Universidad de los Andes - Maestría en Ingeniería de la Información
Profesor: Juan Pablo Reyes
Fecha entrega: 19 de octubre de 2025
Estudiantes: Juan Carlos Tovar Orjuela – Jose Jorge Geles Carrillo – Edgar Javier Toquica Gahona

PROYECTO FINAL ENTREGA 1

1. Definición de la problemática y entendimiento del negocio:

1.1. Problema y contexto: La Fundación Canguro, una organización sin ánimo de lucro especializada en salud neonatal, promueve el método Madre Canguro (MMC), que se basa en tres pilares: contacto piel con piel, lactancia materna exclusiva y seguimiento médico temprano. A lo largo de más de dos décadas, se ha recopilado una fuente de datos longitudinal que comprende más de 70 000 historias clínicas y 600 factores que afectan a aspectos clínicos, gestacionales, neonatales, familiares y de desarrollo.

El desafío principal es convertir esta información en conocimiento predictivo que pueda detectar con anticipación a los niños que corren el riesgo de sufrir alteraciones neurológicas (INFANIB anormal) o cognitivas (IQ < 85) al cumplir un año de vida corregido.

1.2 Objetivos

- *En general:* Crear un modelo de ciencia de datos capaz de pronosticar resultados cognitivos y neurológicos a partir de variables iniciales, así como determinar el impacto de los tres elementos del MMC.
- *Específicos:* Depurar y examinar la base de datos histórica (más de 600 variables).
- Identificar factores de riesgo a nivel clínico, gestacional y familiar.
- Examinar la correlación entre la intensidad de exposición al MMC y los resultados y los desenlaces de desarrollo IQ INFANIB.

1.3 Métricas clave

Indicador	**Meta**
Sensibilidad y especificidad	≥ 80%
Disminución del tiempo requerido para detectar riesgos	≥ 50%
Aumento del seguimiento intensivo	≥ 50%
Validación clínica en unidades MMC, con un mínimo de	≥ 2 pilotos

1.4 Importancia: El proyecto transformará décadas de datos clínicos en una herramienta predictiva para la salud pública, lo que permitirá reforzar la evidencia

científica del MMC y consolidar a la Fundación como un referente global en el análisis aplicado al bienestar neonatal.

1.5 Producto terminado: Un tablero interactivo que reúne registros clínicos, muestra indicadores clave de rendimiento (KPIs), emite alertas sobre riesgos y posibilita la exploración de cohortes a lo largo de la historia.

2. **Ideación**:

2.1 Potenciales usuarios:

Usuario	**Rol e interés principal**	**Problemática actual**
Médicos/Enfermeras (PMC)	Identificar riesgos tempranos y aplicar intervenciones.	La identificación de riesgos es tardía e imprecisa con métodos manuales.
Gerentes/Auditores (EPS)	Garantizar la calidad de la atención y controlar costos a largo plazo.	Falta de evidencia predictiva que justifique la inversión en el programa y que muestre su impacto a 12 meses.
Investigadores/Políticos	Generar evidencia y mejorar los protocolos de cuidado.	Dificultad para analizar el impacto de los protocolos a lo largo del tiempo (por cohortes).

2.2 Requerimientos del producto

Categoría	**¿Qué Necesita Hacer el Monitor?**
Predicción	Clasificar y Predecir la probabilidad de un IQ - INFANIB anormal a 12 meses,
Impacto	Permitir Comparar Cohortes (ej. 2005-2010 vs. 2021-2025) para medir la efectividad de la evolución del protocolo. (Variables nuevas y diferentes instrumentos para realizar la medición)
Seguridad	Garantizar que los datos anonimizados estén seguros y solo sean accesibles por usuarios autorizados.

2.3 Componentes Tecnológicos y analíticos

Componente	**Descripción**
Analítico	Modelos de Machine Learning (ML): Construir modelos que aprendan de los datos históricos para identificar los patrones (IQ - INFANIB)
Tecnológico (La Estructura)	*Base de Datos Robusta:* Migrar la información de Excel a una Base de Datos estructurada para consulta rápida. *Interfaz Web Interactiva:* Desarrollar una plataforma amigable para que el médico vea gráficos en lugar de tablas.

