Universidad de los Andes - Maestría en Ingeniería de la Información
Profesor: Juan Pablo Reyes
Fecha entrega: 30 de noviembre de 2025
Estudiantes: Juan Carlos Tovar Orjuela

PROYECTO FINAL

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.

4.2 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.

8. Preparación de datos:

1. *Carga y Exploración Inicial de Datos*: Carga de KMC-70k-93-2024-Obes 19-conVel-DATA-SPSS-20250322.xlsx y la verificación de dimensiones y tipos de datos.

- **Evidencia**: Referencia a las salidas de used_kmc_df.sample(5), used_kmc_df.shape y used_kmc_df.info().

2. *Manejo de Valores Faltantes (Nulos)*:

- *Eliminación de Columnas por Alto Porcentaje de Nulos*: Explicación de la eliminación de 36 columnas con más del 40% de nulos. Se listarán ejemplos de columnas eliminadas (V313, V356, etc.).
  - **Evidencia:** Salida de nulos_df.head(20) y columnas_a_eliminar.tolist().
- *Identificación y Conversión Forzada a Numérico*: Descripción de cómo las columnas 'object' fueron forzadas a numéricas, tratando #NULL! como NaN y generando nuevos nulos. Mencionar las 72 columnas convertidas y los nuevos nulos creados.
  - ***Evidencia*:** Salida de '--- Conversión Terminada ---' y '--- Verificación de los 5 principales nulos ---' en la celda correspondiente.
- *Imputación de Nulos Restantes*: Detalle de la imputación de todos los NaN restantes (tanto originales como generados) utilizando la mediana para columnas numéricas y la moda para categóricas (aunque en el caso del notebook solo hubo columnas numéricas con nulos restantes).
  - ***Evidencia*:** Salida de '--- REPORTE COMPLETO del Estado Actual de Nulos ---' antes y 'Nulos totales restantes en todo el DataFrame: 0' después de la imputación final.

3. *Tratamiento de Outliers:*

- *Detección de Outliers (Método IQR)*: Descripción de cómo se identificaron outliers en 73 columnas numéricas. Mención de las principales columnas con outliers.
  - ***Evidencia:*** Salida de '--- Top 10 Columnas con Mayor Cantidad de Outliers ---'.
- *Reclasificación de Categorías para Outliers:* Explicación de la conversión de 9 columnas (ej. Sistemadeaseguramiento, edadmatcat) a tipo 'category' para evitar falsos positivos en la detección de outliers numéricos.
  - ***Evidencia:*** Salida de '9 columnas convertidas a tipo 'category''.
- *Aplicación de Capping (Recorte):* Detalle del método de 'capping' aplicado a las columnas numéricas reales para limitar los outliers, resultando en el reemplazo de 152,092 valores.
  - ***Evidencia:*** Salida de ' Tratamiento de Outliers (Capping) completado.' y 'Total de valores outliers reemplazados/tratados: 152092'.

4. *Ingeniería y Codificación de Variables Categóricas:*
   - *Conversión de Alta Cardinalidad 'Object' a Numérico:* Descripción de la conversión a numérico de 19 columnas de alta cardinalidad que representaban valores cuantitativos (ej., CSP_IngresoMensual, CP_PesoMadre), seguida de imputación de NaNs con la mediana.

    - ***Evidencia:*** Salida de 'Conversión a numérico de columnas clave completada. Se han generado nuevos NaNs.' y ' Imputación de los nuevos nulos (NaNs) por la mediana completada.'.
- *Codificación Ordinal (Mapeo Manual):* Explicación del mapeo manual de 4 variables ordinales (CSP_EscolaridadPadre, BMImadrecat, BMIpadrecat, edadmatcat) a representaciones numéricas y la eliminación de las columnas originales.
    - *Evidencia:* Salida de 'Ordinal Encoding (Mapeo Manual) completado.'.
- *Codificación One-Hot para Variables Nominales (Baja Cardinalidad):* Aplicación de One-Hot Encoding a 13 columnas nominales con baja cardinalidad (ej., periodosanalisis, CSP_TipoVivienda).
    - ***Evidencia****:* Salida de 'One-Hot Encoding para variables nominales completado.'.
- *Codificación One-Hot para Categóricas Restantes:* Aplicación final de One-Hot Encoding a todas las columnas categóricas restantes que aún no habían sido procesadas, incluyendo el manejo de dummy_na=True y drop_first=True.
    - ***Evidencia:*** Salida de 'One-Hot Encoding final completado con la versión de Pandas compatible.'.

