Instituto Tecnológico y de Estudios Superiores de Monterrey

Inteligencia Artificial Avanzada para la Ciencia de Datos

Despliegue

Los concentrados v2

Daniel Queijeiro Albo - A01710441

Diego Alfaro Pinto - A01709971

Diego Isaac Fuentes Juvera - A01705506

Jesus Ramirez Delgado - A01274723

Mauricio Anguiano Juarez - A01703337

Luis Adrián Uribe Cruz - A01783129

Índice

1. Introducción

El Campo Agropecuario Experimental del Tec de Monterrey, o CAETEC, es un rancho destinado a fines educativos cuya misión es fungir como un centro de investigación que permita poner en práctica proyectos con la industria, vinculaciones con universidades extranjeras y planes de innovación educativa. Una granja de información donde se busca cambiar de manera positiva a la agricultura y ganadería.
Uno de sus mayores fuertes, y sujeto de este proyecto, es la presencia del sistema de ordeño de DeLaval. Contando con 3 de los 5 robots presentes en el país, poseen un sistema de ordeño automatizado donde las vacas deciden voluntariamente en qué momento ser ordeñadas mientras el robot se encarga de limpiar, extraer y registrar. Esto les permite posicionarse como un referente para empresas que buscan optimizar los procesos productivos relacionados al ganado vacuno, entre ellos, las emisiones de gas metano.
Con una sólida producción diaria de 5,000 a 6,000 litros de leche, repartidos por un promedio de 35 litros por vaca y entregado a empresas lecheras de renombre como Alpura, no se encuentran exentos de problemas y áreas de oportunidad.

2. Objetivos

2.1 Objetivo de negocio

El sistema de DeLaval ya es bastante sólido, reuniendo métricas y datos tanto de la vaca como de la leche en tiempo real durante el ordeño. Por tanto, se busca poder darle uso a estos datos para generar nueva información complementaria.
De este modo, el objetivo principal del proyecto es el generar indicadores para la mejora genética y/o manejo del hato.

2.2. Criterios de éxito

- Se identifican características conductuales o fisiológicas estadísticamente asociadas con:
    - Mayor producción de leche.
    - Menor incidencia de problemas de salud.
    - Mejor adaptación al sistema robotizado.
- Se califican las vacas según sus características genéticas. Debido a la subjetividad de los objetivos, su cumplimiento debe ser aprobado por lo menos por uno de los siguientes expertos:
    - Dra. Guadalupe Lopez Rendón
    - Dr. Ivo Neftali Ayala García
    - Ing. Sergio Sebastián Caballero Chávez

2.3. Objetivos de minería de datos

- **1. Identificación de patrones operativos relevantes**
  La minería de datos es exitosa si genera indicadores cuantificables de las dimensiones con un enfoque que implemente modelos de inteligencia artificial (comportamiento y sanidad) que

permitan distinguir animales de diferente desempeño.
Se considera cumplido cuando estos indicadores permiten construir modelos que clasifican correctamente **al menos 7–8 de cada 10 casos(observaciones)**, demostrando que los patrones detectados son explotables y no ruido.

- **2. Construcción e integración de indicadores que soportan decisiones**
  El proceso es exitoso si los indicadores pueden integrarse en un índice numérico (IMR) capaz de asignar decisiones prácticas (retener, supervisar o descartar) y el índice genera decisiones alineadas con la validación experta y muestra separación clara entre categorías de animales, permitiendo priorizar manejo o mejoramiento genético.

3. Datos

3.1. Descripción de los datos

Todos los datos e información utilizados en este proyecto fueron proporcionados en su totalidad por el equipo del CAETEC, recolectados directamente desde el propio sistema de DeLaval. Estos datos se ponen a disposición libre mientras sean destinados a investigación.
Los datos son archivos separados por comas que reúnen información por vaca de los datos obtenidos durante las sesiones de ordeño, incluyendo datos cuantitativos tanto de la calidad y cantidad de la leche como de valores conductuales del animal.

- idVaca: Registro detallado de una vaca en particular sobre sus sesiones de ordeño. Esto incluye horas, duración, producción y flujo, indicadores de riesgo sanitario y el destino de la leche.
- inventario_total: Registro del inventario ganadero con datos productivos, reproductivos y de alimentación de cada vaca del establo.
- patadas: Registro por animal de sesiones de ordeño enfocado a datos de conducta y errores durante el proceso. Incluye fechas, indicadores de éxito, conteos de patadas y medidores de calidad.
- reporte: Registro Registro por vaca enfocado a su nivel de actividad y estado reproductivo. Incluye pasos por puerta MDI, estado reproductivo, ciclo de producción y conteo desde fechas clave.

Todas las características relacionadas a la lecha se encuentran divididas por cada cuarto de la ubre.

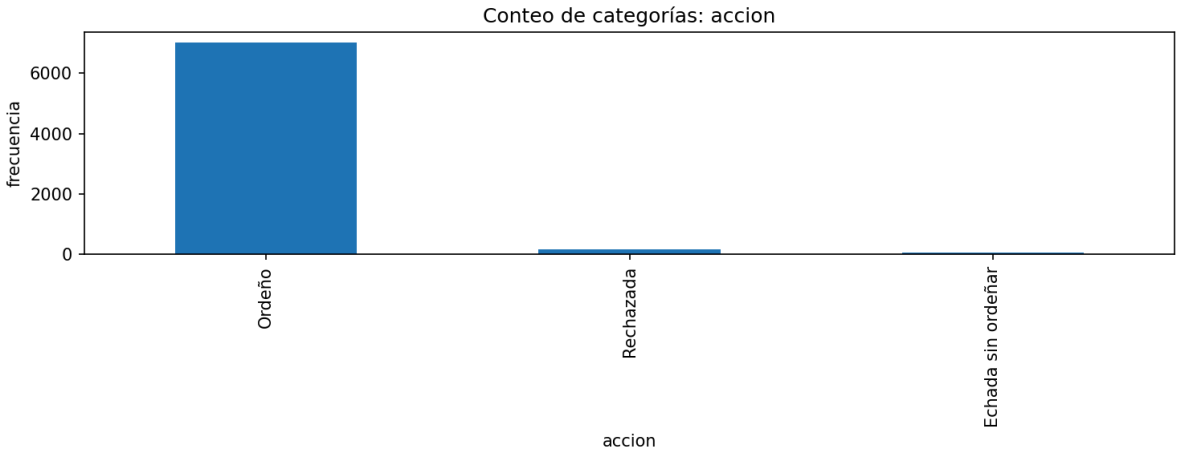
3.1.1. Problemas conocidos

Existe un principal problema con los datos provistos dividido en dos causas. Todos los archivos cuentan con con campos incompletos producto de valores nulos (faltantes) producidos por las siguientes razones:

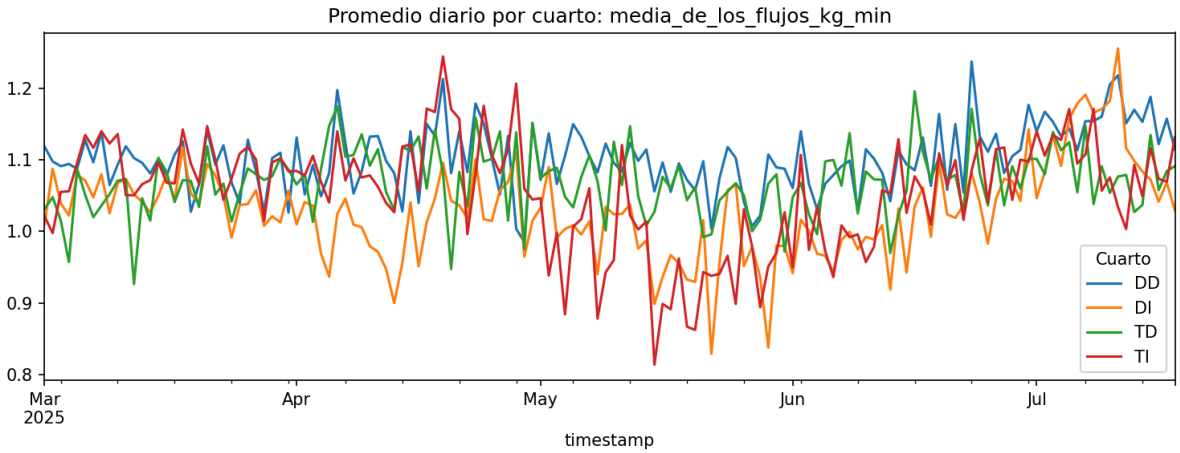
- N/A: Ante campos que no siempre poseen un valor porque “0” y “no aplica” no son intercambiables, no existe una política para llenar con un valor definido. Ante estos sucesos, el dato se deja vacío directamente.
- Error de hardware: Tras validación con operarios, se reveló que hubo fallas en los sensores del robot que provocaron su reemplazo. Esto significa que en un dado lapso de tiempo no es posible discernir de valores faltantes por estos errores, o que mediciones registradas sean incorrectas.

3.1.2. Ejemplos de distribución de los datos

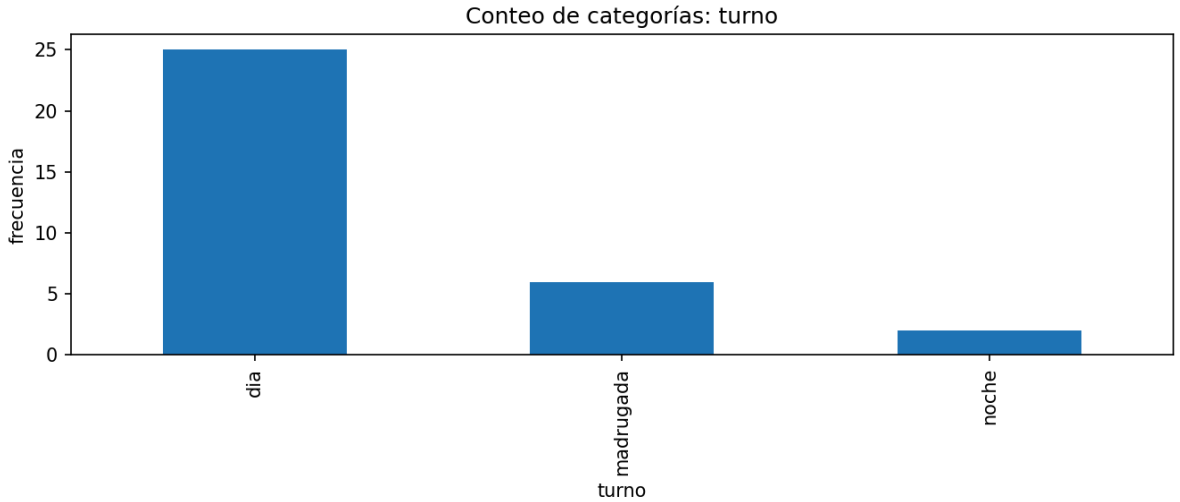
- Distribución de acción de ordeño



- Distribución de flujo promedio por cuarto



- Distribución de turno del día para ordeños



Las variables categóricas en general tienden a presentar fuerte desbalance de clases.

3.2. Preparación de los datos

3.2.1. Datos incluidos

Para las próximas derivaciones de datos se han incluido los archivos previamente definidos salvo el archivo del inventario total, pues se detectaron discrepancias respecto a los datos de las sesiones de ordeño por vaca, por lo que se consideró no sería relevante para los modelos. Del mismo modo, aquellas columnas que poseen casi en su totalidad valores faltantes, como la columnas "Misc." del archivo del registro de sesiones de ordeño pues tenía un 99% de datos vacíos.
De los archivos seleccionados se encuentra información de estado anímico, registros de salud, valores absolutos y medias respecto a la producción, tendencias de comportamiento, actividad de los animales y dificultades del robot.

3.2.2. Datos derivados

1. **Registros de Sesiones de ordeño (registros_sesiones_merged.csv)**
    a. **Origen**: 33 archivos individuales por vaca, fusionados
    b. **Registros**: 7,239 sesiones de ordeño
    c. **Columnas**: 34 atributos
    d. **Transformaciones**: Se integraron los archivos por vaca en un solo dataset y se aplicaron procesos de imputación para manejar valores vacíos, utilizando ventanas de entré más-menos 3 días por vaca junto con sus promedios locales. En el caso de sangre en la leche los valores faltantes se rellenaron con 0 al interpretarse como ausencia de detección
    e. **Justificación**: Estos archivos contienen información detallada de cada sesión de ordeño, con métricas de importancia sobre la producción, calidad de leche y comportamiento por cuarto mamario.
    f. Atributos clave incluidos
        i. ID de vaca: Identificación única del ejemplar
        ii. timestamp: Fecha y hora de la sesión
        iii. Producción por cuarto (DI, DD, TT, TD): Cantidad de leche en kg
        iv. Flujo y promedio máximo: Velocidad de extracción por cuarto
        v. Conductividad eléctrica: Característica de la leche que sirve cómo indicador de mastitis.
        vi. Sangre en leche (partículas por millón ppm): Cantidad de partículas de sangre que hay en la leche, indica posibles lesiones y/o mastitis.
        vii. Estados del ordeño: Patadas, incompleto, pezones no encontrados
        viii. Destino de leche: Tanque, drenaje, divert
2. **Registros de Salud (sessions_health.csv)**
    a. **Origen**: registros_sesiones_merged.csv
    b. **Registros**: 7,240 sesiones
    c. **Columnas**: 20 atributos
    d. **Transformaciones**: Elimina columnas innecesarias para su respectivo modelo, normaliza los nombres de columna. No realiza imputación de datos, registros con valores faltantes directamente vacían toda la fila.
    e. **Justificación**: Dado el enfoque de múltiples modelos especializados, este archivo se genera para el modelo que intenta predecir el nivel de salud de una vaca.
    f. Atributos clave incluidos

i. sange_ppm: Dividido por cada cuarto, es la cantidad de partículas por millón de sangre en la leche. Su presencia por sí sola no indica cuál es el padecimiento, pero discrimina que haya uno en sí o no.
ii. conductividad_ms_cm: Dividido por cada cuarto, es una medida en la leche que indica posible mastitis debido a la presencia de iones derivados de la enfermedad.
iii. media_de_los_flujos_kg_min: Dividido por cada cuarto, indica la velocidad a la que la leche es ordeñada. Un cambio en esta velocidad indica alteraciones en la ubre.
iv. producciones_kg: Dividido por cada cuarto, es la masa total de leche ordeñada. La salud de la vaca afecta directamente a su producción.

3. **Registros de Comportamiento (sessions_behavior.csv)**
   a. **Origen**: registros_sesiones_merged.csv
   b. **Registros**: 7,240 sesiones
   c. **Columnas**: 16 atributos
   d. **Transformaciones**: Elimina columnas innecesarias para su respectivo modelo, normaliza los nombres de columna. No realiza imputación de datos pero sí la generación de una nueva columna.
   e. **Justificación**: Dado el enfoque de múltiples modelos especializados, este archivo se genera para el modelo que intenta predecir el nivel de buen o mal comportamiento de una vaca.
   f. Atributos clave incluidos
      i. dur_seconds: La duración total del ordeño, convertido a segundos.
      ii. main_produccion_kg: Producción total obtenida en el ordeño.
      iii. estado_numero_de_ordenos: Número del ordeño registrado en un determinado lapso.
      iv. label_inquieta: Bandera binaria que indica si en el ordeño la vaca tuvo cualquier tipo de problema en el proceso.

4. **Registros de Mérito productivo (merito_productivo_vacas.csv)**
   a. **Origen**: registros_sesiones_merged.csv
   b. **Registros**: 33 datos
   c. **Columnas**: 5 atributos
   d. **Transformaciones**: Reúne 3 columnas relacionadas a la producción, crea una nueva y las ata al ID de cada vaca.
   e. **Justificación**: Dado el enfoque de múltiples modelos especializados, este archivo se genera para concentrar los méritos de la vaca en relación a su producción de leche histórica.
   f. Atributos clave incluidos
      i. produccion_ajustada_total: La producción de leche total de los datos conocidos.
      ii. n_ordenos: Total de ordeños registrados entre los datos conocidos.
      iii. produccion_media_observada: Promedio de producción por cada ordeña de la vaca.
      iv. merito_productivo: Calificación de la producción histórica de la vaca.

Los 3 archivos derivados del archivo maestro unificado reciben el mismo preprocesamiento, donde valores nulos resultantes son dejados así. El cambio más relevante es la normalización de todos los nombres de columna.

Para esta normalización, se eliminan todos los caracteres especiales de los nombres, todas las letras se vuelven minúsculas y las palabras resultantes se separan por guión bajo. Los caracteres especiales incluyen las vocales con tilde y la letra ñ, las cuales son reemplazadas por su equivalente del sistema de caracteres estándar.

