Proyecto Ciencia de Datos Aplicada

Informe Entrega Final

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.

2. Preparación de datos

Para el proceso de análisis y depuración, las variables se agruparon en cuatro bloques según su significado clínico y operativo, lo que permitió aplicar criterios diferenciados para la detección y corrección de valores atípicos, registros no consistentes con el diccionario de datos y otros errores identificados durante la revisión. Dado que las celdas con valores nulos o no válidos representaban un porcentaje considerable de la base, se decidió recodificarlas de manera sistemática en un valor numérico reservado como categoría sentinela de “dato ausente o no válido”. En todas las variables se generaron visualizaciones exploratorias: en el caso de las variables categóricas se utilizaron diagramas de barras y, para las variables numéricas, se elaboraron boxplots e histogramas. Este ejercicio permitió comprender mejor la distribución y el comportamiento de los datos, así como fundamentar las decisiones de limpieza y transformación.

Durante la preparación de los datos para el entrenamiento de los modelos se llevaron a cabo transformaciones adicionales y la creación de nuevas características. La principal correspondió a la definición de la variable objetivo, entendida como un indicador binario de infección posterior al trasplante hepático, basada en la presencia de episodios de infección asociados a la inmunosupresión consignados en la historia clínica. Asimismo, se revisaron y replantearon las

variables temporales a partir de la trayectoria asistencial típica de un paciente trasplantado, que incluye el ingreso hospitalario, la etapa preoperatoria, el acto quirúrgico, la estancia en cuidados intensivos, la hospitalización en piso y el egreso. Para modelar adecuadamente la duración de la estancia en cada una de estas fases se identificó la necesidad de contar con fechas clave de ingreso, trasplante y egreso, complementadas con indicadores de estancia en unidades de cuidados intensivos y en hospitalización convencional tanto antes como después del trasplante, junto con el número total de días acumulados en cada ámbito. No se consideró necesario registrar fechas específicas de ingreso y alta en cada episodio de cuidados intensivos o de piso, dado que estos pueden ser múltiples y no necesariamente lineales en el tiempo; sin embargo, la combinación de fechas clave e indicadores de estancia permite estimar de manera adecuada la duración total por fase.

En relación con las variables categóricas, se optó por utilizar esquemas de codificación que no imponen un orden artificial entre las categorías. Se eliminaron los identificadores de paciente, por no aportar información clínica relevante y por consideraciones de anonimización, así como las variables que presentaban una alta proporción de datos faltantes y que reflejan el resultado de un estudio clínico particular. Dado que la imputación de estas últimas mediante medidas de tendencia central podría introducir sesgos importantes. Para el algoritmo de agrupamiento se aplicó, además, un proceso de escalamiento de las variables numéricas, con el propósito de evitar que las diferencias de magnitud entre ellas sesgaran el cálculo de distancias sobre el que se basa este tipo de métodos.

Finalmente, se llevó a cabo un análisis bivariado para explorar las asociaciones entre las distintas variables y la presencia de infección posterior al trasplante. En el caso de las variables categóricas se empleó la prueba Chi-cuadrado de Pearson, mientras que para la comparación de variables cuantitativas entre grupos se utilizó la prueba de Kolmogorov-Smirnov. En todas las pruebas de hipótesis se adoptó un nivel de significancia de 0,05. A partir de estos resultados, únicamente las variables que mostraron asociaciones estadísticamente significativas fueron seleccionadas como candidatas para los modelos subsecuentes, de manera que la construcción de estos se sustentara en evidencia estadística y se mitigara el riesgo de sobreajuste o inclusión de predictores irrelevantes.

3. Clustering

Para identificar perfiles de pacientes se empleará una estrategia en dos etapas. Primero, se aplicará una reducción de dimensionalidad mediante análisis de componentes principales sobre las variables numéricas clave, con el objetivo de concentrar la mayor parte de la variabilidad en un número reducido de combinaciones lineales y disminuir la redundancia entre variables altamente correlacionadas. Esta proyección a un espacio de menor dimensión facilita la visualización de la estructura de los datos y reduce el riesgo de que el ruido o las diferencias de

escala distorsionen las distancias entre pacientes. Posteriormente, se aplicarán métodos de agrupamiento no supervisado, de forma comparativa, las variables originales estandarizadas y los componentes principales. El propósito es segmentar la cohorte en grupos clínicamente homogéneos que puedan servir como base para definir estrategias de seguimiento o intervenciones diferenciales, evitando fijar a priori el número de grupos y apoyando esa decisión en criterios cuantitativos.