5. *Creación de la Variable Objetivo:*
    - *Definición de Riesgo_Neurocognitivo_Anual:* Detalle de cómo se construyó esta variable binaria (0 o 1) a partir de CD12 (IQ < 85) e infanib12m (ANORMAL, SOSPECHOSO, RIESGO).
        - *Evidencia:* Salida de ' Variable objetivo 'Riesgo_Neurocognitivo_Anual' CORREGIDA.' y el conteo de casos de riesgo/normales.

6. *Preparación Final de Datos para el Modelo:*
    - *Separación X e Y Exclusión Inicial:* Descripción de la división en variables predictoras (X) y objetivo (Y), excluyendo columnas irrelevantes o de resultados futuros.
    - *Limpieza de X y Eliminación de Varianza Cero:* Proceso de limpieza de X (reemplazo de strings a numérico, manejo de infinitos) y la eliminación de 508 columnas con varianza casi cero (frecuencia > 99.5%).
        - ***Evidencia****:* Salida de 'Eliminando 508 columnas por tener varianza casi cero (frecuencia > 99.5%).' y '--- Listado COMPLETO de Variables Predictoras Restantes para el Modelo 1 ---'.
    - *Filtrado por Multicolinealidad (VIF):* Detalle del proceso de intentar eliminar columnas con VIF extremo para reducir la multicolinealidad, resultando en un número específico de variables finales.
        - ***Evidencia****:* Salida de '---Lista FINAL de Variables Candidatas del Modelo 1 ---'.

7. *División de Datos para Entrenamiento y Prueba:*
    - Detalle de la división de X y Y en conjuntos de entrenamiento y prueba (80/20) usando stratify= Y para mantener la proporción de la clase objetivo.

Mención de la eliminación de columnas no predictoras del conjunto de entrenamiento.
    - *Evidencia:* Salida de 'Dimensiones de X_train: (51840, 540)' y 'Proporción de Riesgo (1) en Y_train: 0.0495'.

8. *Limpieza Final de X_train y X_test:*
    - Descripción de la limpieza de strings especiales (#NULL!) y la conversión final a tipo float para asegurar que todas las características sean numéricas antes del entrenamiento. Se mencionarán las columnas no numéricas excluidas.
        - *Evidencia:* Salida de 'Columnas no numéricas excluidas: [...]' y 'Dimensiones finales de X_train (limpio): (51840, 530)'.

9. Estrategia de validación y selección de modelo:

Muestreo Estratificado (Entrenamiento/Prueba 80/20): Dividimos el dataset en 80% para Entrenamiento y 20% para Prueba. Esta división fue estratificada, asegurando que la proporción de casos de Riesgo approx 5% se mantuviera idéntica en ambos subconjuntos. Esto previno sesgos en la evaluación final.

Validación Cruzada Estratificada (k-Fold CV): Durante la fase de tuning y selección, aplicamos la Validación Cruzada directamente sobre el set de Entrenamiento (o el set balanceado). Esto nos permitió evaluar la estabilidad de los modelos (XGBoost, RF, RL) en diferentes submuestras y hacer el ajuste fino de hiperparámetros antes de tomar una decisión final.

Selección y Prueba Inmutable: El modelo ganador será seleccionado con base en el mejor Trade-Off entre las métricas F1-Score, Precision y Recall

Distribución de riesgo:

Conjunto de Datos	Número Total de Muestras	Muestras de Riesgo (Clase 1)	Proporción de Riesgo
Conjunto Original	64,801	3,363	≈5.19%
Entrenamiento (80%)	51,840	2,566	≈4.95%
Prueba (20%)	12,961	641	≈4.95%

El reporte confirma que la proporción de la clase crítica (Riesgo) se mantiene idéntica (approx 4.95% en ambos sets). Este resultado verifica que la división de datos es válida y asegura que la evaluación final del modelo sobre el conjunto de Prueba es una medida objetiva y no sesgada de su capacidad de generalización.

10. Construcción y evaluación del modelo:

Algoritmos seleccionados:

Algoritmo	Razón de la Elección
XGBoost (eXtreme Gradient Boosting)	Lidera en la precisión por su capacidad para manejar datos complejos y desbalanceados.
Random Forest (RF)	Excelente para manejar grandes conjuntos de características y robusto contra el overfitting.
Regresión Logística (RL)	Ofrece un modelo base de alta interpretabilidad y velocidad.