3.2.3. Datos generados

- Totales: La suma de los cuartos de ubre en todas aquellas columnas que posean tal división
    - Relevancia: Como es de importante disponer de los datos por cuarto para analizar cada sección de la ubre individual, también es útil tener en cuenta el número total por vaca.
- Fecha y hora: Reunidas bajo la misma columna de Fecha, se dividen como características separadas.
    - Relevancia: Diferentes modelos requieren sólo una o la otra, por lo que es mejor tener ambas opciones para elegir la relevante a cada situación.
    - De esta se deriva también una columna de conversión de las horas y minutos a su equivalente en segundos.
- Producción suavizada (Lowess): Se extrae la tendencia general de las producciones diarias.
    - Relevancia: Existen una gran varianza entre datos de día a día y de vaca a vaca, induciendo ruido.
    - Obtención: Se calcula mediante la función Lowess del módulo Statsmodels.
- Tasa de decaimiento: Qué tan bien se mantiene la producción de leche a través del tiempo.
    - Relevancia: Al calcular los ciclos de producción máxima y mínima se puede apreciar la forma en la que crece y decrece la producción en cada ciclo.
    - Obtención: Se calcula desde la curva suavizada obtenida de Lowess.
- IDs: Identificación única de cada vaca
    - Relevancia: No es un dato que se utilice directamente en los modelos, pero es necesario para poder clasificar a las vacas.
    - Obtención: De los nombres de archivo
- Bandera de inquietud: Etiqueta binaria que indica si una vaca posee cualquier forma de mal comportamiento en una sesión de ordeño.
    - Relevancia: Como objetivo de uno de los modelos, es necesario poseer una variable objetiva que indique comportamientos alterados.
    - Obtención: Se marca presente cuando cualquier columna de error posee un valor no nulo.

3.3. Política de datos

Esta política se extiende a todos los datos recibidos, derivados y generados como parte del trabajo de este proyecto y todos los miembros del equipo ya han declarado su conformidad con el mismo. La política se basa en 4 principios fundamentales:

1. **Confidencialidad comercial**
   Los datos son propiedad de CAETEC. No se compartirán fuera del equipo autorizado sin consentimiento expreso.
2. **Transparencia y apertura académica**
   Se promoverá el desarrollo de modelos open source. Los resultados serán compartidos con la comunidad académica respetando la confidencialidad de datos sensibles.

3. **Calidad y confiabilidad**
   Los datos serán validados, limpiados y documentados antes de su uso. Se mantendrán registros de todas las transformaciones realizadas.
4. **Ética y bienestar animal**
   El uso de datos no debe comprometer el bienestar del ganado bovino. Todas las decisiones basadas en datos priorizan la salud y el bienestar animal.

Serán considerados datos confidenciales todos aquellos archivos generados por DeLaval, imágenes de los animales o cualquier otro dato entregado por CAETEC. Por otro lado, códigos, modelos y reportes escritos serán considerados como datos públicos.

3.3.1. Datos confidenciales

El acceso a esta información está restringido hacia los miembros del equipo, profesores de la materia y al personal de CAETEC. Los datos se almacenan de forma segura a través del servicio Amazon S3 de AWS, donde la gestión por parte del servicio IAM garantiza la confidencialidad y control del acceso. IAM estipula permisos mínimos de lectura y escritura sujetos a la generación de logs para cada miembro particular, donde además es necesario la autenticación de dos factores para entrar.
Los datos son de animales y procesos, los cuales no están estipulados en la protección de datos personales al no pertenecer a algún humano identificable a través de dicha información, no requieren de anonimización u otras técnicas para prevenir la inferencia de las poblaciones contenidas. Sin embargo, sí es requisito de ley dictado por SINIIGA/SENASICA que existe trazabilidad e identificación de los animales. Sumado a esto, por intereses comerciales también hay que mantener confidencialidad de datos y procesos que pueden suponer una ventaja injusta por parte de terceros que puedan tener acceso.
Con estos requisitos en cuenta, se garantiza la protección de la información y su acceso para lo estrictamente necesario con la estructura ya establecida y no violar ninguna de estas imposiciones.

3.3.2 Datos públicos

Toda la información considerada pública se encuentra alojada en repositorios abiertos a la lectura del público general mientras que la escritura mantiene las mismas regulaciones limitadas al equipo de trabajo. Cualquiera, entonces, puede ver y reproducir la información para realizar sus propias modificaciones, pero los repositorios originales del proyecto sólo pueden verse alterados por los miembros participantes.

3.4. Ciclo de vida

1. **Obtención de datos crudos**
   a. Se descargan los archivos CSV desde S3/raw/ (subidos previamente por el usuario).
2. **ETL (Limpieza, unión y validación de datos)**
   a. Se realiza el preprocesamiento de datos en **Google Colab** o **PC local (Windows)**.
   b. Se corrigen valores faltantes, se validan formatos y se integran múltiples fuentes.
   c. Los datos procesados se guardan en S3/processed/.
3. **Entrenamiento del modelo (K-Fold o equivalente)**
   a. Se entrena el modelo de forma offline usando los datos de processed/.
   b. Se genera un modelo final con pesos ajustados.
   c. Los pesos se guardan en S3/models/.
4. **Resultados del entrenamiento**

a. Se registran métricas y resultados (accuracy, loss, logs) en S3/outputs/.

3.4.1. Ciclo con el modelo ya entrenado

1. **Frontend estático (CloudFront + S3 Website)**
   a. El usuario accede a la interfaz y realiza solicitudes mediante fetch (HTTP).
2. **Ejecución de inferencias o refresco de resultados**
   a. POST /refresh: Lambda toma los últimos datos de processed/ y el modelo en models/.
   b. Ejecuta el modelo entrenado y genera nuevos resultados (outputs/top.json).
3. **Consulta de resultados**
   a. GET /top: Lambda devuelve al usuario el contenido de outputs/top.json, mostrando los resultados actuales del modelo.

3.4.2. Ciclo actualización continua

1. **Carga de nuevos CSV**
   a. El usuario sube nuevos registros vía POST /upload.
   b. Lambda los almacena en S3/raw/.

2. **Actualización del pipeline**
   a. Los nuevos datos pueden ser procesados nuevamente en la fase ETL.
   b. Se generan versiones actualizadas en processed/.
3. **Ejecución del modelo actualizado**
   a. Lambda usa los nuevos datos para refrescar predicciones sin reentrenar completamente.
   b. Los resultados actualizados se almacenan en outputs/.

4. Modelado

Como parte de los objetivos propuestos, la intención de esta sección es tener la capacidad de establecer una calificación para cualquier animal de acuerdo a su información actual e histórica para poder determinar la calidad de dicho animal. La calificación debe poder ayudar a determinar si una determinada vaca es destacada en su producción y comportamiento, si debe de mantenerse en vigilancia o si por el contrario es necesario establecer otras medidas ante un desempeño inferior.
La tarea se ha dividido en 3 secciones. En lugar de generar un único modelo que abarque todas las columnas a pesar de las diferencias entre sí y deba crecer su arquitectura por la complejidad, se optó por crear en su lugar dos modelos más simples más un cálculo directo. Así el enfoque se reparte de la siguiente manera:

Dimensión	**Tipo de modelo**	**¿Qué mide?**	**Por qué es importante**
Comportamiento	Modelo Supervisado	Eficiencia y estabilidad operativa durante el ordeño	Reduce tiempos muertos y pérdidas
Sanidad (proxy)	Modelo No Supervisado	Probabilidad de eventos de mastitis o complicaciones de ubre	Evita costos futuros y descarte de leche
Mérito productivo	Ajuste estadístico	Capacidad productiva propia de la	Determina valor

ajustado	(no modelo)	vaca sin efecto del ambiente	genético y retención

4.1. Mérito productivo

Tiene como propósito determinar los niveles de una vaca en particular tras eliminar factores ambientales que alteren los resultados. Esta se obtiene directamente a través de fórmulas ya establecidas sin recurrir a ningún tipo de aprendizaje automático y/o profundo. Para obtenerla, primero se calcula la producción de la forma:

$$ProduccionAjustada_i = ProduccionObservada_i - \bar{x}(Produccion \mid robot\ x\ hora\ x\ mes)$$

Donde:
i = Sesión de ordeño

Para entonces poder calcular el mérito productivo así:

$$MeritoProductivo_i = \frac{i}{N_I} \sum_{j=1}^{N_i} ProduccionAjustada_{i,j}$$

Donde:

i = vaca
j = sesión de ordeño de esa vaca
N_i = número de ordeños que tiene esa vaca en el período

Esto representa valor genético operativo, no desempeño circunstancial.

4.2. Comportamiento

Tiene como propósito la identificación de vacas que entorpecen o cancelan las sesiones de ordeño, lo cual genera ineficiencia en el robot y hasta potencialmente daños al mismo. El comportamiento tendrá un enfoque de aprendizaje supervisado, así que su columna objetivo, la bandera de inquietud. Esta bandera usa lógica binaria para construirse:

$$Inquieta = (Patada = 1) OR (Incompleto = 1) OR (PezonesNoRncontrados > 0) OR (Duracion > P95)$$

Como esta etiqueta es binaria, 1 o 0, la salida del modelo se puede interpretar como una probabilidad promedio de mal comportamiento para una vaca.

4.2.1 Modelos seleccionados

Versión	Algoritmo seleccionado	Justificación	Iteración asociada

LR Baseline	Regresión Logística	- Baseline sin tuning para medir mejora posterior. - Simple, interpretable y rápido. - Sirve como punto cero para comparación con RF y MLP. - Bajo recall y F1. No apto para despliegue.	1
RF v1.0	Random Forest - dataset 1.0	- Primer modelo considerado. - Manejo robusto de no-linealidad. - Buen accuracy, pero recall insuficiente. - Identifica patrones, pero aún sub-detecta inquietas.	2
RF v2.0	Random Forest - dataset 2.0	El objetivo de la versión v2.0 no fue optimizar hiper parámetros, sino corregir la base de datos de entrenamiento manteniendo la misma lógica algorítmica, permitiendo medir el impacto puro de la mejora en calidad de datos. Está versión representa el punto de inflexión entre desempeño limitado por datos y desempeño mejorado por estructura y volumen del dataset.	3
RF v2.1	Random Forest 2.1 GridSearchCV	- Ajuste sistemático de hiper parámetros. - Mejor desempeño global en todas las métricas. - Balance ideal entre precisión y F1. - Versión más madura para uso operativo	4
MLP	Red Neuronal MLP v1.0	- Excelente para detectar inquietas (mayor recall). - Aprendizaje no lineal más profundo. - Sensible → detecta más positivos incluso si hay baja precisión. - Preferible cuando la prioridad es no dejar vacas sin revisar. - Punto de referencia para comparar los resultados con los modelos de RF.	5

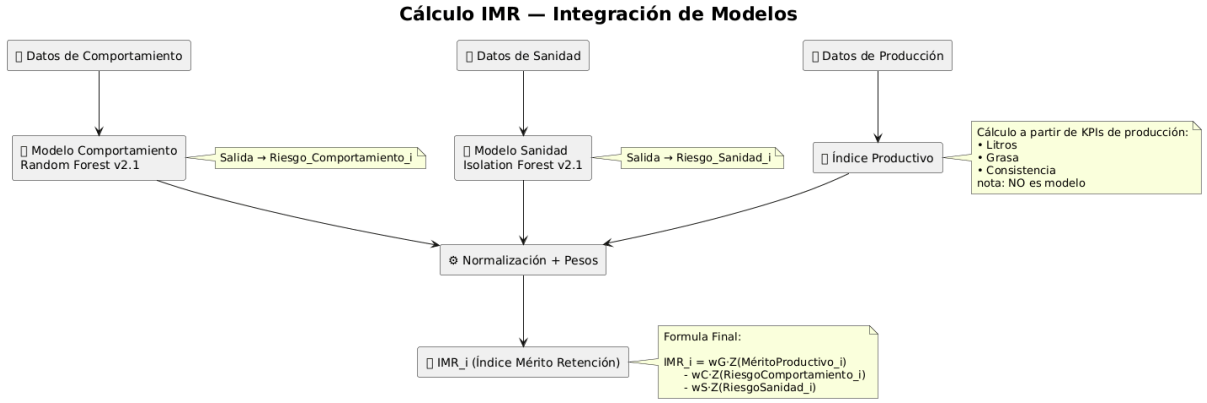
4.3. Sanidad

Tiene el propósito de determinar la susceptibilidad de una determinada vaca a padecer condiciones de salud, las cuales involucran infecciones o afectaciones en las ubres. Un animal enfermo es un animal cuya producción además de disminuir, no puede utilizarse. Un animal que es, entonces, más vulnerable a tener problemas de salud es un animal al que se le debe de invertir más tiempo en cuidados y a cambio tendrá un peor desempeño.
Este modelo requiere de un acercamiento por aprendizaje no supervisado, debido a que la columna que debería proporcionar esta información se encuentra vacía y no hay una forma de generar una etiqueta de riesgo sanitario.
Por ello, se usan modelos de agrupamiento y detección de anomalías para identificar grupos normales y poder identificar a las vacas que son anormalmente más enfermizas. La salida de este paso es el promedio de puntuación de anomalía.

4.3.1 Modelos seleccionados

Versión	Algoritmo seleccionado	Justificación	Iteración asociada
DBSCAN	Clustering no supervisado	- Detecta grupos densos y ruido como anomalías. - Sin supuestos estadísticos. Útil en sanidad real. - Alta interpretabilidad visual vía PCA. - Sirve como referencia base para métodos futuros	1
Isolation Forest v1.0	Isolation Forest - dataset 1.0	- Detecta anomalías raras sin etiquetas. - Ideal para datos tabulares de salud (numeros+binarios). - Da score continuo de riesgo sanitario. - Identifica patrones inusuales en DI/DD/TI/TD. - Limitación: mayor variación entre folds y % anomalías.	2
Isolation Forest v2.0	Isolation Forest - dataset 2.0	El objetivo de la versión v2.0 no fue optimizar hiper parámetros, sino corregir la base de datos de entrenamiento manteniendo la misma lógica algorítmica, permitiendo medir el impacto puro de la mejora en calidad de datos. Está versión representa el punto de inflexión entre desempeño limitado por datos y desempeño mejorado por estructura y volumen del dataset.	3
Isolation Forest v2.1	Isolation Forest Optimizado	- Mantiene beneficios de la versión anterior. - Usa One-Hot Encoding y preprocesamiento más limpio. - CLI configurable + autodetección de tipos. - Column Transformer + pipelines separados para num+cat. - Versión más estable y con mejor calibración (5%).	4

4.4. Integración de conceptos



Los resultados de las 3 secciones del problema se integran en el Índice de Mérito para a Retención (IMR), El IMR es una métrica propuesta con rango continuo [0,1] donde valores más cercanos a 1 indican un mejor desempeño general de la vaca. Para calcular el IMR, se utilizan ponderaciones definidas para cada término de los modelos en la próxima ecuación:

$$IMR_I = w_G \cdot Z(MeritoProductivo_i) - w_C \cdot Z(RiesgoComportamiento_i) - w_S \cdot Z(RiesgoSanidad_i)$$

Donde:

U

w_G = peso del mérito productivo (recomendado: 0.5)

w_C = peso del riesgo de comportamiento (recomendado: 0.2)

w_S = peso del riesgo sanitario (recomendado: 0.3)

Una vez obtenido el IMR, además del propio número, se permite evaluar el resultado contra la tabla de rangos propuesta para facilitar el criterio.

Rango del IMR (basado en percentiles poblacionales)	**Condición estadística**	**Decisión recomendada**	**Interpretación operativa**
IMR ≥ p75	Top 25% del hato	Retener / Reproducir	Alta prioridad genética y productiva — candidata para inseminación o mantenimiento prolongado.
p40 ≤ IMR < p75	Segmento medio (p40–p75)	Supervisar / Manejo dirigido	Vacas de comportamiento intermedio — conservar, monitorear y optimizar manejo antes de decidir reproducción o descarte.
IMR < p40	Bottom 40% del hato	Descartar / Secado / Venta	Bajo desempeño relativo — adecuada para retiro, secado o priorizar reemplazo.