3.1. Estrategia de validación y selección de modelo

La estrategia de segmentación se basó en el algoritmo K-Means, un método de agrupamiento no supervisado que busca particionar n observaciones en k clústeres donde cada observación pertenece al clúster cuyo centroide le es más cercano. La selección del número óptimo de clústeres, k, se abordó de manera cuantitativa e iterativa, evitando una fijación a priori. Se evaluó la estabilidad y calidad de la partición para k en el rango [2, 10], empleando la métrica de silueta como criterio principal. Esta métrica interna evalúa qué tan bien se agrupan los objetos dentro de su propio clúster en comparación con otros clústeres. Además, para validar la estabilidad de la estructura de agrupamiento, el análisis se replicó en un subespacio de dos componentes principales (PCA). En el subespacio reducido a dos componentes principales, la métrica de silueta alcanza su valor máximo k = 5 (aproximadamente 0.71).

3.2. Construcción y evaluación del modelo

Las variables numéricas fueron sometidas a un escalamiento estándar, procedimiento esencial que normaliza la distribución de las variables a una media de cero y desviación estándar de uno, mitigando así el riesgo de que las diferencias de escala distorsionen las distancias euclidianas utilizadas por el algoritmo de K-Means. Las variables categóricas se transformaron utilizando un codificador ordinal. Esta elección, aunque simple y estable, introduce una métrica de distancia arbitraria entre las categorías, lo cual es una limitación metodológica. Una vez realizado el preprocesamiento, se transforma los datos con PCA a dos componentes.

En conjunto, los dos componentes principales explican 75.48% (PC1:55,70%, PC2: 19,78%) de la variabilidad del *dataset*. Esto confirma que el plano de visualización (PC1 vs. PC2) de la figura 1 es altamente representativo de la estructura subyacente de los datos. El primer componente principal PC1 se define casi exclusivamente por la primera etiología del trasplante. La variabilidad primaria de los perfiles de pacientes está determinada por la causa subyacente que llevó al trasplante hepático. El segundo componente principal PC2 se define principalmente por la variable que codifica el primer antibiótico utilizado en el trasplante. Esto sugiere que, después de la etiología, la estrategia antibiótica inicial es el segundo factor más importante que diferencia los perfiles de los pacientes en el conjunto de datos.

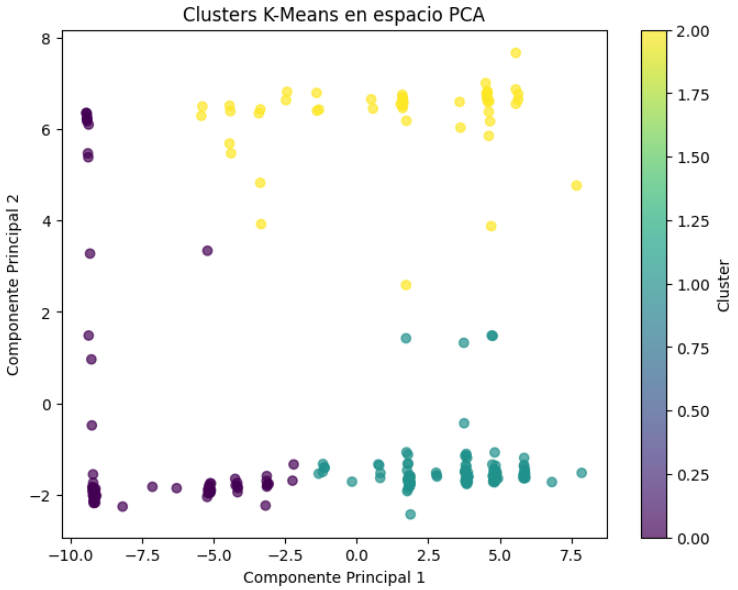

Figura 1 Clusters encontrados con K-means en espacio PCA

4. Clasificación

Acorde con nuestro objetivo principal, identificar factores asociados a riesgos de infección post trasplante hepático, se construyeron 2 modelos de clasificación binaria que permiten predecir si un paciente pudiese sufrir de infecciones post operatorias: regresión logística y random forest. Dado que previamente se había realizado la selección de variables los pasos siguientes están relacionados más hacia la búsqueda de hiperparámetros óptimos de los modelos y posterior evaluación.