Estrategia de Hiperparámetros y Muestreo

- Dada la naturaleza crítica del desbalance 19:1, la optimización no solo se centró en los parámetros internos de los algoritmos, sino también en la manipulación del conjunto de datos para presentar un patrón de entrenamiento más claro.
- Ajuste de Pesos: Uso de scale_pos_weight (XGBoost) o class_weight='balanced' (RF y RL) para penalizar los errores en la clase minoritaria (Riesgo).
- Estrategia Híbrida de Muestreo: Se aplicó una técnica combinada de Undersampling (para reducir la clase Normal) y Oversampling (SMOTE) (para generar ejemplos sintéticos de Riesgo), creando un conjunto de entrenamiento balanceado (1:1).
- Ajuste Fino de Hiperparámetros: Se exploraron parámetros clave para optimizar la estructura del modelo, incluyendo: la profundidad de los árboles (max_depth), la tasa de aprendizaje (learning_rate), y el número de estimadores (n_estimators).
- Optimización del Umbral: Finalmente, se ajustó el Umbral de Decisión (del estándar 0.50 al óptimo 0.56) para maximizar directamente el F1-Score en el conjunto de prueba.

Evaluación Cuantitativa de Rendimiento

Modelo	Estrategia de Balanceo / Ajuste	F1-Score	Precision	Recall	AUC-ROC Score
Random Forest (Base)	Sin ajuste (class_weight='balanced')	0.23	0.14	0.67	N/A
Regresión Logística (Lasso)	class_weight='balanced'	0.23	0.13	0.78	8.453
XGBoost (Ajuste de Peso)	scale_pos_weight (Peso Real ≈19)	0.28	0.18	0.63	8.438
XGBoost (Undersampling 3:1)	Muestreo Aleatorio	0.31	0.22	0.49	8.496
XGBoost (Muestreo Híbrido)	Undersampling + SMOTE	0.31	0.23	0.51	8.481
XGBoost (Ajuste de Peso Fino)	scale_pos_weight = 10	0.32	0.23	0.49	N/A
XGBoost (Híbrido Optimizado)	Umbral de Decisión (0.56)	3.228	2.555	4.384	8.481

Evaluación Cualitativa:

Robustez del Modelo: El modelo XGBoost Híbrido Optimizado es el mejor predictor disponible, pero su robustez práctica tiene limitaciones claras:

- Rendimiento Limitado: Aunque es el máximo alcanzado, un F1-Score de 0.32 indica que el modelo solo es moderadamente funcional. No es un predictor perfecto y no debe reemplazar el juicio clínico.
- Confiabilidad Práctica Aceptable: La Precision del 25.5% significa que solo una de cada cuatro alarmas de "Riesgo" es correcta. Esto permite una intervención dirigida, pero genera un alto volumen de Falsos Positivos (niños clasificados con riesgo sin tenerlo), lo cual es costoso.

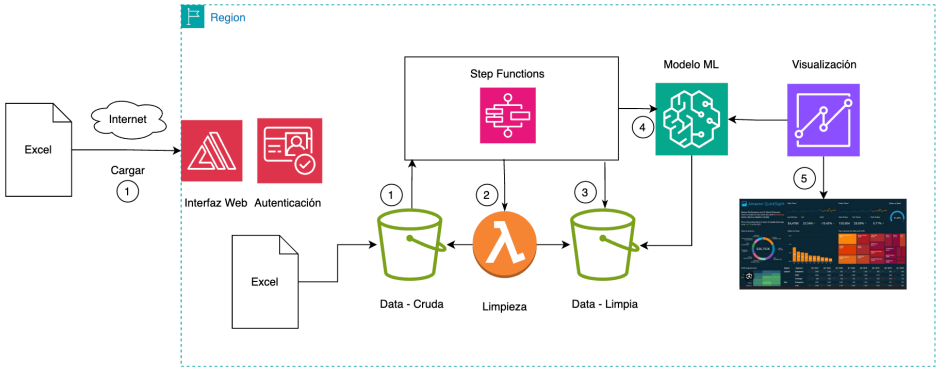
- Poder Discriminatorio Sólido AUC-ROC approx 0.85: La capacidad del modelo para distinguir entre clases es alta, lo que sugiere que la información predictiva está presente, pero es difícil de aislar debido al ruido de los datos.

Oportunidades de mejora de los modelos.