5. Evaluación

Para poder evaluar el desempeño de todos los modelos de forma equitativa, todos recibieron el mismo Pipeline, todas las inicializaciones aleatorias tienen la misma semilla definida y a todos se les aplican las mismas métricas dentro de su propia área, es decir, estos campos están divididos para el modelo supervisado y el no supervisado.
Para ambos modelos, se tienen las siguientes métricas junto a sus respectivos criterios:

Modelo	Métrica	Criterio aceptable	Criterio ideal
Comportamiento	Accuracy	≥ 0.80	≥ 0.85–0.90
	Precision	≥ 0.75	≥ 0.80–0.85
	Recall	≥ 0.70	≥ 0.80
	F1 Score	≥ 0.72	≥ 0.80
Sanidad	% anomalías por fold	diff ≤ 5 puntos	diff ≤ 3 puntos
	% anomalías global	4%–7%	≈ 5%
	Histograma score	cola visible	outliers muy definidos

5.1. Resultados

5.1.1. Modelo supervisado de Comportamiento

Modelo	Accuracy	Precisión	Recall	F1 Score
Regresión Logística (baseline)	0.7746	0.6639	0.6713	0.6674
Random Forest v1.0	0.909091	0.3333	0.3333	0.3333
Random Forest v2.0	0.8629	0.8910	0.6766	0.7689
Random Forest v2.1 (Optimizado)	0.8699	0.8964	0.6943	0.7824
MLP Optimizado	0.8505	0.8068	0.7351	0.7682

Observaciones:

- El experimento comparó cuatro modelos para identificar vacas inquietas. La **Regresión Logística** sirvió como punto de partida, pero quedó por debajo de los criterios mínimos en todas las métricas, funcionando únicamente como referencia comparativa.
- La primera versión de **Random Forest (v1.0)** mostró resultados artificialmente altos en accuracy pero con métricas inestables y poco confiables debido a un dataset pequeño y desbalanceado, por lo que fue descartada y el modelo se reconstruyó con más datos.
- La versión **Random Forest v2.0** logró un desempeño sólido con alta precisión, pero aún dejó escapar casos reales (recall bajo), mostrando oportunidad de mejora.
- La versión **Random Forest v2.1**, optimizada mediante búsqueda de hiper parámetros, presentó el mejor equilibrio global entre precisión, recall y estabilidad entre folds, convirtiéndose en el modelo más maduro y recomendable del enfoque basado en árboles.
- Finalmente, el **MLP (red neuronal)** prioriza la detección de casos inquietos y fue el único que superó el criterio mínimo de recall, siendo útil cuando la prioridad es capturar el mayor número posible de positivos aunque implique sacrificar algo de precisión.

5.1.2. Modelo no supervisado de Sanidad

Modelo	Variación entre folds	% Anomalías global	Histograma (distribución)	PCA (separación espacial)
Modelo base (DBSCAN v1.0)	1.98 p.p. (mejor estabilidad)	4.97% (óptimo)	Cola clara + corte 5% definido	Clúster compacto + outliers limpios
Isolation Forest 1.0	60 p.p. (muy inestable – no cumple)	16.11% (fuera del rango – no cumple)	Cola marcada pero con extremos amplios (0–60%) — separación inconsistente	Outliers presentes pero dispersos; separación aceptable pero no homogénea
Isolation Forest 2.0	~4 p.p. (inestable vs otros)	7.22% (sobre-detección)	Frontera difusa - menos separación	Separación visual aceptable
Isolation Forest 2.1	2.42 p.p. (ideal)	5.03% (equilibrado)	Corte limpio top 5%, más definido que v2.0	Cluster normal más compacto + outliers aislados

Observaciones:

- El experimento evaluó 4 enfoques no supervisados para detectar anomalías sanitarias: DBSCAN, Isolation Forest 1.0, Isolation Forest 2.0 e Isolation Forest 2.1. El análisis buscó medir estabilidad, claridad en la separación de casos atípicos e interpretabilidad de resultados.
- **DBSCAN v1.0** sirvió como referencia sólida. Mostró la mayor estabilidad entre pliegues, detectó anomalías en el rango ideal (~5 %) y evidenció una segmentación limpia de outliers, funcionando como curva base para comparar el desempeño del resto de modelos.

- **Isolation Forest v1.0**, entrenado con un dataset pequeño, resultó inestable y sobreajustado: la tasa de anomalías fluctuó drásticamente entre folds y el promedio global (16.11 %) se alejó del objetivo. Aunque sugería cierta separación visual, sus resultados eran inconsistentes, por lo que se descartó y se reentrenó con una base ampliada.
- **Isolation Forest v2.0** ofreció mejoras, pero mostró sobre-detección (7.22 % de anomalías) y menor definición geométrica, lo que confirmó la necesidad de refinar estabilidad y calibración del umbral.
- Finalmente, **Isolation Forest v2.1** alcanzó el equilibrio deseado: proporción ideal de anomalías (~5 %), variabilidad controlada entre folds (~2.42 p.p.), histograma más limpio y mayor separación espacial de atípicos en PCA. Esto lo convierte en el modelo más maduro e interpretable, recomendado para despliegue operativo y como base para futuros métodos avanzados.

5.2. Modelo final

Las versiones **v2.1** tanto para comportamiento como para sanidad representan las primeras versiones que:

1. Cumplen los criterios aceptables o ideales en **al menos tres métricas clave**,
2. muestran **estabilidad** ante validaciones cruzadas,
3. mantienen **comportamiento consistente** visual y estadísticamente,
4. superan los problemas observados en versiones anteriores (dataset insuficiente, sobre detección, baja sensibilidad).

Por lo anterior, ambas v2.1 son seleccionadas como modelos candidatos finales para integración y despliegue.

6. Despliegue

A modo de poder utilizar la arquitectura ya establecida dentro de AWS, originalmente se pensaba entregar una aplicación web integrada directamente en ella, sin embargo, esta idea fue descartada en pos de no tener que generar costos adicionales a CAETEC. En su lugar, todos los archivos se concentraron en una única aplicación de escritorio, de forma que corre localmente y funciona con recursos propios, sin tener que adaptar una implementación en AWS ni tener que pagar por ello.
La interfaz permitirá subir los datos por vaca, facilitando la alimentación de los modelos localmente sin depender de la nube. El flujo de trabajo procesa los archivos exportados por el sistema DeLaval para ejecutar los modelos seleccionados: el modelo Random Forest v2.1 para la clasificación de comportamiento y el Isolation Forest v2.1 para la detección de anomalías de sanidad. El objetivo de este plan es que la aplicación se encargue de la visualización de los resultados de los datos subidos, integrando las dimensiones de comportamiento, salud y capacidad productiva para calcular el Índice de Mérito Reproductivo (IMR).
Además de la aplicación, se entrega un manual de usuario que enseñe fácilmente la utilización del programa y un compromiso a 2 horas dedicadas a mantenimiento y mejoría de la aplicación.
Esta aplicación finalmente no tiene la intención de competir o reemplazar a DelPro, sino usar sus propios datos procesados para auxiliar a CAETEC en la identificación del potencial en cada animal o de su posible ineficiencia.

7. Aportaciones y reflexiones individuales

7.1. Daniel Queijeiro Albo

Al iniciar este semestre tenía muy altas expectativas sobre todo el contenido que veríamos durante la materia debido al enorme auge que estamos viviendo actualmente con la IA. Esto no era solamente porque me interesaba añadir a mi CV que sé de IA, sino porque también quería aprender cómo es que realmente funcionan estas herramientas como ChatGPT que surgieron “de la nada” y parecieran seres omnipresentes que te ayudan con todo.

Estando ahora en la última semana del semestre puedo decir que ha sido un viaje intenso. Sin duda, el módulo de estadística representó el reto inicial más grande, principalmente por tantos temas que vimos y yo que era un completo principiante. Enfrentarme a una avalancha de conceptos como CLT, distribuciones, pruebas de hipótesis, regresiones y series de tiempo fue intimidante, pero al mismo tiempo revelador al notar que es un área de estudio que sustenta a la inteligencia artificial.

Para el módulo de machine learning creía yo que al realizar el proyecto del bloque pasado sería fácil también realizar este. Pero al correr las pruebas de mi primer modelo y ver que “overfitteado” le quedaba corto a los resultados comprendí que aquí tendría que aplicar todo el conocimiento base que había adquirido. Honestamente creo que fue mejor que mi primer modelo fracasara tan duro porque así pude cimentar los conceptos base para que mi siguiente iteración del modelo no tuviera los mismos problemas. Tuve que diagnosticar el origen del “overfit”, finalmente llegando a la conclusión que necesitaba ampliar mi dataset sin llegar a desbalancearlo. Después de hacer todo esto, los resultados de mi segundo modelo fueron una diferencia enorme, logrando un 96% de accuracy con datos del mundo real.

El módulo de Procesamiento de Lenguaje Natural fue clave para ver como conectaban los conceptos que veíamos en machine learning a todo el “boom” reciente de la IA, conceptos como los embeddings (que gracias a las clases estoy aplicando en mis prácticas) fueron justo lo que tenía en mente ver cuando inscribi la concentración. Ver, aprender y entrenar un modelo de GPT2 para la entrega del módulo me encantó porque fue entender esta “magia negra” que antes ocultaba cómo funcionaba ChatGPT y similares.

Además, los módulos de Big Data y Nube fueron una sorpresa muy grata que no esperaba ver en la concentración. Desde 5° semestre que hice mi primera cuenta en AWS he querido aprender más sobre Cloud Engineering y este módulo fue justo lo que deseaba. El acceso a la plataforma de AWS Academy me permitió aprender muchísimo sobre servicios de AWS que no sabía que existían. Además el aprender PySpark y que en la plática que nos dieron sobre el mundo laboral de Ciencia de datos lo mencionaran como una herramienta del día a día terminó de conectar este módulo no un extra sino como una parte esencial para dedicarme a esto.

Finalmente, el reto con el CAETEC fue ideal para aprender sobre la metodología CRISP-DM. Como buena metodología, esta nos obligó a mí y a mi equipo a mantener un orden y una estructura rigurosa para garantizar el éxito. Sin la aplicación de CRISP-DM, el resultado del proyecto sería sin lugar a duda otro. Gracias a la disciplina de entender primero el negocio, luego los datos y finalmente los objetivos de minería, nos vimos en la necesidad de realizar varias iteraciones de los modelos para cumplir con las metas que nos propusimos. Además, la retroalimentación constante de los profesores y socios formadores nos ayudó a plantear un proyecto que, estoy seguro, logra generar conocimiento, tal como debe ser el objetivo final de todo proyecto de minería de datos.

En cuanto a mis contribuciones individuales, debo mencionar que durante las fases iniciales mi enfoque principal estuvo en resolver los retos de mi proyecto individual del módulo de Machine Learning, por lo que mi participación inicial en el reto fue menor, aunque siempre tome una tarea por lo menos. En la fase de Business Understanding definí los objetivos de negocio e identifiqué las plataformas existentes del socio formador, mientras que en Data Understanding me encargué de documentar ejemplos de los datos que nos fueron otorgados al inicio del bloque de las vacas del rancho. En Data Preparation, documenté los atributos derivados, explicando su cálculo y justificación.
En el Modelado desarrollé el modelo de comportamiento v2.1 con Random Forest, donde usando GridSearch busqué los mejores hiperparametros para el modelo. Así obtuve el modelo que fue seleccionado para integrarse en la aplicación final. En Evaluación evalué los resultados de los modelos desarrollados para el comportamiento y sanidad y, finalmente, en Deployment generé los planes de despliegue, monitoreo y mantenimiento. También quiero mencionar que desarrollé un frontend y backend en la nube para el proyecto, aunque al final descartamos la aplicación web a favor de una aplicación de escritorio, donde apoye en el componente que describe los resultados otorgados por los modelos.

Como pequeña conclusión, termino este semestre con mucha satisfacción al ver que todas mis expectativas se cumplieron. Entré como alguien sin la menor idea de IA y salgo con un conjunto de habilidades técnicas y conocimientos que considero invaluables. Todos los módulos me enseñaron algo único y gracias a sus entregables puedo decir que me llevo la seguridad de saber enfrentar problemas complejos, la resiliencia para iterar ante el fallo y la certeza de que cuento con las herramientas necesarias para aportar valor real en la industria de los datos.

Evidencias en el PVG

Plan Data preparation

Fase del proyecto	Entregable	Actividad	Personas	Estado
		Documentar los atributos derivados, cómo se calcularon y por qué son relevantes (basarse en las notas de lo que nos dijo aurelio)	Daniel Queijeiro Albo	Revisado

Plan Modeling

Fase del proyecto	Entregable	Actividad	Personas	Estado
	Modelo Comportamiento V2.1	Descripcion del modelo	Daniel Queijeiro Albo	Revisado
	Modelo Comportamiento V2.1	Configuracion de parametros	Daniel Queijeiro Albo	Revisado
	Modelo Comportamiento V2.1	Codificacion del modelo	Daniel Queijeiro Albo	Revisado

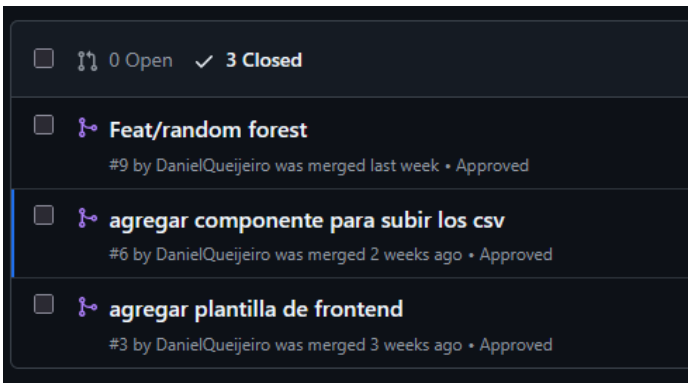
Tabla_3

Fase del proyecto	Entregable	Actividad	Personas	Estado
Evaluacion	Evaluacion de resultados		Nombre	Listo para revisión
		Evaluacion de resultados de modelado	Mauricio Anguiano Juárez Daniel Queijeiro Albo	Listo para revisión

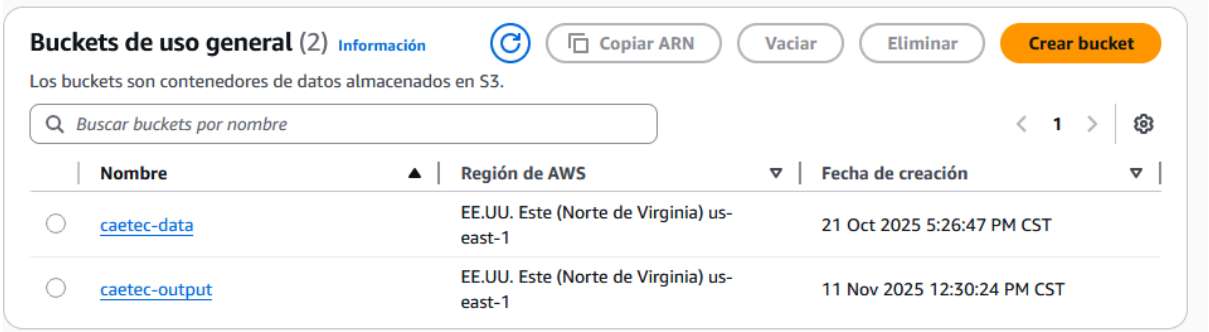
Deployment

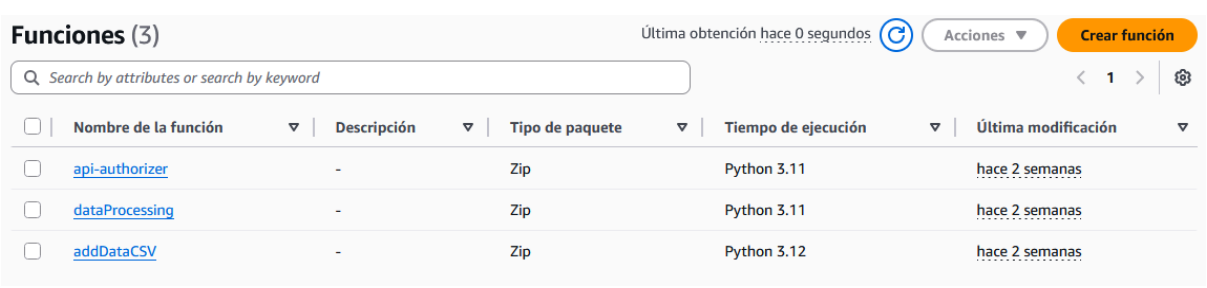
Fase del proyecto	Entregable	Actividad	Personas	Estado
		Generar plan de despliegue	Daniel Queijeiro Albo	Revisado
		Generar plan de monitoreo y mantenimiento	Daniel Queijeiro Albo	Revisado
		Crear buckets de S3 para almacenamiento de datos	Diego Ricardo Alfaro Pinto Daniel Queijeiro Albo	Deprecated
		Crear funcion lambda para el procesamiento de dat	Diego Ricardo Alfaro Pinto Daniel Queijeiro Albo	Deprecated
		Crear funcion lambda para la carga de archivos	Diego Ricardo Alfaro Pinto Daniel Queijeiro Albo	Deprecated
		Crear funcion lambda para ejecucion de prediccion	Diego Ricardo Alfaro Pinto Daniel Queijeiro Albo	Deprecated
		Crear funcion lambda para obtener los resultados d	Diego Ricardo Alfaro Pinto Daniel Queijeiro Albo	Deprecated
		Crear funciones lambda	Diego Ricardo Alfaro Pinto Daniel Queijeiro Albo	Deprecated
		Cargar archivos CSV a la página	Diego Ricardo Alfaro Pinto Daniel Queijeiro Albo	Deprecated
		Conectar página a los endpoints	Diego Ricardo Alfaro Pinto Daniel Queijeiro Albo	Deprecated
		Probar la API con Postman	Diego Ricardo Alfaro Pinto Daniel Queijeiro Albo	Deprecated
		Enseñar resultados en la página	Diego Ricardo Alfaro Pinto Daniel Queijeiro Albo	Deprecated
		Manejo de errores	Diego Ricardo Alfaro Pinto Daniel Queijeiro Albo	Deprecated
		Crear el cuadro con la interpretación de los resultados	Diego Isaac Fuentes Juvera Daniel Queijeiro Albo	En progreso

Evidencias en Github



Evidencias en AWS





7.2. Diego Alfaro Pinto

Al inicio del semestre sentía miedo al haber tomado la decisión de cursar esta concentración, ya que mis habilidades y conocimientos sobre IA eran muy limitados, gracias a todo lo que he aprendido puedo decir que fue la decisión correcta, todo lo que aprendí, todo lo que vivi, se que es conocimiento que será crucial en mi vida laboral y en mi futuro.