4.1. Estrategia de validación y selección de modelo

Se realizó la partición de la base de datos en subconjunto de entrenamiento y validación. Adicionalmente, posteriormente se desplegó búsqueda de hiperparámetros óptimos mediante el algoritmo de GridSearch para cada uno de los modelos. Para el caso de random forest de buscó el número de estimadores, la máxima profundidad y el mínimo de muestras, mientras que para la regresión logística se varía la penalidad, la regularización, la fuerza de regularización, el balanceo de clases y número de iteraciones. Dado el contexto, nuestra métrica a priorizar para la selección del modelo es el recall, ya que requerimos que se logre identificar correctamente a las personas que ya presentan la condición infecciosa.

4.2. Construcción y evaluación del modelo

Se construyeron a matrices de confusión para evidenciar la clasificación de ambos modelos, de los cual obtuvimos lo siguiente métricas muy similares, en promedio para ambos modelos se obtuvo un recall de 0.9, en donde la clase positiva: tiene infección, se tiene un recall superior a

0.9, indicando un buen rendimiento. Ahora bien, se elegirá el modelo de regresión logístico por simplicidad de modelo y claridad en su análisis por parte del stakeholder. Se destaca de la variable de mayor impacto para la clasificación positiva fueron las siguientes: el requerimiento de diálisis, la transfusión de GRE y el tiempo de estancia hospitalaria.

5. Construcción del producto de datos.

A partir de los tres productos definidos en la fase de ideación (*Entrega 1 – Sección 2.1*), se propone la construcción de un prototipo funcional que integre: la aplicación de captura de datos en PowerApps, el modelo de *Machine Learning* para estratificación de riesgo de infección postrasplante y un tablero interactivo en Power BI para el seguimiento clínico y estratégico.
En primer lugar, se implementó la aplicación de recolección de datos en PowerApps, que actualmente funciona como el punto de entrada del sistema. La aplicación registra la información clínica y demográfica de los pacientes trasplantados mediante formularios con validaciones automáticas (fechas coherentes, variables obligatorias, rangos permitidos) y reglas de consistencia. El personal médico y administrativo autorizado accede a la aplicación según su rol, y todos los registros se almacenan directamente en la base de datos central en SharePoint Online Lists, asegurando trazabilidad y estandarización. Se hizo de esta forma para poder acceder a toda la información de manera correcta desde cualquier punto de la arquitectura en construcción.