2.4 Mockup Conceptual

Elemento del Mockup	**Utilidad Práctica para el Usuario**
Panel de Riesgo Predictivo	Usuario - Médico: Introduce los datos iniciales de un bebé y recibe una alerta temprana sobre la probabilidad de problemas a 12 meses.
Módulo de Trayectoria Longitudinal	Usuario - EPS/Gerente: Observara un gráfico de la curva de crecimiento del bebé. Si la curva se sale del carril ideal de la OMS, es una señal de que la nutrición o el cuidado deben ajustarse

Dashboard de Intervención	Para el Investigador/Gerente: Muestra el impacto directo de las acciones (ej. horas de Posición Canguro).

3. Implicaciones éticas, de privacidad, confidencialidad, transparencia y aspectos regulatorios.

La naturaleza de los datos clínicos implica que son confidenciales y altamente sensibles, protegidos por la legislación colombiana. El Artículo 15 de la Constitución garantiza la intimidad, y la Ley 1581 de 2012 obliga a tratar los datos personales y de salud con reserva y seguridad. Las Leyes 23 de 1981 y 35 de 1989 reservan la historia clínica exclusivamente al titular o autorizados. Adicionalmente, el equipo y la Fundación Canguro han firmado un acuerdo de *confidencialidad* que prohíbe la difusión de los datos proporcionados. *El dataset está anonimizado*, y se asume que la recolección se hizo bajo el consentimiento expreso de los padres/representantes, quienes deben tener garantizado su derecho a retirar los datos (Decreto 1377 de 2013). Más allá de la ley, el pronóstico del neurodesarrollo exige *transparencia algorítmica*, anonimización rigurosa y consentimiento completo para evitar sesgos y la reidentificación de pacientes.

Consideraciones éticas: La información implícita debe protegerse, por lo que los resultados se deben presentar solo como agregaciones o datos estadísticos, sin particularizar a ningún individuo. Un aspecto ético central es la no discriminación: el modelo debe ajustarse cuidadosamente para evitar sesgos predictivos basados en situaciones sociales (estrato, etnia) entre otros.
Transparencia metodológica: Se deben describir explícitamente los pasos, el algoritmo usado y la composición del dataset, indicando las variables influyentes y señalando los riesgos o limitaciones probabilísticas del modelo. Esto garantiza que otros investigadores puedan validar los resultados y permite que el personal médico tome decisiones informadas sobre las recomendaciones médicas.

4. Método y perspectiva analítica

4.1 Preguntas e hipótesis

- ¿Qué factores familiares, maternos, gestacionales o clínicos están asociados con un IQ menor a 85 al año o con INFANIB anormal?
- ¿La adherencia a los tres elementos del MMC tiene un impacto importante en los resultados nutricionales y neurológicos?

Hipótesis nulas (H_0): No hay una correlación importante entre variables y resultados.
Hipótesis alternativas (H_1): Hay correlaciones que son estadísticamente significativas.

- *Fases de análisis :Fase 1: De exploración:* Correlaciones (Pearson/Spearman), análisis descriptivo, PCA y mapas de calor. Segmentación por medio de K-Means y verificación a través del método del codo y de Silhouette.

- *Fase 2: Inferencial:* Chi cuadrado, t de Student, ANOVA, Mann-Whitney y Kruskal-Wallis son algunas de las pruebas.
- *Fase 3: Predictiva:* (1) Modelo que se puede interpretar: regresión lineal, árboles de decisión y regresión logística. (2)Regularización Ridge y entrenamiento/validación (70% de entrenamiento y 30% de validación).
- *Etapa 4: Integración:* (1) Tablero clínico para observar en tiempo real, supervisar riesgos y tomar decisiones.(2) Resultados esperados: desarrollar mapas de riesgo, determinar predictores relevantes y sugerir intervenciones clínicas con base en la evidencia (Dashboard).

5. **Recolección de datos**:

La principal fuente de datos es un conjunto de datos multidimensional con 64.801 registros y 753 variables, que abarcan información clínica, sociodemográfica y ambiental para el estudio del neurodesarrollo en bebés prematuros.