Una de las piezas de aprendizaje mas fundamental que tuve este semestre fue mi modelo de clasificación de especies de hongos, este proyecto fue muy interesante ya que al principio mi idea era hacer una clasificación de si los hongos eran comestibles o no, pero debido a que no tenia etiquetas correctas, y etiquetar las miles de imagenes del dataset iba a tomar mas tiempo del que tenia disponible, para este proyecto utilice una CNN de 5 capas para la clasificación de imagenes, luego de muchas horas de entrenamiento logre obtener un accuracy del 17%, que aunque no parece mucho, a comparacion de la capacidad de una persona promedio de clasificar una especia de hongo especifica, este 17% resulto ser bastante bueno, no obstante decidi por tambien utilizar transfer learning para poder obtener un modelo muchísimo mejor, lo que me llevo a usar EfficientNetB0, ya que este modelo esta entrenado justamente para clasificacion de imagenes, con este modelo logre obtener un accuracy del 60%, luego de el entrenamiento y el testing, decidi hacer predicciones descargando imagenes de internet de otros hongos, incluso el hongo de mario bros, y a pesar de que algunos hongos no estaban incluidos en las etiquetas, los modelos lograban predecir hongos que se parecian mucho a los que se estaba probando. Esto fue posible gracias al módulo 2 de la concentración, donde pude aprender cómo funcionaban algunas arquitecturas de deep learning en mayor detalle y no solo pensar en ellas como una magia y entender que es lo que estaba sucediendo internamente. Saber programar y evaluar estos modelos es de las cosas más valiosas que me llevo de este semestre y que estoy seguro que me ayudaran mucho en el campo laboral.

Uno de los descubrimientos más personales y potentes fue redescubrir el valor de los procesos "analógicos" en un mundo digital. La obligatoriedad de tomar notas a mano para los exámenes transformó mi método de estudio. Me di cuenta de que escribir me permitía estructurar la lógica y visualizar la arquitectura de los problemas con una claridad que el teclado no ofrece. Esta práctica fue determinante, no solo para aprobar la materia, sino como una ventaja competitiva real durante mi preparación para una entrevista técnica con Amazon.

A nivel de procesos, la metodología CRISP-DM se convirtió en mi brújula. Entendí que antes de escribir una sola línea de código, es imperativo definir objetivos claros y tratar los datos adecuadamente. La tecnología es inútil si no responde a una pregunta de negocio bien formulada.

Otro de los grandes descubrimientos para mi fue la importancia de la estadística en estas tecnologías, ya que al trabajar con modelos que usaban muchos conceptos estadísticos como lo puede ser el ARIMA para predecir cosas como el PIB de méxico usando el PIB de estados unidos, donde usando series temporales pude sacar varias conclusiones sobre qué correlaciones tienen estos dos mercados. También fue muy importante el aprender a usar la estadística descriptiva para poder entender la calidad de los datasets que estamos utilizando, ya que estas herramientas hacen mucho más fácil el poder tomar decisiones sobre qué datos usar, que hacerle a los datos etc.

Los datos crudos del establo tenían historias que solo los expertos podían interpretar. La simbiosis entre nuestro análisis técnico y su conocimiento empírico fue lo que permitió limpiar los datos correctamente y generar valor. Fue un ejercicio de humildad intelectual y colaboración interdisciplinaria.

Una de las secciones del semestre donde más aprendí fue en el módulo de big data, ya que me ayudó a entender más a fondo qué es big data, como se identifica cuando un proyecto cae en el uso de este, que se necesita para desarrollar esto proyectos y cómo manipular grandes cantidades de datos. Aprender sobre diferentes tipos de bases de datos se me hizo muy refrescante, ya que no conocía muchas afuera de las SQL y NoSQL normales, también el uso de librerías como pyspark y sobre todo el uso de herramientas de visualización administradas como tableau fueron cruciales para poder generar conocimiento.

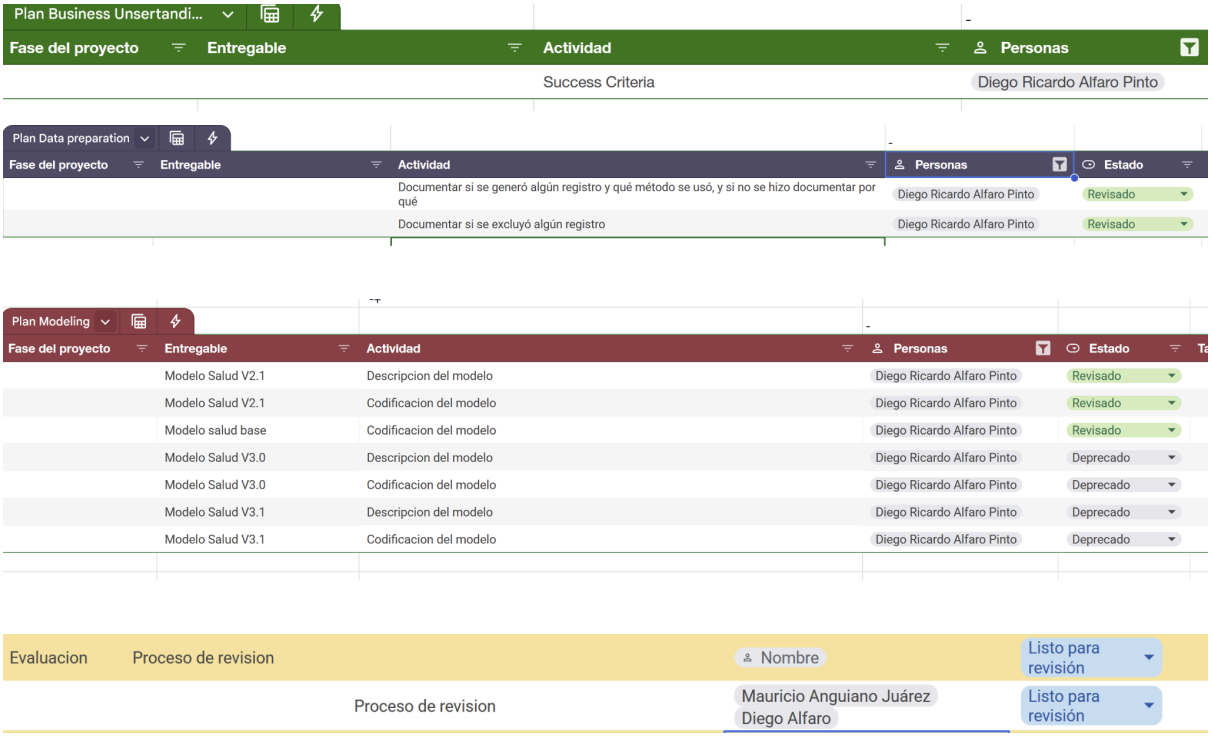
Todo esto cobró vida cuando nos enfrentamos al reto con el CAETEC, donde teníamos como objetivo predecir un índice para apoyar a la toma de decisiones de un establo de vacas productoras de leche. Uno de los grandes aprendizajes fue el aprender a escuchar a expertos en el área, ya que sin estos no podríamos haber llegado al resultado al que llegamos, todo esto unido por el pegamento que mencione arriba, la metodología CRISP-DM para poder sobrellevar este proyecto.

Mis aportaciones para este proyecto, aunque no son muchas, logre apoyar a mis compañeros cuando más lo necesitaban, aparte de poder crear 3 modelos, el modelo de sanidad base y la versión 2.1 del modelo de sanidad, este último siendo el modelo que se estará desplegando para el cliente, además de esto pude aportar estando a carga de crear los criterios de éxito del

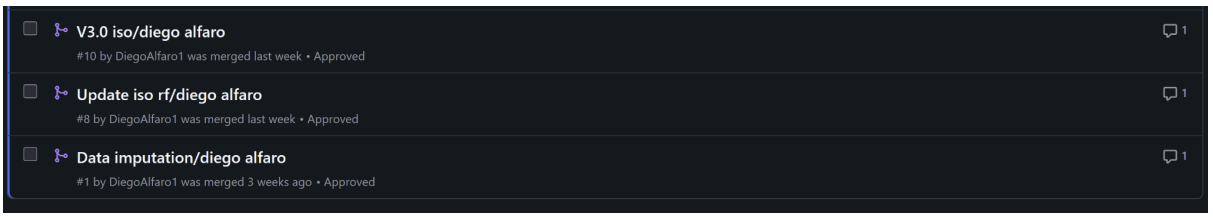
proyecto, al igual que el reporte final y la presentación final del proyecto, agregado a estas aportaciones, estuve encargado de desarrollar la página web y la arquitectura del sistema, aunque el equipo decidió no seguir adelante con este artefacto ya que significarán costos extras de infraestructura y capacitación extra que podría ser remediada con una aplicación de escritorio, también puede implementar un modelo usando un autoencoder, pero debido a la complejidad y bajo rendimiento se decidió por no ser utilizado.

Más allá de las aportaciones, haber trabajado este semestre con mi equipo, haber podido conocer personas nuevas y también todo el aprendizaje que obtuve fue muy significativo y marcó un antes y un después en mi carrera tanto académica como profesional, y estoy seguro que podré poner en práctica todos mis conocimientos en un futuro.

Evidencia



Fase del proyecto	Entregable	Actividad	Personas	Estado	Tamaño	Costo	Fecha planeada	Valor ganado	Costo real	Fecha real
Deployment	Aplicación web		Daniel Queijeiro Albo Diego Ricardo Alfaro Pinto	Deprecated						
		Crear buckets de S3 para almacenamiento de datos	Diego Ricardo Alfaro Pinto Daniel Queijeiro Albo	Deprecated						
		Crear funcion lambda para el procesamiento de dat	Diego Ricardo Alfaro Pinto Daniel Queijeiro Albo	Deprecated						
		Crear funcion lambda para la carga de archivos	Diego Ricardo Alfaro Pinto Daniel Queijeiro Albo	Deprecated						
		Crear funcion lambda para ejecucion de prediccion	Diego Ricardo Alfaro Pinto Daniel Queijeiro Albo	Deprecated						
		Crear funcion lambda para obtener los resultados d	Diego Ricardo Alfaro Pinto Daniel Queijeiro Albo	Deprecated						
		Crear funciones lambda	Diego Ricardo Alfaro Pinto Daniel Queijeiro Albo	Deprecated						
		Crear diseño en figma	Daniel Queijeiro Albo Diego R	Deprecated						
		Crear plantilla base con Vite	Daniel Queijeiro Albo Diego Ricardo Alfaro Pinto	Deprecated						
		Pasar diseño de figma a React	Daniel Queijeiro Albo Diego Ricardo Alfaro Pinto	Deprecated						
		Cargar archivos CSV a la página	Diego Ricardo Alfaro Pinto Daniel Queijeiro Albo	Deprecated						
		Conectar página a los endpoints	Diego Ricardo Alfaro Pinto Daniel Queijeiro Albo	Deprecated						
		Probar la API con Postman	Diego Ricardo Alfaro Pinto Daniel Queijeiro Albo	Deprecated						
		Enseñar resultados en la página	Diego Ricardo Alfaro Pinto Daniel Queijeiro Albo	Deprecated						
		Manejo de errores	Diego Ricardo Alfaro Pinto Daniel Queijeiro Albo	Deprecated						
		Reporte Final Socio	Diego Ricardo Alfaro Pinto	En progreso						
		Generar presentacion final	Diego Ricardo Alfaro Pinto	En progreso						



7.3. Diego Isaac Fuentes Juvera

Desde que iba en la prepa la IA ha sido un tema que me ha interesado profundamente. Siempre me pareció fascinante el cómo podíamos hacer que una computadora aprendiera prácticamente cualquier cosa de forma automática. En ese tiempo veía el tema como algo muy misterioso y que me infundía tanto respeto como miedo pues sentía que era una herramienta muy poderosa que nunca iba a ser capaz de comprender, o por lo menos no al nivel de las tecnologías del estado de arte.

Mis primeros acercamientos fueron los videos de 3 Blue 1 Brown sobre redes neuronales hace más de 6 años; logré entender las partes más sencillas como el forward pass o el hecho de que todo se pueda describir con funciones, pero al momento de entrar a cosas como el Back propagation mis conocimientos de matemáticas y programación hasta ese momento me limitaban. Sabía que tenía (y aún tengo) un largo camino por recorrer en mi misión de vida: Entender el mundo.

Más tarde, ya en mi tercer semestre de la universidad, mientras realizaba unas prácticas profesionales me encomendaron la tarea de crear un módulo de IA para analizar curriculums en una plataforma llamada Odoo. Mi jefe, quien me encargó la tarea, no conocía prácticamente nada del tema; la única ayuda que tenía eran videos de youtube y un libro enorme de 2014 con técnicas de machine learning como árboles de decisión, redes neuronales básicas y técnicas de procesamiento de lenguaje natural como conteo de palabras. Este reto fue un punto de inflexión pues me forzó a estudiar muchísimo sobre machine learning haciendo que me diera cuenta que el proyecto era inviable sin muchísimos curriculums muy diversos y clasificados con los que entrenar un modelo. Antes de dejar la empresa terminé el diseño del módulo bajo el supuesto que consiguieran suficientes datos de entrenamiento para ejecutarlo en el futuro, junto a algunas alternativas para ejecutar el proyecto sin machine learning. No voy a mentir, lo que aprendí en ese tiempo no solo me ayudó a entender el mundo si no que también me indujo mucho miedo de los modelos de lenguaje pues sentía que me iban a reemplazar, pero lo más importante es que este miedo se terminó convirtiendo en una semilla y en una curiosidad infinita por comprender la tecnología, desarrollarla y usarla para bien mío y de la humanidad.

Cuando me enteré que la concentración existía supe que era mi camino, pero al hacer el proceso de inscripción me dijeron que por procesos burocráticos que nadie entiende no había quedado dentro y me habían asignado a una concentración que no me serviría de nada para mi formación. Tuve que hablar con todo mundo, preguntar a todos los profes y al final mi director de carrera, Ricardo, me ayudó a conseguir un lugar dentro de la concentración. Viendo al pasado no podría estar más agradecido con ese conjunto de coincidencias y de buenas voluntades que me permitieron estar aquí, escribiendo estas palabras.

La concentración para mí fue un círculo de pasión. Pasión por aprender y pasión por superarme. Aunque a veces no se notara o perdiera el norte sentí todas las clases como una oportunidad de hacer lo que me gusta y de superarme. Aprendí muchísimo tanto a nivel técnico cómo a nivel de gestión de proyectos, y hoy estoy orgulloso de presentar mis aprendizajes más importantes y mis aportaciones al éxito de este proyecto.

Módulo de estadística.

Antes de este periodo no sabía ni un cuarto de lo que hoy sé de estadística. Sabía lo que era la moda, la desviación estándar, había visto distribuciones normales, pero nunca la había tratado como la herramienta tan poderosa que és. Aprender sobre coeficientes estadísticos como la R^2, el P-Value,lal T-Statistic y las distribuciones de probabilidad me hizo entender muchas cosas más allá de la inteligencia artificial, por ejemplo el cómo saber si los resultados de mis experimentos y de mi trabajo

son relevantes y confiables y saber cómo sustentarlos con bases matemáticas, entender cómo se pueden manipular las estadísticas incluso para fines políticos, y en general a cómo saber qué cosas creer y que nó en base a interpretaciones correctas. Esto me ayudó mucho en el proyecto al evaluar mis modelos, entender mis métricas, y saber si mi trabajo aportaba el valor que se esperaba.

Mi aprendizaje sobre las series temporales con modelos como ARIMA/SARIMA y conceptos como estacionalidad y temporalidad me ayudaron no solo a entender el presente, sino también a mejorar mi toma de decisiones con respaldo matemático en temas de economía y del rendimiento de mis equipos de trabajo, que se puede ver afectado por cosas como el día de la semana.

