-Smirnov (muestra > 50). Es importante notar que el nivel significancia, para las pruebas de hipótesis en los pasos descritos previamente, será establecido en $p < 0.05$.

Machine learning: En primer lugar, se propone un modelo para predecir el riesgo de infección postrasplante (clasificación binaria: sí/no). Se experimentará inicialmente con regresión logística, por su interpretabilidad y utilidad como línea base para estimar probabilidades y obtener razones de momios (odds ratio) de factores de riesgo clínicos, y con Random Forest, por su capacidad para modelar relaciones no lineales y su robustez frente a outliers, aportando importancia de variables más allá de la correlación lineal. En segundo lugar, para identificar perfiles de pacientes se aplicará reducción de dimensionalidad con PCA sobre variables numéricas clave y, posteriormente, agrupamiento no supervisado (k-means o clustering jerárquico) usando las variables originales o los componentes principales, con el fin de segmentar la cohorte en grupos clínicamente homogéneos que orienten intervenciones diferenciales.

5. Recolección de datos

Los datos fueron recolectados por el personal médico de la Fundación Santa Fe entre 2009 y 2025 e incluyen 60 variables para 557 pacientes. Con el apoyo del equipo clínico y a partir de la información inicial,

se clasificaron las variables en cuatro grupos según su relación (ver tabla) y se elaboró un diccionario de datos que registra el nombre y tipo de cada variable (numérica o categórica) e incorpora, cuando aplica, su codificación y descripción. Para asegurar un entendimiento preciso, se sostuvieron reuniones de validación con la parte experta, dado que se identificaron inconsistencias respecto del diccionario original entregado. Como resultado, se consolidó el diccionario de datos que se puede consultar aquí.

G1: Caracterización preoperatoria y comorbilidades (baseline)	Edad, Sexo, Etiología_#1, Etiología_#2, Child_Pugh_Score, Meld_Score, Diabetes_Mellitus, Tabaquismo, Alcoholismo, Hipertensión_Arterial, Peso_previo_Cx, Talla, IMC, Antecedente_UCI_6_meses_PreOP, UCI_al_Momento_Tx.
G2: Intraoperatorio, profilaxis antibiótica y metadatos de registro	Año_Tx; Fecha_Tx; Tiempo_Cx; Tiempo_Isquemia_Fria; Tiempo_Isquemia_caliente; Tipo_Reconstrucción_Biliar; Antibiótico_Profiláctico_VS_Terapéutico; Antibiotico_previo_al_Tx; Días_Tratamiento_Antibiótico_Previo_A_Tx#1; Tiempo_De_Dosis_Hasta_Tx#1; Antibiótico_1__Tx#1; Antibiótico_2_Tx#1; Antibiótico_3_Tx#1; Antifúngico_Tx#1; Infección_Al_Momento_Tx
G3: Infección y curso postoperatorio temprano	Localización de la infección; Germen/Microorganismo aislado; Nutrición_Enteral; Días_Nutrición_Enteral; Requerimiento_de_diálisis; Trasfusión_GRE_hasta_1m_POP; Reintervención_Quirúrgica_hasta_1m_POP; Complicaciones_Técnicas; Días_En_UCI_POP; Días_En_Hospitalización_Piso; Días_Totales_Intrahospitalarios; Fecha_Egreso_UCI; Fecha_Egreso_Hospitalario; Retrasplante.
G4: Estancias globales, inmunosupresión y desenlaces:	Inmunosupresor_1_Postx; Inmunosupresor_2_PostTx; Inmunosupresión_con_Anticuerpos; Inmunosupresor_1_1mesPostTx; Inmunosupresor_2_PostTx_1mesPx; Inmunosupresor_1_6mesesPostx; Inmunosupresor_2_6mesesPostTx; Vivo_Hoy; Fecha_Control/Muerte; SOBREVIDA_DIAS; SOBREVIDA_MESES; SOBREVIDA_AÑOS; Dias_Estancia_Hospitalaria; Días_Hospitalización_UCI

6. Entendimiento de los datos

En el notebook del repositorio se encuentra el análisis y limpieza de cada variable del dataset. Para esto, las variables se dividieron en 4 grupos según su significado, y de esta manera, se hizo el tratamiento de valores atípicos, no consistentes con el diccionario y demás errores que se encontraron. Las celdas con valores nulos o no validos representaban un alto porcentaje de los datos, estos se mapearon a -1. En todas las variables, se realizaron gráficas para entender el comportamiento de estas, dependiendo si eran categóricas, se hizo un diagrama de barras, o si eran numéricas, un boxplot e histograma. Con esto, se pudo realizar una mejor exploración y visualización de los datos.