El dataset captura:

- *Indicadores médicos directos*: medidas antropométricas, puntuaciones de neurodesarrollo y cálculos estadísticos.
- *Factores contextuales:* nivel educativo de los padres, nivel socioeconómico y entorno vital.
- El rango temporal de los datos (variable “Iden_FechaParto”) es de 15 años (25/ene/2008 a 3/ene/2023).
- Se dispone de un diccionario de datos detallado (Excel) que documenta cada variable con atributos.

Las variables de desenlace del neurodesarrollo a los 12 meses de edad corregida son “infanib12m” e “IQ12cat”. El subconjunto analítico se centrará en los registros con información completa sobre estas variables desenlace para asegurar la validez.

Adicionalmente, se utilizan siete artículos médicos como fuentes cualitativas secundarias, que proporcionan la base teórica y clínica esencial.

6. **Entendimiento de los datos**:

- *Reporte de Análisis Exploratorio y Calidad de Datos (EDA)*

Este informe detalla la fase de Análisis Exploratorio de Datos (EDA) posterior a la ingesta y transformación inicial. El objetivo es validar la calidad de los datos, comprender su distribución y detectar relaciones estructurales.

A. Gestión de Valores Ausentes (“Null”): Se aplicó una reducción de dimensionalidad a las variables con Nulos > 40%, eliminando aquellas con información insuficiente para evitar la inyección de sesgo.

B. Consistencia de Tipos de Datos (Dtype): Corregimos errores de tipado, enfocándonos en las 72 columnas forzadas a numérico. Usamos *pd.to_numeric*

para garantizar que solo haya números (int o float) y eliminamos caracteres especiales.

C. tansformación de las variables de código a tipo category. Una vez convertimos los tipos de datos con *pd.to_numeric*, aparecieron nuevos nulos. El porcentaje de nulos no superaba el 5% en ninguna variable, aplicamos *imputación simple*: usamos la mediana para los números (para proteger de los outliers) y la moda para las variables categóricas.

- *Análisis Univariado y Tratamiento de Outliers*

A. Distribución de Variables Numéricas: Evaluamos la simetría y dispersión usando Medidas de Tendencia Central (Media vs. Mediana) y Gráficos de Caja (Boxplots).

B. Detección y Tratamiento de Outliers (Valores Atípicos): Identificamos una alta incidencia de outliers (289,171 valores), señal de distribuciones sesgadas o de cola pesada. Estos se estabilizaron usando el método de Capping*.(Técnica: Rango Intercuartílico (IQR) con un factor de 1.5*).Se aplicó Winsorización (Capping), reemplazando los valores que excedían los límites del IQR con dichos límites. Esta técnica preserva el tamaño muestral, reduce la varianza extrema y mejora la estabilidad de los modelos.

- *Análisis Multivariado y Relaciones Estructurales:*

A. Codificación Categórica: Las variables object y category se transformaron usando *One-Hot Encoding*, para que los algoritmos puedan procesarlas y evitar la trampa de la variable ficticia.

B. Escalado Numérico: Todas las variables numéricas se someterán a Estandarización (StandardScaler) para lograr una media de cero y una desviación estándar de uno.

7. Conclusiones iniciales:

El data set está en una condición óptima de completitud y coherencia interna, con las transformaciones de capping y tipado aplicadas, listo para el escalado final y el entrenamiento de modelos.

Es importante destacar que las variables desenlace de nuestro proyecto no se eliminaron, se imputaron los valores, *imputación simple*: usamos la mediana para los números (para proteger de los outliers) y la moda para las variables categóricas.

Los pacientes con un resultado documentado de INFANIB e IQ fueron 26.155 de los cuales el 4,16% presentaron un diagnóstico anormal en la prueba INFANIB a los 12 meses. En cuanto al IQ se documentó en el 19% un IQ por debajo de 90. Las variables desenlaces mostraron una distribución no normal y se usó una correlación de Spearman para ver si las variables están correlacionadas y se concluyó que no están relacionadas.