Estos conocimientos fueron clave al integrarlos para mi proyecto del módulo de deep learning donde desarrolle un modelo generativo de lenguaje que está basado y sustentado completamente en distribuciones probabilísticas al no tener solo una respuesta correcta y en series temporales al seguir una secuencia lógica donde la posición de cada palabra puede cambiar por completo el resultado.

Módulo de machine learning y deep learning.

Para mí estos módulos fueron de los más divertidos e interesantes de toda la concentración, pues fue lo que hizo que dejara de ver a la IA como una caja negra y me sintiera capaz de usarla para resolver muchos problemas. El entender cómo funciona una red neuronal e implementarla desde 0 fue lo que me permitió seguir comprendiendo todos los avances de años recientes que han permitido logros tan increíbles como los modelos de lenguaje modernos y me ayudó a construir un conjunto de herramientas que me han permitido resolver problemas complejos como crear mi propio modelo de lenguaje.

El deep learning está tan integrado en nuestra vida diaria que lo vemos como algo común cuando en realidad es un milagro de la ingeniería. Entender cómo diferentes modelos atacan diferentes problemas y conocer una gran variedad de modelos diferentes me permitió superar un reto muy grande en el proyecto con el CAETEC: la falta de etiquetas, pues pudimos implementar modelos de aprendizaje no supervisado con el objetivo de encontrar anomalías para sustentar nuestros resultados.

Este módulo también me enseñó que lo más importante para un buen modelo son buenos datos, pues un modelo nunca puede ser mejor que sus datos de entrenamiento. Técnicas de normalización, de aumentación, y de generación de datos derivados permitieron que mis modelos fueran mucho más robustos y que pudieran ajustar sus parámetros de forma mucho más efectiva. Ahora soy capaz de evaluar la viabilidad de un proyecto de machine learning antes de siquiera empezar a programar, solo con estudiar los datos tanto en su cantidad como su calidad y diversidad, permitiéndome generar proyectos exitosos en conjunto a mi abanico de herramientas.

Módulo de Big Data y Cloud Computing

Habiendo trabajado como data engineer en PPG este módulo resonó mucho conmigo, pues pude constatar que todos y cada uno de los temas que vimos en clase son aplicados en la industria real en empresas de nivel global. A pesar de que ya había trabajado con tecnologías como Spark o AWS, nunca había recibido una educación formal y este módulo me permitió comprender por qué esas tecnologías son la elección de muchísimas empresas y cómo usarlas siguiendo las mejores prácticas. Gracias a este módulo me siento capaz de tomar decisiones de infraestructura y tecnología de datos en cualquier empresa a la que llegue, teniendo la certeza de que podré aportar el mayor valor posible para lograr cualquier objetivo de negocio que me proponga.

Herramientas como Tableau me permitirán hacer propuestas de valor en el futuro que cualquier interesado no técnico pueda entender, ya sean directivos o gente de otras áreas de negocio, y siento que me aporta mucho valor como profesionista al ahora ser capaz de comunicar de qué manera le beneficia a las personas el trabajar conmigo. El uso de cloud computing y Spark también me permitirán aumentar muchísimo la escala y el valor de mis proyectos al poder hacer uso de Big Data para proyectos de deep learning en entornos mucho más masivos y con más impacto que los modelos que puedo entrenar localmente en mi computadora.

Módulo de Procesamiento de Lenguaje Natural.

Este módulo fue el más intensivo tecnológica y matemáticamente hablando, pero permitió afianzar mis conocimientos como ingeniero al darme un entendimiento total de las tecnologías que están en boca de todos: Los LLM. Aquí fue donde puse a prueba los conocimientos que obtuve durante el módulo para poder entender tecnologías de estado de arte como las redes con mecanismos de atención, los tokenizadores tan robustos creados por empresas como OpenAI, o los embeddings que permiten a una computadora entender a los humanos, y esto me sirvió mucho en mi implementación manual de un modelo de lenguaje ya que además de poder implementar una arquitectura robusta pude mejorar mucho mi modelo usando técnicas de NLP previas al entrenamiento, principalmente utilicé el tokenizador de ChatGPT-2 que funciona a nivel de morfema y lexema, permitiéndome pasar de un modelo que generaba algo 'parecido' al inglés a un modelo con el que se puede conversar de forma coherente aunque limitada. El módulo también me ayudó a entender cómo sacarle el mayor provecho a los modelos de lenguaje, pues ahora sé en su mayoría cómo funciona el cobro por token y cómo optimizar mis prompts para obtener algún resultado esperado, aunque aún así evito mucho usarlos.

CRISP-DM

A lo que yo me quiero dedicar cuando salga de la carrera es a la administración de proyectos, por lo que me hizo muy feliz ver que la concentración también tenía un enfoque en gestionar nuestra forma de trabajo para poder tener un portafolio de proyectos basto y exitoso, abriendome la puerta a trabajar con cualquier industria aunque su conocimiento técnico sea prácticamente nulo. Lo que más aprendí sobre la metodología son todas las cosas que se tienen que tener en cuenta para que un proyecto no sea un fracaso y la importancia que tiene cada fase para aportar a esto, desde definir buenos objetivos de negocio y buenos criterios de aceptación que permitan darle valor al cliente como objetivos de minería de datos que permitan tener un proyecto en tiempo y forma. Algo que me gustó también fue que tuvimos la oportunidad de adaptar la metodología a un enfoque más ágil, iterando en el proceso de todas las fases a medida que teníamos un mejor entendimiento de la problemática, de los datos con los que contamos, de las herramientas a nuestro alcance, y de nuestra forma de trabajo. Todo el semestre fue un ir y venir de fases que terminó por refinar el proyecto y consolidar el producto final como algo de lo que estar orgullosos.

Reto

El reto fue el lugar indicado para integrar todos estos aprendizajes y demostrar nuestra capacidad de hacer ingeniería. Haber seguido cada una de las fases de la metodología y haber aplicado los conocimientos técnicos de todos los módulos me ha permitido sentirme listo para afrontar proyectos reales, desde comprender la problemática del socio, entender los datos, explorar los modelos, encontrar valor en nuestros resultados y entregar una plataforma funcional. Construimos

sobre nuestros errores y eso nos permitió tener un conocimiento más fundamental y afianzado de la importancia de todo lo que aprendimos.

Conclusión

Este semestre me sorprendió para bien y me dejó encantado con todo lo que aprendí. Me siento mucho más capaz de aportar valor a cualquier ambiente en el que llegue a trabajar, y por fin se me quitó parte del miedo que tenía a que la IA me reemplazara al entender cómo puedo seguir aportando valor de formas que son mucho más difíciles de replicar. Agarrar al toro por los cuernos me ayudó a trabajar en mi formación personal y profesional, y me siento contento de lo que aprendí de mis compañeros y orgulloso de lo que les aporté durante el semestre. ¡Muchas gracias!

Actividades completadas durante la realización del reto:

1. Business understanding:

Actividad	Personas	Estado
	Diego Isaac Fuentes Juvera	Revisado
Hacer PVG completo	Diego Isaac Fuentes Juvera	Revisado
Hacer plan de business understanding	Diego Isaac Fuentes Juvera	Revisado
Hacer plan de data understanding	Diego Isaac Fuentes Juvera	Revisado
Hacer plan de data preparation	Diego Isaac Fuentes Juvera	Revisado
Hacer plan de modeling	Diego Isaac Fuentes Juvera	Revisado
Definir cuál es el objetivo principal del negocio	Diego Isaac Fuentes Juvera	Listo para revis...
Definir cuales son los criterios de éxito de ese obejtivo	Diego Isaac Fuentes Juvera	Listo para revis...
Validar con el socio formador	Diego Isaac Fuentes Juvera	Listo para revis...
Revisar video de CRISP-DM	Diego Isaac Fuentes Juvera	Revisado
Añadir y quitar tareas de las fases de CRISP-DM	Diego Isaac Fuentes Juvera	Revisado

2. Data understanding:

Actividad	Personas	Estado
Estadisticas basicas	Diego Isaac Fuentes Juvera	Revisado

3. Data Preparation

Actividad	Personas	Estado
Generar etiquetas finales	Diego Isaac Fuentes Juvera	Revisado
Obtener pendiente de decaemiento de producción de leche	Diego Isaac Fuentes Juvera	Revisado
Obtener tiempo promedio de ordeña	Diego Isaac Fuentes Juvera	Revisado
Obtener el total de patadas por cuarto	Diego Isaac Fuentes Juvera	Revisado
Obtener las patadas por día, por sesión de ordeño, y por hora en la máquina.	Diego Isaac Fuentes Juvera	Revisado
Obtener el indice de mastitis promedio y total	Diego Isaac Fuentes Juvera	Revisado
Obtener los pezones no encontrados por cuarto	Diego Isaac Fuentes Juvera	Revisado
Documentacion de datos calculados	Mauricio Anguiano Juárez Diego Isaac Fuentes Juvera	Revisado
Normalizar datos	Diego Isaac Fuentes Juvera	Revisado
Dar formato a los datos	Diego Isaac Fuentes Juvera	Revisado

4. Modeling

Entregable	Actividad	Personas	Estado
Modelo Comportamiento V1.0	Configuracion de parametros	Diego Isaac Fuentes Juvera	Revisado
Modelo Comportamiento V1.0	Codificacion del modelo	Diego Isaac Fuentes Juvera	Revisado
Modelo Salud V1.0	Configuracion de parametros	Diego Isaac Fuentes Juvera	Revisado
Modelo Salud V1.0	Codificacion del modelo	Diego Isaac Fuentes Juvera	Revisado

5. Evaluation

No tuve actividades en la fase de evaluación.

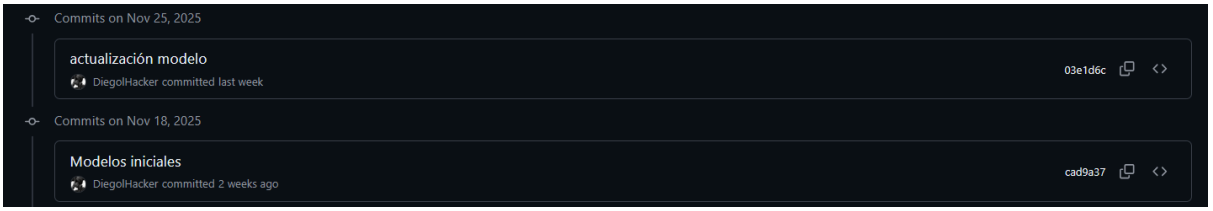
6. Deployment

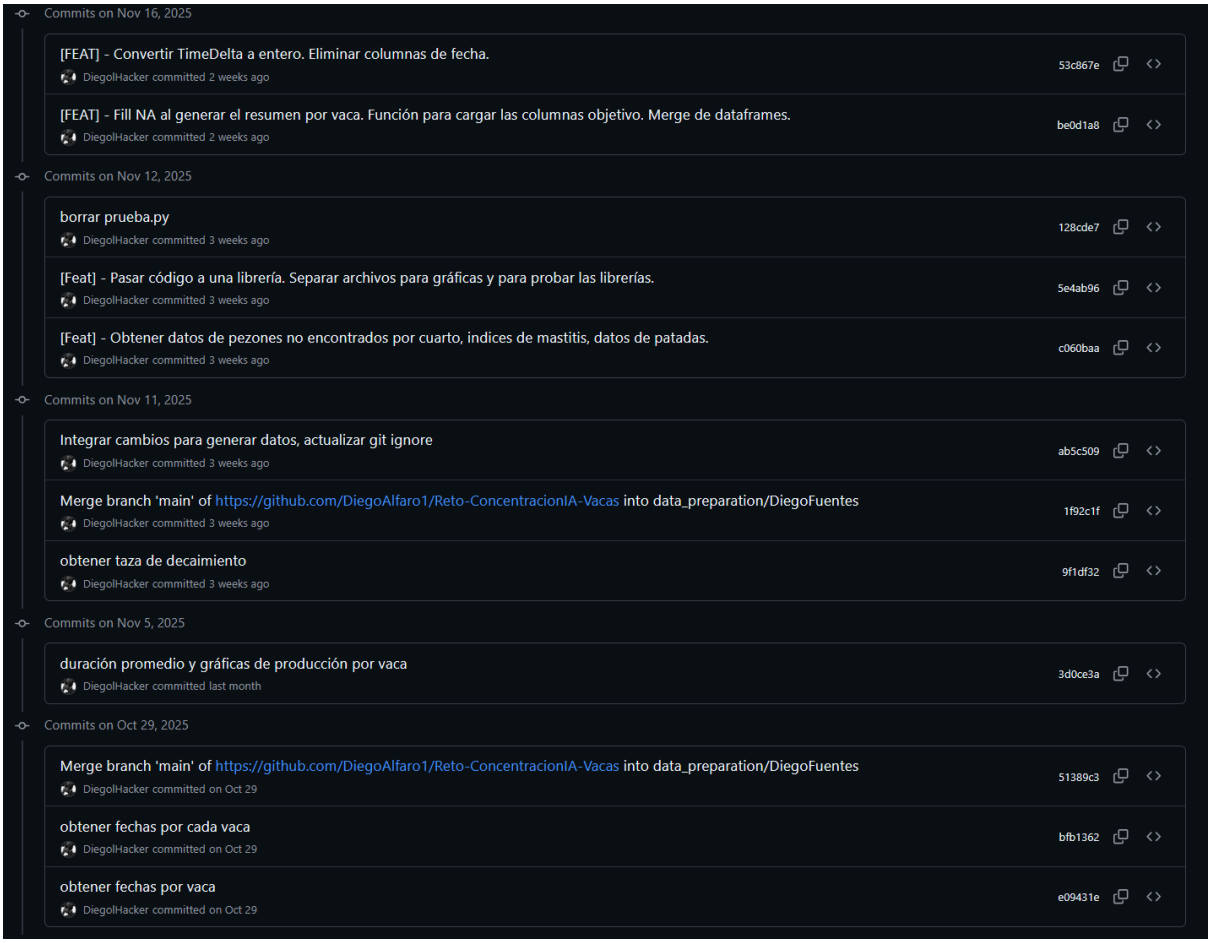
Entregable	Actividad	Personas	Estado
Aplicación/Entregable Final		Diego Isaac Fuentes Juvera	
	Definir que métricas se mostrarán en la interfaz	Diego Isaac Fuentes Juvera Mauricio Anguiano Juárez	Listo para revisión
	Generar diseño de la aplicación en Figma	Diego Isaac Fuentes Juvera Mauricio Anguiano Juárez	Listo para revisión
	Crear esqueleto de la aplicación basado en MVC	Diego Isaac Fuentes Juvera	Listo para revisión
	Programar barra superior	Diego Isaac Fuentes Juvera	Listo para revisión
	Generar componentes de interfaz reutilizables	Diego Isaac Fuentes Juvera	Listo para revisión
	Ensamblar componentes de la aplicación sin contenido	Diego Isaac Fuentes Juvera	Listo para revisión
	Llenar campos de texto para métricas y resultados	Diego Isaac Fuentes Juvera	Listo para revisión
	Implementar carga de CSV	Diego Isaac Fuentes Juvera	Listo para revisión
	Separar archivos necesarios para que funcione el modelo	Diego Isaac Fuentes Juvera	Listo para revisión
	Cambiar las funciones de S3 por funciones locales	Diego Isaac Fuentes Juvera	Listo para revisión
	Conectar el FrontEnd con el modelo	Diego Isaac Fuentes Juvera	Listo para revisión
	Manejar errores en el CSV y al correr en modelo	Diego Isaac Fuentes Juvera	Listo para revisión
	Crear el cuadro con la interpretación de los resultados	Diego Isaac Fuentes Juvera Daniel Queijeiro Albo	En progreso
	Generar ejecutable	Diego Isaac Fuentes Juvera	No empezado
	Generar manuales de usuario de la aplicación	Mauricio Anguiano Juárez Diego Isaac Fuentes Juvera	Listo para revisión