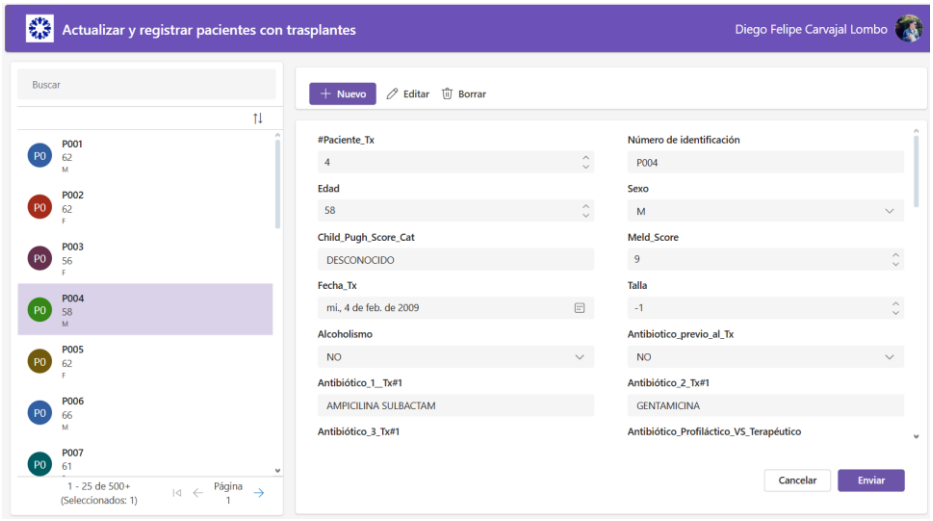

Figura 2 Aplicación de Power Apps

Adicionalmente, se desarrolló el tablero interactivo en Power BI, conectado directamente a la base de datos central y a la tabla donde se almacenan las predicciones generadas por el modelo. El dashboard presenta los KPIs definidos: tasa de infección, distribución por tipo de microorganismo, días promedio en UCI, comorbilidades, estado vital y evolución temporal. Adicionalmente, este reporte incluye información relevante al contexto como vistas individuales por paciente, KPIs por fechas a elección y análisis agregado de todos los pacientes que han tenido un transplante hepático. En esta última vista se construye una visualización haciendo

uso del estimador Kaplan-Meier para por tipo de infección, cual es el comportamiento en cuanto a muertes de los pacientes.

Este reporte incluye segmentaciones dinámicas por período, haciendo filtrado por fechas, tipo de infección y nivel de riesgo, además de la opción de exportación de los reportes en PDF o Excel. Este tablero ya se encuentra disponible para jefatura de departamento y médicos autorizados.