- *Eliminación o Tratamiento del Sesgo de Falla de Registro:* El alto AUC-ROC indica que el modelo puede ser mejor, pero la presencia de predictores basados en el #NULL! (falla del sistema, no variable clínica) genera un sesgo que afecta la Precisión práctica. Se debe re-evaluar la imputación de los campos críticos para que el modelo dependa de la información clínica real, no de la ausencia del dato.
- *Reducción de la Dimensionalidad (Feature Selection):* Reducir el *dataset* de 753 columnas a un subconjunto más pequeño (por ejemplo 30 a 50). Esto mitigará el ruido que dificulta el aislamiento del patrón y reducirá los costos de recopilación de datos.

11. Construcción del producto de datos:

Arquitectura propuesta AWS



Descripcion de la arquitectura:

Los numeros dentro de la arquitectura indican el flujo de los datos.

- (1) Se carga el excel al S3 (Data – Cruda)
- (2) Step Functions recibe una alerta del S3 y envía la información a la lambda(Limpieza), este se encarga de ejecutar un proceso para limpiar la información/transformar.
- (3) Step Functions recibe una notificación que el proceso del paso (2) ya termino y pasa la información limpia al S3 (Data-Limpia)
- (4) Step Functions se encarga de activar el modelo XGBoost contenido dentro de SageMaker
- QuickSight espera la señal de finalización del modelo para presentar los resultados en el tablero.

Despliegue del modelo y creación de la infraestructura.

1. Para la creación de la infraestructura se proporciona el archivo *AWSTemplateFormatVersion.yaml*, vaya a la consola de AWS y busque el servicio AWS SAM CLI, pegue el contenido del archivo y ejecute.
2. El modelo se encuentra dentro del paquete *deployment_package.zip*, descomprímalo y dentro encontrara los archivos necesarios para construir el modelo: (*inference_script.py, modelo_xgb.joblib, requirements.txt, variables_candidatas.json*)
3. Busque en la consola de AWS el servicio SageMaker, después de una configuración básica, adjunte los archivos del punto 2. Ya el modelo esta configurado.
   3.1 Conecte el SageMaker a StepFuctions (paso 1)

4. En la consola de AWS busque el servicio QuickSight, servicio de visualización e interacción con el modelo. Asegúrese de conectarlo con el paso 3.
5. Una vez tenga la infraestructura completa asegúrese de cargar un archivo al Bucket s3 (Data – Cruda) para hacer la prueba completa.

Nota: Todos los archivos estan contenidos dentro de la carpeta: ***EntregaCD***

12. Retroalimentación por parte de la organización:

Se adjuntan las tres bitácoras como resumen de la interaccion con el docente Jose Tiberio Hernández.

Acta N° 001 - 10/03/2025: Definición de la Problemática	
Categoría	**Detalle de Acuerdos y Entregables**
Objetivo de la Sesión	Entendimiento profundo del negocio y definición de la problemática central: Detección temprana y predicción del riesgo neurocognitivo en la población infantil.
Producto de Datos	NA
Enfoque Analítico	Se puso en evidencia el carácter longitudinal de los datos, y sus características de calidad
Compromiso / Tarea Pendient	El stakeholder se compromete a entregar el Dataset completo y el Diccionario de Datos oficial al día siguiente para iniciar la fase de exploración.
Aprobado por:	Jose Tiberio Hernández
Firma/Sello	
Estudiante:	Juan Carlos Tovar Orjuela

Acta N° 002 - 20/10/2025: Diseño y Estrategia Analítica	
Categoría	**Detalle de Acuerdos y Entregables**
Objetivo de la Sesión	Presentación de la visión del proyecto (Análisis Exploratorio Inicial y Arquitectura) y obtención de retroalimentación crítica.
Producto de Datos	Documento consolidado con: Análisis Descriptivo, Objetivos Detallados, KPIs de Éxito y Reporte de Limpieza de Datos.
Enfoque Analítico	Se presenta y se acuerda la necesidad de implementar una estrategia híbrida de balanceo de clases (Undersampling + SMOTE) para abordar el desbalance. Sin entrar en detalle de la estrategia de análisis
Compromiso / Tarea Pendient	Se solicita y se espera la retroalimentación formal del equipo y la aprobación del Diagrama de Arquitectura de Aplicación para proceder con el entrenamiento de modelos.
Aprobado por:	Jose Tiberio Hernández
Firma/Sello	
Estudiante:	Juan Carlos Tovar Orjuela