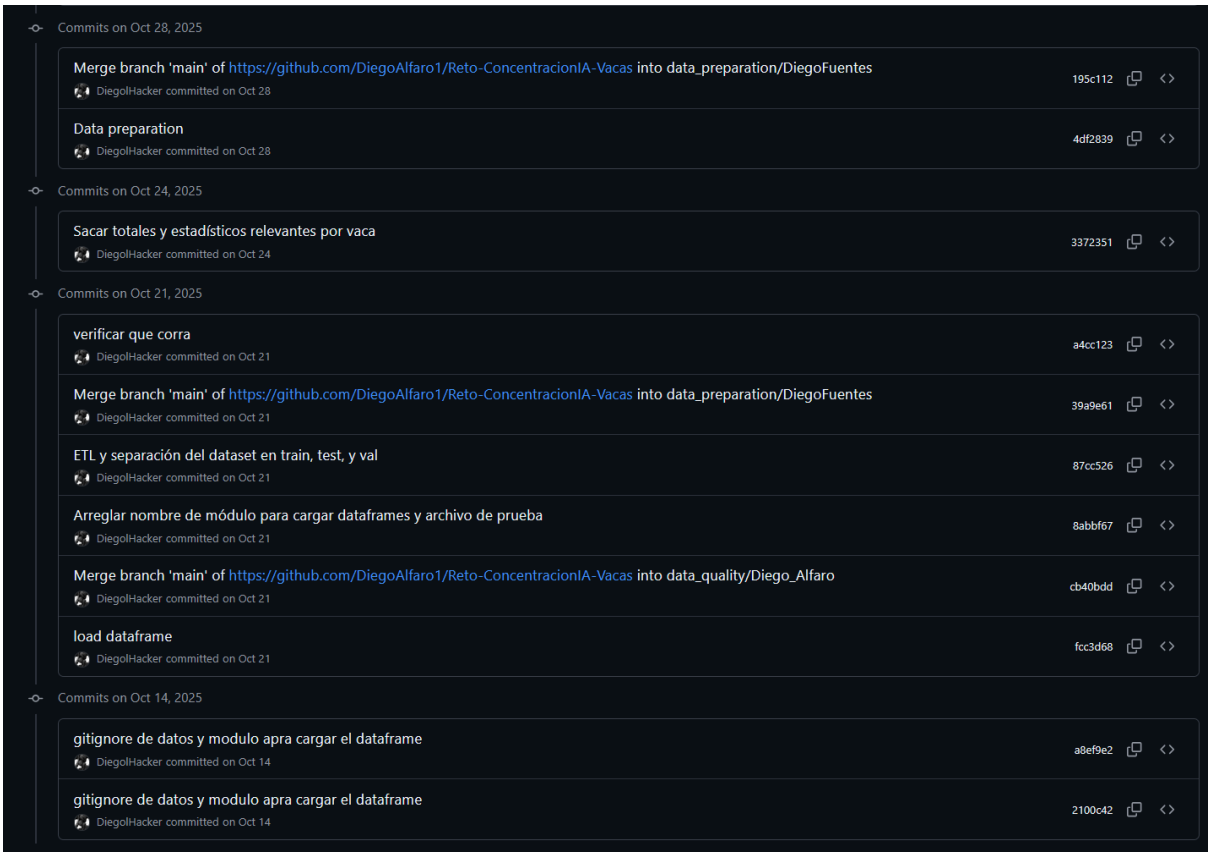
Commits realizados al repositorio de GitHub del proyecto:

Los commits se pueden consultar en el repositorio del proyecto.
Mis principales tareas fueron:

- Generar el .gitignore
- Generar una librería para cargar, limpiar, y formatear el data frame (Generar columnas multi-index, rellenar valores vacíos, generar columnas de fecha y de hora, cambiar formato de datos de string a formatos de pandas)
- Generar librería para partir el dataset en entrenamiento, prueba, y validación.
- Generar columnas de totales y promedios para las columnas que estaban separadas por cuarto (DI, DD, TI, TD).
- Generar estadísticos de todas las columnas numéricas para cada vaca.
- Generar notebook de data preparation donde se generaron valores derivados: Duración promedio de ordeño, frecuencia de ordeño, cantidad de pezones no encontrados, cantidad de patadas, patadas por ordeño, por hora dentro del máquina, índice de mastitis promedio y total, tasa de decaimiento en la producción de leche después del punto de producción máxima del ciclo productivo, etc.
- Generación de gráficas de producción para cada vaca.
- Generar la primera versión de los modelos de salud y comportamiento usando Isolation Forest.
- Crear la aplicación de escritorio que se entregará como producto final al socio formador.




Commits on Nov 16, 2025
[FEAT] - Convertir TimeDelta a entero. Eliminar columnas de fecha.
DiegolHacker committed 2 weeks ago
53c867e
[FEAT] - Fill NA al generar el resumen por vaca. Función para cargar las columnas objetivo. Merge de dataframes.
DiegolHacker committed 2 weeks ago
be0d1a8
Commits on Nov 12, 2025
borrar prueba.py
DiegolHacker committed 3 weeks ago
128cde7
[Feat] - Pasar código a una librería. Separar archivos para gráficas y para probar las librerías.
DiegolHacker committed 3 weeks ago
5e4ab96
[Feat] - Obtener datos de pezones no encontrados por cuarto, indices de mastitis, datos de patadas.
DiegolHacker committed 3 weeks ago
c060baa
Commits on Nov 11, 2025
Integrar cambios para generar datos, actualizar git ignore
DiegolHacker committed 3 weeks ago
ab5c509
Merge branch 'main' of https://github.com/DiegoAlfaro1/Reto-ConcentracionIA-Vacas into data_preparation/DiegoFuentes
DiegolHacker committed 3 weeks ago
1f92c1f
obtener taza de decaimiento
DiegolHacker committed 3 weeks ago
9f1df32
Commits on Nov 5, 2025
duración promedio y gráficas de producción por vaca
DiegolHacker committed last month
3d0ce3a
Commits on Oct 29, 2025
Merge branch 'main' of https://github.com/DiegoAlfaro1/Reto-ConcentracionIA-Vacas into data_preparation/DiegoFuentes
DiegolHacker committed on Oct 29
51389c3
obtener fechas por cada vaca
DiegolHacker committed on Oct 29
bfb1362
obtener fechas por vaca
DiegolHacker committed on Oct 29
e09431e


Commits on Oct 28, 2025
Merge branch 'main' of https://github.com/DiegoAlfaro1/Reto-ConcentracionIA-Vacas into data_preparation/DiegoFuentes
DiegolHacker committed on Oct 28
195c112
Data preparation
DiegolHacker committed on Oct 28
4df2839
Commits on Oct 24, 2025
Sacar totales y estadísticos relevantes por vaca
DiegolHacker committed on Oct 24
3372351
Commits on Oct 21, 2025
verificar que corra
DiegolHacker committed on Oct 21
a4cc123
Merge branch 'main' of https://github.com/DiegoAlfaro1/Reto-ConcentracionIA-Vacas into data_preparation/DiegoFuentes
DiegolHacker committed on Oct 21
39a9e61
ETL y separación del dataset en train, test, y val
DiegolHacker committed on Oct 21
87cc526
Arreglar nombre de módulo para cargar dataframes y archivo de prueba
DiegolHacker committed on Oct 21
8abbf67
Merge branch 'main' of https://github.com/DiegoAlfaro1/Reto-ConcentracionIA-Vacas into data_quality/Diego_Alfaro
DiegolHacker committed on Oct 21
cb40bdd
load dataframe
DiegolHacker committed on Oct 21
fcc3d68
Commits on Oct 14, 2025
gitignore de datos y modulo apra cargar el dataframe
DiegolHacker committed on Oct 14
a8ef9e2
gitignore de datos y modulo apra cargar el dataframe
DiegolHacker committed on Oct 14
2100c42

7.4. Jesus Ramirez Delgado

Este semestre representó uno de los periodos académicos más significativos en mi formación profesional, no solo por el nivel técnico abordado, sino por la claridad y estructura mental que adquirí sobre el mundo de la inteligencia artificial, su construcción y su aplicación práctica. A través de los módulos cursados entre las dos unidades de formación, entendí que la inteligencia artificial no es solo programación o algoritmos, sino un ciclo metódico, estructural, humano y ético, donde la capacidad de comprender problemas, transformar información y evaluar resultados es tan importante como saber programar modelos.

Fundamentos de la inteligencia artificial

El recorrido inició con estadística, donde aprendí que cualquier modelo solo puede ser tan bueno como el conocimiento que se tenga de los datos. Comprender distribución, hipótesis, significancia, regresión y cómo interpretar coeficientes fue clave para empezar a ver la inteligencia artificial como un proceso matemático con propósito, más allá del software. Este aprendizaje sirvió como plataforma conceptual para entender todos los demás módulos.

Posteriormente, el módulo de aprendizaje de máquina funcionó como mi puerta conceptual a la IA. Más que aprender algoritmos, descubrí cómo "funcionan" y por qué existen: backpropagation, el papel de la función de pérdida, sesgo y varianza o cómo diagnosticar un modelo. Construir modelos desde cero y reportar resultados fue un ejercicio que no solo fortaleció mi comprensión técnica, sino mi capacidad de documentación y análisis crítico.

El módulo de software para ciencia de datos permitió aterrizar la teoría anterior. Aquí comprendí que sin un buen tratamiento de datos ninguna técnica sirve, y que herramientas como Pandas, NumPy o frameworks como Scikit-Learn o TensorFlow habilitan experimentación, pero no sustituyen la necesidad de entender qué se está resolviendo. También aprendí el valor práctico del ETL, del "ensuciarse las manos" con datos reales y de visualizar resultados.

Por otro lado, el módulo de integración hardware–software amplió mi visión sobre cómo los modelos se consumen o ejecutan en el mundo real. Desde el uso de modelos preentrenados hasta la implementación y uso de herramientas no exploradas antes, este módulo se sintió como un conector entre teoría y aplicación práctica.

Sobre el aprendizaje de máquina, deep learning, y módulos de parte II

En la segunda parte de la concentración mi interés y gusto crecieron exponencialmente. El módulo de Big Data me enseñó a identificar cuándo realmente es necesario un enfoque de procesamiento masivo, entendiendo las 5V's y conociendo tecnologías habilitadoras como PySpark, infraestructuras cloud y herramientas como Tableau. Sin embargo, también identifiqué que no todos los problemas requieren Big Data, algo que aplicamos en el reto del semestre.

El siguiente módulo, enfocado en deep learning, fue uno de los más enriquecedores. Pude conectar arquitecturas con problemas reales, entender principios del aprendizaje profundo, técnicas que se diseñaron para resolver problemas de entrenamiento y ver cómo redes neuronales generan representaciones inteligentes del mundo. Conceptos como autoencoders, convoluciones, embeddings, RNNs, GRUs, LSTMs o VAEs dejaron de ser abstractos y cobraron sentido, no sólo como modelos sino como formas de pensar. Este módulo también me ayudó a ser más crítico con tecnologías como

los modelos de lenguaje actuales, reconociendo su complejidad, sesgos éticos y las bases históricas de su evolución.

El módulo de Procesamiento de Lenguaje Natural fue una extensión natural del módulo de deep learning, donde convergen temas como: embeddings, CNN 's, RNN y son el puente para ver arquitecturas y algoritmos más complejos como transformers que actualmente conforman las base de los LLM’s comerciales como chatgpt.

Finalmente, el módulo de computación en la nube permite conectar la IA con contexto operativo: diseño de arquitectura, seguridad, permisos, almacenamiento y despliegue. Pude aplicar todo esto en nuestro proyecto, lo cual consolidó mi aprendizaje.

Metodología

CRISP-DM fue más que una metodología; fue la columna vertebral del reto del semestre. Me enseñó que la IA no se trata solamente de entrenar modelos, sino de comprender el negocio, evaluar el impacto y tener disciplina de trabajo. Adaptamos la metodología a nuestro proyecto, y esto reforzó una lección muy valiosa: las herramientas funcionan solo cuando se ajustan al contexto real.

Reto

El reto integró absolutamente todo. Desde entender la necesidad del cliente, explorar los datos reales provenientes de CAETEC, corregir errores en el procesamiento, diseñar modelos, evaluar, fallar, replantear y entregar resultados — todo implicó aplicar teoría, trabajo en equipo y reflexión crítica sobre lo aprendido. No fue solo un proyecto técnico: fue un ejercicio formativo donde cada tropiezo y retroalimentación construyó conocimiento.

Reflexión Final

Este semestre se convirtió en uno de mis favoritos. Me exigió técnica, claridad mental y madurez profesional. Me dio un criterio sólido, me hizo crítico sobre la tecnología que hoy se utiliza sin cuestionar y me demostró que la disciplina y metodología son tan valiosas como el conocimiento técnico. Aprecio profundamente haber sido guiado por profesores con experiencia académica y profesional real, quienes nutrieron mi aprendizaje tanto en conocimiento como en conversaciones humanas.

Termino este semestre orgulloso por comprender cómo funciona la inteligencia artificial desde fundamentos matemáticos hasta arquitecturas complejas, por aplicar teoría a un problema real y por ver el resultado tangible de mi propio crecimiento. Hoy me considero no solo más competente, sino también más consciente del valor y la responsabilidad que implica construir inteligencia artificial.

Dejo el enlace a una reflexión más detallada por cada módulo de ambas partes (I y II) que hice sobre el curso por si es de interes: Click aquí

Aportaciones al reto

En este apartado enlistare mis aportaciones al reto.

Actividades de la metodología

1. Business Understanding.

Terminology	A01274723 Jesús Ramír...	Revisado
Agregar magnitud y responsables en plan de riesgos y contingencias	A01274723 Jesús Ramír...	Revisado
Determinar objetivos de minería de datos	A01274723 Jesús Ramír...	Revisado
Hacer objetivos de negocio 1:1 con objetivos de mineria de datos	A01274723 Jesús Ramír...	Revisado
Definir criterios de exito de minería de datos	A01274723 Jesús Ramír...	Revisado
Definir criterios de exito de minería de datos	A01274723 Jesús Ramír...	Revisado
Hacer plan de evaluation	A01274723 Jesús Ramír...	Revisado
Hacer plan de deployment	A01274723 Jesús Ramír...	Revisado
Agregar adaptaciones de CRISP-DM	A01274723 Jesús Ramír...	Listo para revisión

2. Data Understanding.

Fuente de datos	A01274723 Jesús Ramírez ...	Revisado
Condiciones de recoleccion	A01274723 Jesús Ramírez ...	Revisado
Problemas encontrados	A01274723 Jesús Ramírez ...	Revisado
Estructura de datos	A01274723 Jesús Ramírez ...	Revisado
Diccionario de datos	A01274723 Jesús Ramírez ...	Revisado
Distribuciones (histogramas, boxplots, conteo de categorías)	A01274723 Jesús Ramírez ...	Revisado
Relaciones entre variables (scatterplots, correlaciones)	A01274723 Jesús Ramírez ...	Revisado

Patrones temporales o espaciales (tendencias, estacionalidad, anomalías)	A01274723 Jesús Ramírez ...	Revisado
Hipótesis iniciales	A01274723 Jesús Ramírez ...	Revisado

Hipotesis general	A01274723 Jesús Ramírez ...	Revisado

Agregar adaptaciones de CRISP-DM	A01274723 Jesús Ramírez ...	Revisado
Agregar logs y acceso a los datos a Política de datos	A01274723 Jesús Ramírez ...	Revisado

3. Data Preparation.

Dataset 2.0	A01274723 Jesús...	Revisado

4. Modeling.

	Introduccion	A01274723 J...	Revisado

Modelo Comportamiento	Tecnica de modelado	A01274723 J...	Listo para revisión
Modelo Salud	Tecnica de modelado	A01274723 J...	Listo para revisión

Estrategia de división de datos (train/validation/test o k-fold)	A01274723 Jesús ...	Revisado
Codigo - Hacer el split	A01274723 Jesús ...	Revisado

Criterios de evaluación derivados de los Data Mining Success Criteria	A01274723 Jesús ...	Listo para revisión
Métricas de evaluación modelo(s) comportamiento	A01274723 Jesús ...	Listo para revisión
Metricas de evaluacion modelo(s) sanidad	A01274723 Jesús ...	Listo para revisión
Plan general de entrenamiento, validación y comparación de modelos	A01274723 Jesús ...	Listo para revisión

Modelo	Descripcion del modelo	A01274723 J...	Listo para

Comportamiento V2.0			revisión
Modelo Comportamiento V2.0	Configuracion de parametros	A01274723 J...	Listo para revisión
Modelo Comportamiento V2.0	Codificacion del modelo	A01274723 J...	Listo para revisión
Modelo Salud V2.0	Descripcion del modelo	A01274723 J...	Listo para revisión
Modelo Salud V2.0	Configuracion de parametros	A01274723 J...	Listo para revisión
Modelo Salud V2.0	Codificacion del modelo	A01274723 J...	Listo para revisión

Modelo Comportamiento MLP	Descripcion del modelo	A01274723 J...	Listo para revisión
Modelo Comportamiento MLP	Configuracion de parametros	A01274723 J...	Listo para revisión

Modelo Comportamiento base	Resultados de evaluacion del modelo	A01274723 J...	Listo para revisión
Modelo Comportamiento base	Obvservaciones	A01274723 J...	Listo para revisión
Modelo Salud base	Resultados de evaluacion del modelo	A01274723 J...	Listo para revisión
Modelo Salud base	Obvservaciones	A01274723 J...	Listo para revisión
Modelo Comportamiento V1.0	Resultados de evaluacion del modelo	A01274723 J...	Listo para revisión
Modelo Comportamiento V1.0	Obvservaciones	A01274723 J...	Listo para revisión
Modelo Salud V1.0	Resultados de evaluacion del modelo	A01274723 J...	Listo para revisión
Modelo Salud V1.0	Obvservaciones	A01274723 J...	Listo para revisión
Modelo	Resultados de evaluacion del modelo	A01274723 J...	Listo para

Comportamiento V2.0			revisión
Modelo Comportamiento V2.0	Obvservaciones	A01274723 J...	Listo para revisión
Modelo Salud V2.0	Resultados de evaluacion del modelo	A01274723 J...	Listo para revisión
Modelo Salud V2.0	Obvservaciones	A01274723 J...	Listo para revisión
Modelo Comportamiento V2.1	Resultados de evaluacion del modelo	A01274723 J...	Listo para revisión
Modelo Comportamiento V2.1	Obvservaciones	A01274723 J...	Listo para revisión
Modelo Salud V2.1	Resultados de evaluacion del modelo	A01274723 J...	Listo para revisión
Modelo Salud V2.1	Obvservaciones	A01274723 J...	Listo para revisión
Modelo Comportamiento MLP	Resultados de evaluacion del modelo	A01274723 J...	Listo para revisión
Modelo Comportamiento MLP	Obvservaciones	A01274723 J...	Listo para revisión

Agregar adaptaciones de CRISP-DM	A01274723 Jesús ...	En progreso

5. Evaluación

Aprovacion de modelos	A0127472...	En progreso
Prediccion de modelos	A0127472...	En progreso

Adaptaciones CRISP-DM	A0127472...	Listo para revisión

.