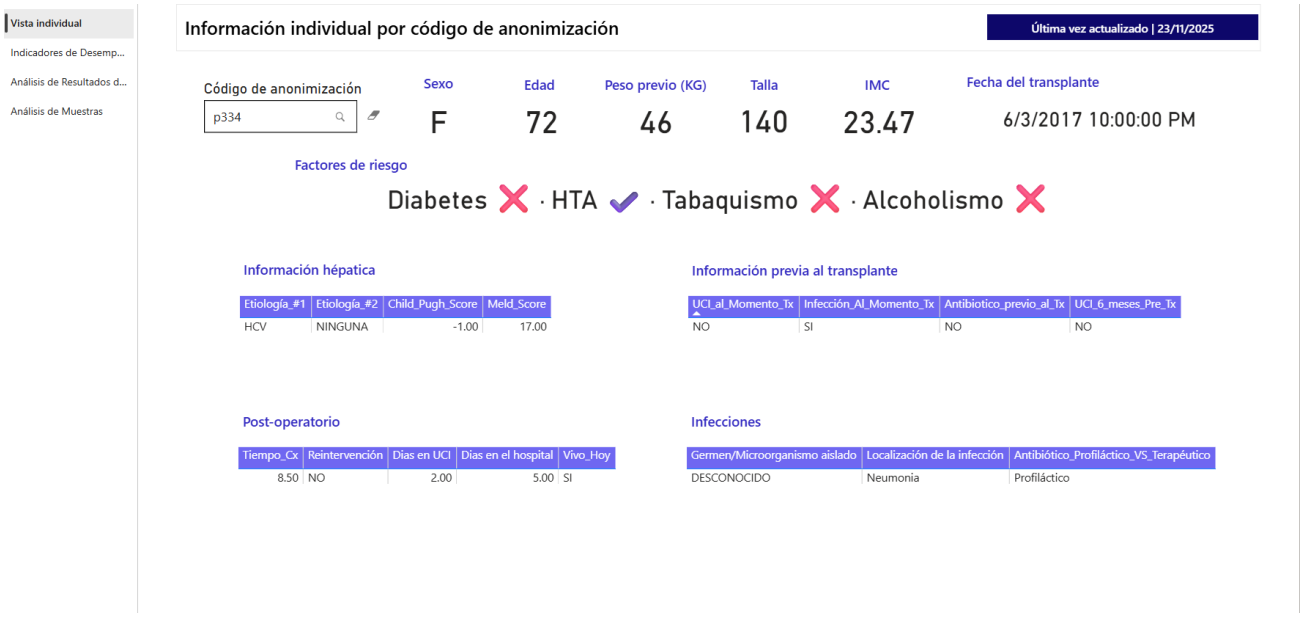

Figura 3 Reporte de PowerBI

Sobre esta base consolidada se construyó y entrenó el modelo de Machine Learning, para el despliegue, se planeó hacer mediante PowerAutomate, con el conector HTTPS. Esta herramienta ejecuta un flujo cada vez que se solicite, y dentro del él envía una petición HTTPS, recibe la respuesta para luego ser procesada. La idea inicial es mandar los datos con esta solicitud, y luego procesar la predicción del modelo. Sin embargo, este conector requiere licenciamiento premium, por lo que no se pudo hacer.

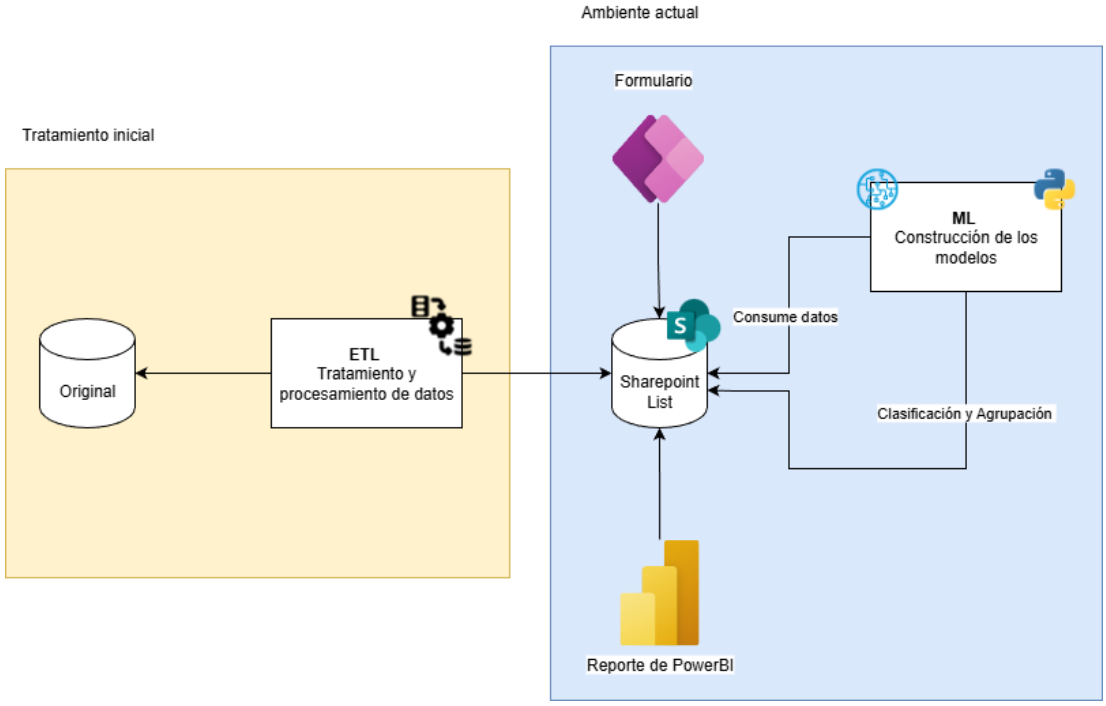

Figura 4 Arquitectura de solución

6. Retroalimentación por parte de la organización:

Durante el semestre tuvimos 4 interacciones con nuestro stakeholder: La Fundación Santa Fé. En la primera reunión ocurrida el 10 de septiembre, se nos informó acerca del proyecto y las expectivas que se tenían con este. Principalmente se buscaba analizar la data que tenían almacenada en bases de datos sobre la evolución de los pacientes a los cuales se les había realizado un trasplante hepático y sus características asociada; en específico, el objetivo era ahondar en los factores de riesgo de infección post trasplante. Posterior a esta reunión se nos fue enviado el acuerdo de confidencialidad, para luego tener acceso a la base de datos del proyecto.