El 51% de los pacientes son hombres mientras que el 49% son mujeres. El 35% reporta alcoholismo, 30,8% tabaquismo, el 29,4% diabetes, 19,3% hipertensión arterial. La distribución de la edad se caracteriza por tener una media en 55 años, pero una oblicuidad con tendencia a valores mayores, por lo que los datos se agrupan en edades más altas, lo que es consistente con lo que se esperaba de estos. La media de la estatura fue de 160 cm y del peso previo al trasplante de 68 kg. El 17,2% de los pacientes presentó infección al momento del trasplante, mientras que el 82,8% no. Hay registro de 27 enfermedades (etiología) distintas que conllevaron a la enfermedad hepática siendo la más prevalente NASH con un 25,7% de los pacientes.

7. Conclusiones iniciales

Se pudo evidenciar que la manera en la que los doctores de la fundación Santa Fé registran la información es poco eficiente, dado que comparten un único archivo Excel, el cual llenan todos a la vez, sin ningún tipo de validación de los datos, ni uniformidad; lo que genera que, en repetidas ocasiones, no se guarde la información correcta. Este es un verdadero problema al momento de hacer cualquier tipo de análisis o construir un producto, dado que los datos no son consistentes con la realidad. Como consecuencia, es indispensable emplear un método para mejorar la calidad de estos.

Además, se identificó que, pese a la riqueza clínica del conjunto de datos, existen vacíos e inconsistencias que dificultan el aprovechamiento de la información para análisis avanzados. La ausencia de estandarización en variables clave los tipos de infecciones, los días calculados de los pacientes en ciertos estados clínicos y los estados postoperatorios limita la trazabilidad del paciente.

8. Bibliografía

Cleveland Clinic. (2025, agosto 14). MELD score: Calculating & interpreting results. https://my.clevelandclinic.org/health/diagnostics/meld-score

Facultad de Medicina, Universidad Francisco Marroquín. (s. f.). *Escala de Child-Pugh*. https://medicina.ufm.edu/eponimo/escala-de-child-pugh/

Christian van Delden, Susanne Stampf, Hans H Hirsch, Oriol Manuel, Pascal Meylan, Alexia Cusini, Cédric Hirzel, Nina Khanna, Maja Weisser, Christian Garzoni, Katja Boggian, Christoph Berger, David Nadal, Michael Koller, Ramon Saccilotto, Nicolas J Mueller, Swiss Transplant Cohort Study , Burden and Timeline of Infectious Diseases in the First Year After Solid Organ Transplantation in the Swiss Transplant Cohort Study, Clinical Infectious Diseases, Volume 71, Issue 7, 1 October 2020, Pages e159–e169, https://doi.org/10.1093/cid/ciz1113

Fundación Santa Fe de Bogotá. (s. f.). *Sobre nosotros*. Recuperado el 17 de octubre de 2025, de https://fundacionsantafedebogota.org/intelecto/nosotros/quienes-somos

Tezcan H, Altunsoy A, Turan Gökçe D, Gökcan H, Arı D, Aydın O, Bostancı EB, Akdoğan Kayhan M. Multidrug-Resistant Infections After Liver Transplantation, Etiology and Risk Factors: A Single-Center Experience. Exp Clin Transplant. 2023 Dec;21(12):952-960. doi: 10.6002/ect.2023.0081. PMID: 38263782.

República de Colombia. (1991). *Constitución Política de la República de Colombia*. Recuperada de https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=4125

Superintendencia de Industria y Comercio. (2024, 21 de agosto). Circular Externa No. 002: Lineamientos sobre el tratamiento de datos personales en sistemas de inteligencia artificial.