6. Deployment

Generar presentacion final	TODOS	En progreso
Adaptaciones CRSIP-DM	A01274723 Jesús Ra…	Listo para revisión

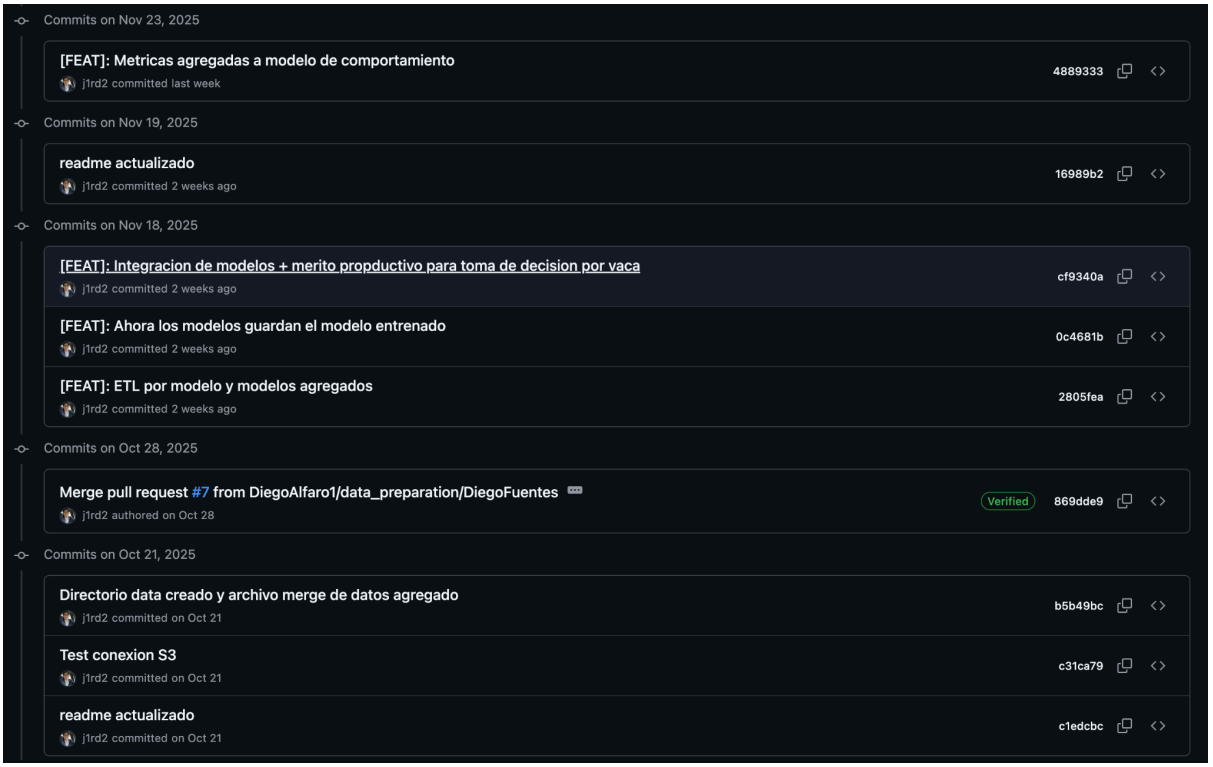
Actividades de codificación/modelos

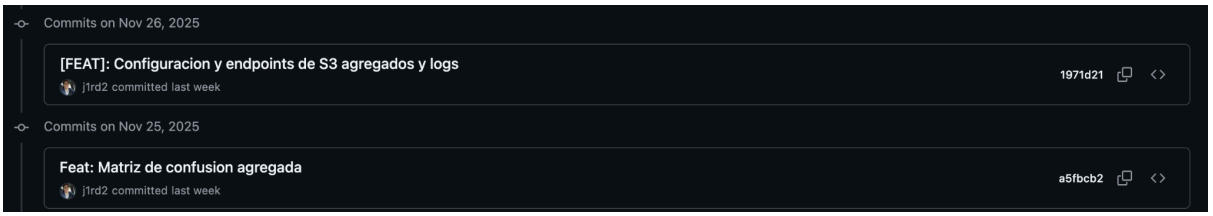
Los commits se pueden consultar en el repositorio del proyecto: Click aqui.

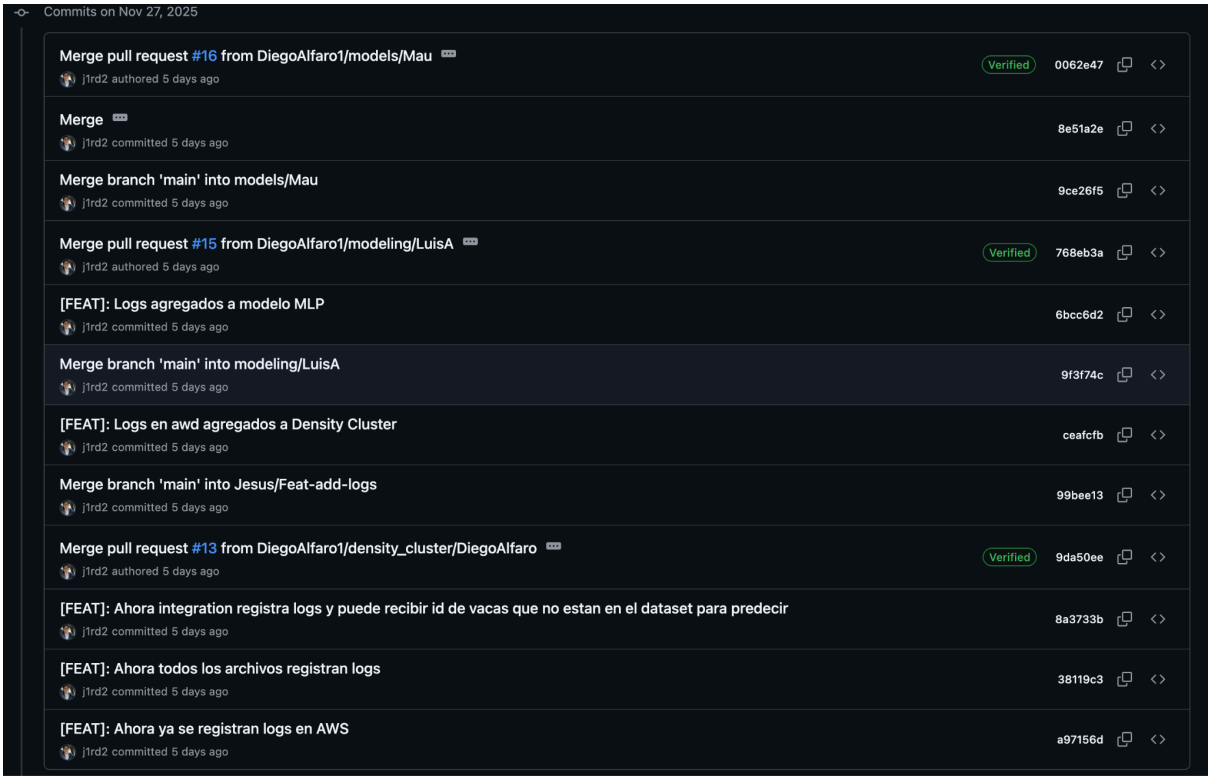
Las actividades principales que hice fueron:

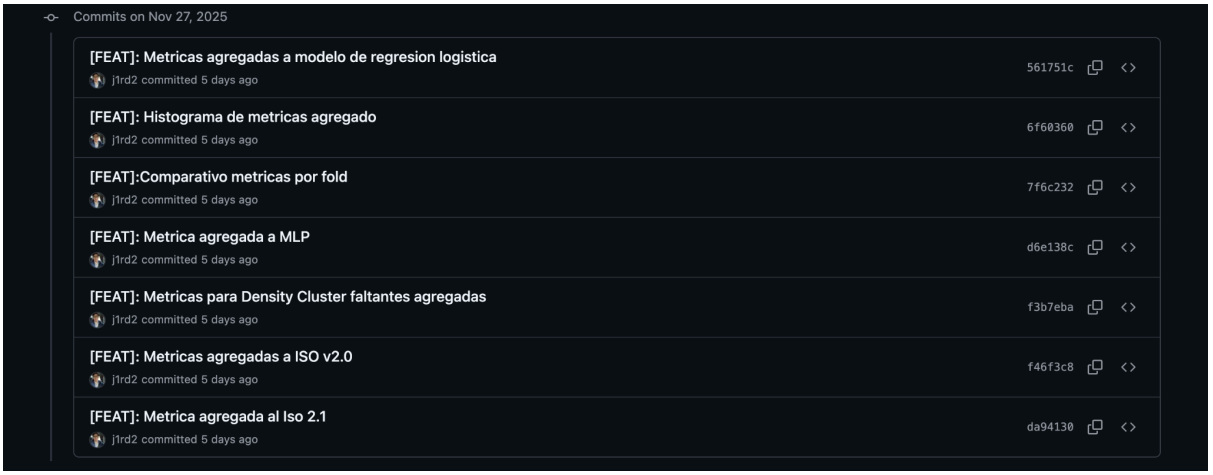
- Merge de datos: Unificar csv individual por vaca en uno solo.
- Arquitectura del IMR.
- Modelos:
  - Random Forest 2.0.
  - Isolation Forest 2.0.
- ETL 2.0: Script que genera 2 datasets para el modelo de comportamiento y sanidad.
- Índice de mérito productivo.
- Integración de modelos 1.0.
- Integración de modelo 2.0.
- Estandarizar métricas de acuerdo a nuestras métricas estipuladas en todos los modelos y los que les faltaban.
- Aplicar el sistema de logs a todos los modelos, y archivos que hagan alguna lectura y escritura de los datos.

A continuación evidencia de commis en orden cronológico:


Commits on Nov 23, 2025
[FEAT]: Metricas agregadas a modelo de comportamiento
j1rd2 committed last week
4889333
Commits on Nov 19, 2025
readme actualizado
j1rd2 committed 2 weeks ago
16989b2
Commits on Nov 18, 2025
[FEAT]: Integracion de modelos + merito propductivo para toma de decision por vaca
j1rd2 committed 2 weeks ago
cf9340a
[FEAT]: Ahora los modelos guardan el modelo entrenado
j1rd2 committed 2 weeks ago
0c4681b
[FEAT]: ETL por modelo y modelos agregados
j1rd2 committed 2 weeks ago
2805fea
Commits on Oct 28, 2025
Merge pull request #7 from DiegoAlfaro1/data_preparation/DiegoFuentes
j1rd2 authored on Oct 28
Verified
869dde9
Commits on Oct 21, 2025
Directorio data creado y archivo merge de datos agregado
j1rd2 committed on Oct 21
b5b49bc
Test conexion S3
j1rd2 committed on Oct 21
c31ca79
readme actualizado
j1rd2 committed on Oct 21
c1edcbc


Commits on Nov 26, 2025
[FEAT]: Configuracion y endpoints de S3 agregados y logs
j1rd2 committed last week
1971d21
Commits on Nov 25, 2025
Feat: Matriz de confusion agregada
j1rd2 committed last week
a5fbcb2


Commits on Nov 27, 2025
Merge pull request #16 from DiegoAlfaro1/models/Mau
j1rd2 authored 5 days ago
Verified
0062e47
Merge
j1rd2 committed 5 days ago
8e51a2e
Merge branch 'main' into models/Mau
j1rd2 committed 5 days ago
9ce26f5
Merge pull request #15 from DiegoAlfaro1/modeling/LuisA
j1rd2 authored 5 days ago
Verified
768eb3a
[FEAT]: Logs agregados a modelo MLP
j1rd2 committed 5 days ago
6bcc6d2
Merge branch 'main' into modeling/LuisA
j1rd2 committed 5 days ago
9f3f74c
[FEAT]: Logs en awd agregados a Density Cluster
j1rd2 committed 5 days ago
ceafcfb
Merge branch 'main' into Jesus/Feat-add-logs
j1rd2 committed 5 days ago
99bee13
Merge pull request #13 from DiegoAlfaro1/density_cluster/DiegoAlfaro
j1rd2 authored 5 days ago
Verified
9da50ee
[FEAT]: Ahora integration registra logs y puede recibir id de vacas que no estan en el dataset para predecir
j1rd2 committed 5 days ago
8a3733b
[FEAT]: Ahora todos los archivos registran logs
j1rd2 committed 5 days ago
38119c3
[FEAT]: Ahora ya se registran logs en AWS
j1rd2 committed 5 days ago
a97156d



PR's aprobados:

- PR-5: Models V1 test/Jesus
- PR-7: [FEAT]: Metricas agregadas a modelo de comportamiento
- PR-12: Feat: Matriz de confusion agregada
- PR-14: Jesus/feat add logs
- PR-17: Jesus/add metrics lr
- PR-19: [FEAT]: Modelos v1.0 comportamiento y sanidad pasados a .py

7.5. Luis Adrián Uribe Cruz

Esta concentración en particular fue el motivo exacto por el que entré a esta carrera en primer lugar. no sólo la inteligencia artificial, sino en sí internet, computadoras, programación, todos esos términos que los medios de comunicación suelen vender como varitas mágicas y mundos de potencial ilimitado. Entré entonces con entusiasmo infantil, y aún tras acabar todo el semestre y ver que un montón de matrices están detrás de toda la magia, sigo viendo todo este mundo con la misma emoción.

No sólo que ya comprendo qué hay dentro de estas cajas negras, desde sus componentes hasta la matemática detrás de ellos, sino que ahora también comienzo a tener el conocimiento de las herramientas que me permiten poder armar mis propias "cajas mágicas". Total, si ya había entrado a la carrera sin saber ni qué era un lenguaje de programación, venir completamente en blanco a esta concentración no sería muy diferente.

Al ver en los primeros semestres la forma más básica de aprendizaje automático, la regresión lineal, ya me daba una ligera idea de qué podría esperar en este semestre, lo principal fue un gran desagrado a la estadística, la cual percibía difícil de comprender y quizá no tan importante para lo que querría hacer. Primer día de concentración, la estadística era la columna vertebral de la IA.
Por lo que empezando con estadística, fue un apoyo enorme que el curso empezará desde lo más básico de nuevo. Con clases más largas y especializadas, al fin pude ir comprendiendo los

mecanismos por los que se entrelaza a sí misma y a otras materias, como la que aquí compete, y pude ir dejando atrás la aversión. A pesar de que todo el tema de IA venía mucho más en blanco que la estadística, este fue el tema en el que sentí que ví el mayor aprendizaje a pesar de no conseguir los mejores resultados. Aprendí a analizar datos, a no dejarse llevar por primeras conclusiones y buscar apoyo a afirmaciones, cómo ver distintas situaciones desde el punto de vista estadístico para un mejor entendimiento y, por supuesto, modelos de aprendizaje automático que no sólo funcionaran y ya, sino que tuviera bases sólidas para decir por qué funciona y qué es lo que explica, donde mi cúspide fue en el trabajo final, creación de modelos sARIMAX para el PIB mexicano, donde pude aplicar varias de las técnicas aprendidas para obtener un modelo altamente preciso, explicable y útil para defender mis posturas.