Acta N° 003 - 19-22/11/2025: Resultados y Conclusiones Finales	
Categoría	**Detalle de Acuerdos y Entregables**
Objetivo de la Sesión	Entrega de los resultados finales de modelado y discusión de la viabilidad de implementación del producto de datos.El objetivo inicial y el modelo se plantean a 12 meses como se evidencia en la primera entrega, sin embargo, el Stakeholder solicita que el modelo se haga con 4 diferentes modelos y cortes de tiempo, especificados así (1) Día primer examén hasta el nacimiento. 2) Valoración (40 semanas) + 1 3) Valoración (3 a 6 meses) + 2 4) Valoración (6 a 9 meses) + 3)
Producto de Datos	Documento de resultados que incluye: el Proceso de Entrenamiento y Validación de Modelos, Comparativa de Algoritmos (XGBoost vs. RF vs. RL), y las Conclusiones definitivas.
Enfoque Analítico	Se presenta el modelo ganador: XGBoost Híbrido Optimizado (F1-Score=0.3228), destacando la Precision de 25.5% como la mejor confiabilidad obtenida.
Oportunidad de Mejora	Se identifica que la principal barrera de rendimiento son los datos nulos (#NULL!). Se recomienda Inversión en la calidad del registro como paso crítico para alcanzar un F1-Score funcional (superior a 0.50). Se pone en evidencia la importancia de tener en cuenta el modelo de fases de los datos para considerarlos fase por fase (no todas las variables a la vez).
Aprobado por:	Jose Tiberio Hernández
Firma/Sello	
Estudiante:	Juan Carlos Tovar Orjuela

13. Conclusiones: Resumen ejecutivo con los resultados más relevantes del proyecto.

El proyecto cumplió su objetivo general de crear un modelo de ciencia de datos capaz de pronosticar el riesgo neurocognitivo y determinar los factores influyentes a partir de los datos históricos de la Fundación Canguro. El modelo final, el XGBoost Híbrido Optimizado, alcanzó un F1-Score máximo de 0.3228 con una Precisión del 25.5%. Si bien este resultado valida la existencia de patrones predictivos, NO cumple la meta cuantitativa de Sensibilidad/Especificidad >= 80%.

La principal barrera fue la calidad de los datos: a pesar de que se realizó imputación de datos el modelo se vio obligado a utilizar la ausencia de registro (#NULL!) y como su principal predictor de riesgo, revelando fallas en el sistema de captura de datos (Ictericia, Sifilis, Nutrición) más que factores clínicos. Por lo tanto, el modelo es un excelente prototipo de apoyo a la decisión, sin embargo, no es suficiente para ser la solución definitiva, ya que su baja Precisión generará un uso subóptimo de los recursos de seguimiento intensivo.

El impacto inmediato del producto de datos es positivo: una vez desplegado, se estima que la Detección de Riesgo (Recall) aumentará al 43.8%, impactando favorablemente en la *disminución del tiempo requerido para detectar riesgos* y el *aumento del seguimiento intensivo*, sin embargo en este momento tampoco se cumplen los objetivos planteados en un inicio. Para obtener mejores resultados y un modelo verdaderamente funcional (con un F1-Score superior a 0.50), es indispensable una mejora en la estratégica de calidad y completitud de los datos.

Se requiere eliminar el sesgo del #NULL! mediante la *re-colección* de datos para las variables críticas y realizar una ingeniería de características avanzada que cree *scores* sintéticos a partir de los factores maternos y sociales (IMC, Nutrición), convirtiendo las variables iniciales en conocimiento de alto valor predictivo, ya que el modelo aprendió cuales son las variables más relevantes es decir, Si el modelo XGBoost determinó que esta característica binaria tiene una importancia del 6.3% (HD_C_Ictericia_#NULL!), significa que el XGBoost aprendió que, si esa columna binaria es 1 (y si el dato original faltó), la probabilidad de riesgo es alta.

El dataset tiene un desbalance de 19 a 1, por esta razón se trató haciendo un undersampling para bajarlo a una relación 3 a 1 y crear un SMOTE para incrementar los datos sintéticos y lograr una relación 3 a 3. Se pudieron observar mejoras pero se recomienda trabajar en disminuir la cantidad de variables que hay por un solo niño, 753 variables por registro generan una alta probabilidad de redundancia (varias columnas miden el mismo concepto) y ruido. Esto confunde al modelo y diluye su poder de predicción.

Como conclusión el mejor modelo obtenido después de pasar por 3 diferentes modelos (Random Forest, Regresión logística y XGBoost) y finalmente profundizar con XGBoost con 3 diferentes versiones más de solo este modelo, se ha llegado a la conclusión que no es suficiente para dar solución el problema.