En la segunda reunión que se sostuvo el 15 de octubre se realizaron preguntas sobre la consistencia e integridad de los datos, el significado de las variables médicas, los flujos de tiempo asociados a un paciente, variables relevantes para analizar y los análisis estadísticos realizados; esto nos permitió llevar a cabo procesos de limpieza de datos más efectivos y congruentes con el proyecto. Se mencionaron los productos a realizar dentro del proyecto y qué sugerencias deberíamos tener en cuenta dentro de ellos.

El tercer encuentro ocurrió el 19 de noviembre se hizo presentación de los productos principales, o de mayor interacción con los médicos tratantes, los cuales eran: el tablero de control en PowerBI, en el cual se podía evidenciar discriminación de la data dadas las recomendaciones

de la reunión previa, por ejemplo, frecuencia por bacteria encontrada, distribución de factores de riesgo como el alcoholismo o tabaquismo, fechas del año que sucedieron los eventos infecciosos, entre otras; asimismo, se presentó el boceto del formulario realizado con las PowerApps para el nuevo ingreso de información de los pacientes respetando la consistencia e integridad de los datos. En esta ocasión se nos sugirió incluir dentro del tablero una gráfica Kaplan-Meier para evidenciar la evolución de supervivencia o el tiempo hasta un evento (por ejemplo: muerte, recaída, complicación, seguimiento clínico) con distintos grupos de tratamiento.

Finalmente, el último encuentro que tuvimos con los integrantes de la Fundación fue el 24 de noviembre, en el cual se le fue presentada el tablero de Power BI, con los cambios solicitados al jefe de cirugía de trasplantes hepáticos de la fundación. En general, la retroalimentación fue positiva y se enmarcaron pasos a seguir dentro del proyecto, como la creación de variables de grupos de tratamiento a visualizar (un proceso interno de la fundación), la habilitación de una vista que permita la comparación entre dos periodos distintos, la cual logramos integrar en el tablero, y vistas que permitan visualizar la información del paciente.

7. Conclusiones

Del trabajo realizado podemos concluir que se entregaron tres productos de datos que ayudarán a la Fundación Santa Fe a mejorar el ciclo de vida de la información de los pacientes que requieren trasplante.

En primer lugar, el paso de un archivo de Excel sin controles a un esquema de gestión estructurado permitió reducir de forma notable errores de registro, valores nulos, duplicados e inconsistencias. Esto hizo que la base de datos fuera más confiable para el análisis y el modelamiento. Además, las reglas incorporadas en la aplicación para ingresar, actualizar y borrar registros evitan que reaparezcan los problemas de calidad asociados a la edición manual del Excel.

En segundo lugar, gracias a la implementación del dashboard, la Fundación cuenta ahora con una herramienta que permite analizar y tomar decisiones que pueden impactar directamente la vida de los pacientes: identificar características de los casos de fatalidad, conocer la cantidad de días promedio en UCI y explorar otras variables que contribuyen a mejorar la calidad del servicio.

De esta forma, se considera que los objetivos del proyecto se cumplieron, ya que estas herramientas fortalecen el ciclo de vida de los datos y aumentan su valor, dejando atrás la visión de un simple registro en Excel.

La principal dificultad encontrada fue la calidad de los datos y la ausencia de un diccionario de datos claro. Como se ha mencionado, la información se registraba manualmente en Excel por cada cirujano, quien definía sus propias columnas, lo que hacía extremadamente complejo cualquier esfuerzo de limpieza, análisis o construcción de modelos de machine learning.

Para mejorar las condiciones y obtener mejores resultados, es fundamental que la Fundación adopte de manera consistente la aplicación desarrollada, de modo que pueda construir una base de datos más estandarizada y limpia. Sin embargo, consideramos también que puede existir un sesgo importante debido a la limitada accesibilidad de la Fundación para pacientes de bajos recursos, lo que podría afectar los resultados.

Por último, es necesario aclarar que los modelos obtenidos se vieron fuertemente condicionados por la calidad de los datos. Una vez se utilice la aplicación y se consolide una base más estandarizada, será posible retomar el proyecto y reentrenar o construir nuevos modelos de machine learning para mejorar sus métricas.

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.