En segundo lugar estaría el Big Data y la computación en la nube. Big Data es un reto de mayor nivel para estadística, pues se deben crear sistemas que además de fiables sean rápidos y escalables. Al tener que manejar volúmenes colosales y dinámicos de información, hay que aprovechar muy bien cada recurso disponible, pues querer depurar a prueba y error se paga con tiempo. Por ejemplo, para mi proyecto de esta materia, puede crear un modelo de Bosque Aleatorio para predecir el precio de materias primas en la India antes de poder conocer el modelo sARIMA. Allí, un error humano tan sencillo como no configurar bien un hiper parámetro para que el modelo pudiera abarcar toda la cardinalidad de mi información me costó en varias horas de entrenamiento que fracasaron al sobrepasar este límite. De este mismo proyecto aprendí a confiar en la intuición y probar antes de confiar ciegamente en métricas. Evaluando el modelo por RMSE y R2, los números arrojaban un modelo tan malo que era incluso peor que adivinar el precio por medio del promedio y ya, pero al aplicar el modelo a los datos reales, se veía un acercamiento bastante notorio con los precios correctos donde las métricas se iban empujadas por la presencia de outliers. Al graficar y ver que efectivamente modelo y datos reales seguían la misma tendencia, pude demostrar la efectividad del modelo y de mi aprendizaje.

Por la parte de computación en nube, no tuve la oportunidad de tener un aprendizaje tan profundo como con Big Data, sin embargo, puede crear bases y conocer conceptos que me permitirán realizar mis propias investigaciones y aprender todo aquello que aquí no me terminó de quedar claro. El curso de AWS fue vital, pues te permitía aprender sobre la plataforma real y con ejemplos aplicables en lugar de sólo leer teoría y contestar cuestionarios. Pude aprender los distintos servicios que Amazon ofrece, la forma de crearles robustez, alternativas y cómo decidir si conviene o no usar estos.
Antes de hablar del elefante en la habitación quisiera repasar la condición que definió el semestre completo: la computadora. Mi computadora personal sufrió de defectos en la placa madre que la dejaron en una reparación dolorosamente lenta hasta que finalmente fue dada por muerta. Eso provocó un gran impacto, el tener que tomar una clase de computadoras sin una. Aunque podía mantenerme gracias a sitios como Colab a los que puedo acceder desde mi celular, tuve que adaptarme a estar al día con todo haciendo lo mínimo que se pidiera. No había oportunidad de avanzar con tiempo, conforme terminaba algo a último momento, ya tenía que empezar lo siguiente. Mi aprendizaje no fue “gracias a” sino “a pesar de” y por eso tengo el sentimiento general de no haber aprendido lo suficiente o de haber hecho lo suficiente. Fue hasta la quinta semana que pude tener acceso a una computadora: una muy vieja e incompatible con mucho, pero con la suficiente memoria para compensarlo con múltiples programas y pestañas simultáneas. Con esta computadora prestada, aunque no podía instalar cosas cómodamente por tener que dejarla en cualquier momento, que no pasó, me permitió sobrevivir el semestre y cumplir con casi todos los trabajos.
Finalmente, el tema principal de la concentración. Totalmente maravillado, y abrumado, por la gran variedad de tipos y arquitecturas disponibles, donde se pueden adaptar a problemas que normalmente

no podrían o se puede también optar por la adecuada cuando sea posible, gracias a eso se me ocurrió reemplazar un sARIMA con un bosque aleatorio, o por el contrario, decidir para el proyecto de las vacas que una red neuronal no era superior a dicho bosque. Por supuesto, no pude aprender absolutamente todo de todos, pero tener nombres e ideas es suficiente para comenzar a explorar y armar bloque a bloque para comprenderlos. Pongo especial atención en los autoencoders, en los que basé mi proyecto de esta materia. Cuando recién los aprendí, no les hallaba utilidad, o sea, solamente sirven para sacar lo mismo que les entra. Claro, resumido así no suenan tan impresionantes, pero esa capacidad hace que sea una arquitectura ampliamente usada en una gran variedad de tareas y la base de algunas de las más usadas, como U-Net de la cual nacieron múltiples otras aplicaciones.

Este proyecto se centra en la tarea de Inpainting, la recreación de zonas totalmente ocultas de una imagen a través del contexto extraíble de su alrededor y de la imágen completa. Supuse que algo que es capaz de recrear su entrada tendría que ser el indicado para la labor, y la intuición no me falló. Al leer los múltiples papers de gente que ya lo había hecho, casi todos utilizaban de base los autoencoders, y U-Net, cada uno con un enfoque distinto para poder reunir el contexto global y local. Para este proyecto, decidí usar esa desventaja de la computadora en mi enfoque. Esas arquitecturas usaban técnicas y capas muy complejas que daban excelentes resultados, un proyecto en el que el poco tiempo y mal hardware no podría acercarse aunque copiara sus ideas. Entonces pensé en que mi proyecto sería justamente el mínimo esfuerzo: "¿Qué tan lejos se puede empujar una arquitectura sencilla, con dataset limitado y con pocos recursos computacionales?" Pues estos grandes modelos operan con sets de millones de imágenes y grandes magnitudes de poder en GPUs, qué tan cerca se podría llegar.

Por supuesto, los resultados no son visualmente sorprendentes: poseen aliasing, baja resolución y un pequeño detalle por el cuál era imposible probar el modelo con imágenes más grandes, pero los resultados sí eran bastante buenos. El modelo consiguió la capacidad de obtener el contexto de la imagen incluso con grandes regiones eliminadas por la máscara y poder recrearlas de forma coherente. Es decir, a pesar de que se notaran los problemas al verla, el parche generado era totalmente creíble y se integraba bien con la imagen a pesar de. Se pudo demostrar lo lejos que se puede llegar sin los recursos más óptimos o las técnicas más revolucionarias.

Pero de IA no sólo se aprendió a hacerla, también a usarla. Aprendimos de agentes, cómo darle herramientas a los modelos para realizar cosas por su cuenta, pero también de los riesgos, el cómo la IA aún no es perfecta y no deberíamos confiar ciegamente en ella. A usarla responsablemente, con los profesores siempre haciendo hincapié a no abusar de ella y comprender siempre lo que entrega para poder comprender también lo bueno y lo malo. Al final del día, es una herramienta: es tan buena o mala como el uso que se le dé.

Hubo una sección enfocada en procesamiento de lenguaje natural, incluso se tuvo la oportunidad de entrenar un LLM de pequeña escala, aunque por el nombre suene contradictorio. Lamentablemente aquí no tuve la oportunidad de involucrarme en tan amplia medida por tener que estar dedicando mis limitados recursos en los proyectos, y de por sí me estaba costando comprender bien desde la base, los embeddings, pero como ya he reiterado varias veces en este escrito, las bases son lo que me permitirán explorar por mi cuenta y reponer todo lo que aquí no pude aprovechar en su totalidad.
Por último, mis aportaciones al modelo. CRISP-DM fue mi primer acercamiento a una metodología establecida de trabajo, y me costó bastante adaptarme a un trabajo estructurado. Sobre todo, cuando en el primer documento, casi todo lo que yo escribí estuvo rotundamente mal, me dió pena seguir contribuyendo así y sentía la inseguridad de estar más estorbando que ayudando. Eso y el tiempo que

pase atrasado y recuperándome me mantuvieron apartados del proyecto y tengo una inseguridad aún mayor de no haber contribuido lo suficiente. Por eso, aunque sí tengo trabajo establecido y evidenciado, muchas de mis aportes fueron más informales, como ideas, correcciones o añadidos en los documentos en secciones que ya habían sido redactados por otros y que por ello mi nombre no podía aparecer allí, especialmente en lo relacionado a los modelos donde podía apoyar explicando tipos, métricas y lo relacionado. Mi mayor aporte es este documento, el reporte final, y uno de los modelos candidatos para la solución, aunque finalmente se haya descartado.

En conclusión, estoy orgulloso de haber podido llegar al cierre de este semestre y muy emocionado por todo lo que he aprendido y estoy por aprender. Mi opinión de la estadística ha cambiado radicalmente y creo que poder dedicarme a la IA y datos no me parece tan mal después de todo. Solamente me queda ese pequeño sabor amargo de lo más que pudo ser de tener un mejor equipo y poder ver el pizarrón, pues tampoco llegaron en todo el semestre mis lentes para poder leer a más de medio metro de distancia. Finalizo con lo mismo con lo que comencé: A pesar de todo lo que esperaba, aprendí y me faltó, salgo contento, y con la misma emoción infantil de poder "jugar" con todo lo que el mundo de la IA tiene para ofrecer.

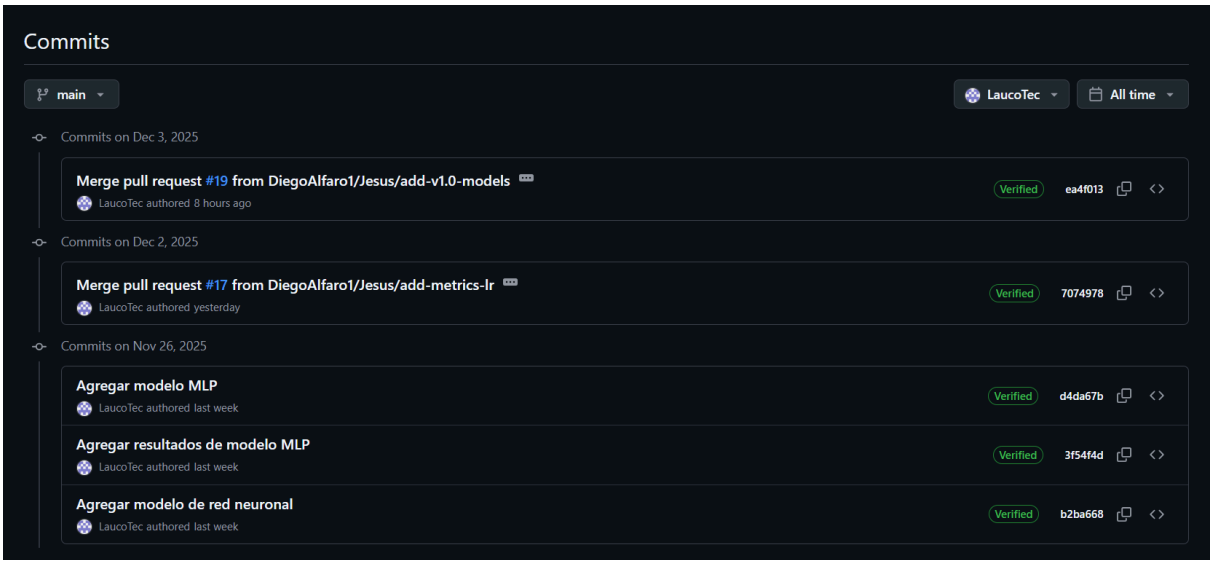
Evidencias:

Actividad	Personas	Estado
Observaciones iniciales	Luis Adrián Uribe Cruz	Revisado
Valores faltantes	Luis Adrián Uribe Cruz	Revisado
Duplicados	Luis Adrián Uribe Cruz	Revisado
Inconsistencias	Luis Adrián Uribe Cruz	Revisado
Outliers	Luis Adrián Uribe Cruz	Revisado
Precisión y coherencia	Luis Adrián Uribe Cruz	Revisado
Completitud	Luis Adrián Uribe Cruz	Revisado
Accesibilidad	Luis Adrián Uribe Cruz	Revisado

Actividad	Personas	Estado
Documentar datos adicionales que vamos a recolectar/solicitar e integrar para el problema	Luis Adrián Uribe ...	Listo para revisión

Entregable	Actividad	Personas	Estado
Modelo Salud V2.1	Configuracion de parametros	Luis Adrián U...	Listo para revisión
Modelo Comportamiento MLP	Codificacion del modelo	Luis Adrián U...	Listo para revisión
Modelo Salud V3.0	Configuracion de parametros	Luis Adrián U...	Deprecado
Modelo Salud V3.1	Configuracion de parametros	Luis Adrián U...	Deprecado

Actividad	Personas	Estado
Reporte Final Profesores	Luis Adrián Uribe Cruz	En progreso



Técnicamente faltaría el PR de mi rama MLP a main, pero tuve problemas con git y alguien más me hizo el favor.

7.6. Mauricio Anguiano Juarez

Cuando empecé este 7mo semestre en agosto contaba con un conocimiento intermedio/avanzado de Python y aunque no estaba convencido al cien por ciento de que esta concentración era para mí, hoy a varios meses después veo hacia atrás y me puedo dar cuenta que ahora cuento con conocimiento fundamental para mi vida laboral futura gracias a varias lecciones que aprendí tanto de código como de trabajo en equipo. Completando varios “mini proyectos”, entrenando varios modelos diferentes y siendo capaz de procesar grandes cantidades de información de datos reales, sin embargo todos esos números no son los que cuentan, lo que realmente cuenta de todo esto es que mi forma de pensar cambió totalmente.

Uno de los proyectos que más considero que me enseñó a lo largo de este semestre fue el clasificador de imagenes de con ResNet18 en el cuál utilicé Transfer Learning, apliqué varias técnicas que investigué para una normalización robusta, mixup, label smoothing, dropout, TTA, etc. Logré tener un 99.98% de accuracy en el test set, teniendo así solamente un error de 6,321 imágenes, fue ahí cuando probé imágenes externas descargadas de Google y gameplays de Youtube cuando el modeló falló, imágenes con shaders modernos eran clasificados como “NO MINECRAFT”, ahí fue cuando me pude dar cuenta que mi modelo había podido clasificar correctamente las características de casi todas las imágenes de los datasets utilizados, sin embargo no caracteristicas de Minecraft en realidad, dando como resultado que el siempre se debe de verificar los resultados que uno tiene, ya que un 99% no significa nada si falla con datos reales. Este proyecto fue en este periodo, pero durante el

primer periodo tuve que aprender y entender de verdad cómo es que funcionan los modelos, implementando una regresión logística desde cero para dar una predicción basada en datos si es que realmente un juego de League of Legends se puede decidir en los primeros diez minutos de partida, el entender lo que estaba haciendo me llevó a trabajar ahora si con librerías las cuales me facilitaron la vida, programando un random forest con scikit learn, pude apreciar profundamente lo que las bibliotecas hacen por nosotros pero también entendiendo realmente qué es lo que hay detrás de cada modelo.

Para los temas de estadística el trabajar con datos reales sobre la economía me sacó de mi zona de confort, esto porque jamás había trabajado con este tipo de datos o documentos, trabajar con series temporales fue un reto, ya que el orden es importante, no puedes mezclar datos, la estacionariedad era un requisito, como lo era aprender sobre las pruebas de Dickey-Fuller, diferenciación estacional, etc. Pero lo que más me gustó de este módulo fue el usar el modelo ARIMAX donde utilizamos el PIB de USA como variable para predecir el PIB de México, confirmando cuantitativamente que tenemos una interdependencia económica con ellos, viendo que como se dice comúnmente "Cuando a Estados Unidos le da gripa a México le causa pulmonía", conectar las matemáticas con el mundo real fue satisfactorio.

Hablando de inicios de semestre, este semestre conocimos la metodología de trabajo CRISP DM, antes solamente era un acrónimo y un entendimiento de que así es como se tienen que trabajar los proyectos de minería de datos, pero al enfrentar el proyecto en el que nos asociamos con el CAETEC pude entender la razón por la cuál existe este tipo de metodología y la razón por la cuál es importante seguirla para así asegurar tener un proyecto exitoso, primero como buen programador y con algunas malas prácticas todavía quería empezar a programar lo que yo creía que teníamos que hacer, pero al no entender la metodología al principio sabía que esto tomaba tiempo, el tener que darle el tiempo al *business understanding* para poder entender lo que realmente necesitaba nuestro cliente. Después de 2 semanas entendiendo el proyecto tuvimos que entender los datos que nos estaban proporcionando, entrando a la etapa del *data understanding*, pasamos una semana entendiendo que es lo que las columnas que tenían los csv mandados por el CAETEC tenían, descubriendo que es lo que varios indicadores eran, familiarizarnos con el lenguaje con el que tendríamos que trabajar durante este proyecto, aquí fue donde nos empezábamos a preguntar que modelo podríamos usar, que hacerle a los datos para que nos sean útiles, estas respuestas empezaban a llegar conforme fuimos pasando de fase, como bien lo dice la siguiente fase el *data preparation* en esta fase nos dedicamos a limpiar los datos, imputar en dado caso de ser necesario, transformarlos y crear features, la importancia de esta fase es crucial ya que en dado caso de no hacerlo bien, ninguno de los modelos que implementamos a lo largo del semestre hubieran podido capturar los patrones de salud y comportamiento que queríamos para poder darle una solución a nuestro cliente. Cuando llegamos a la fase del *modelling* experimentamos algo que ya habíamos sentido a lo largo del semestre, no saber que modelos usar, cuáles eran los que serían los más adecuados para este tipo de proyectos y cuantas iteraciones tendríamos que hacer para que se puedan lograr nuestros objetivos. Para todos nuestros modelos documentamos hiperparametros, resultados, tiempo de entrenamiento,

razones para continuar o descartar. Esto evitó que nos perdiéramos en el caos de no saber qué hacer y tener métricas exactas de qué modelo sería el "ganador", sin embargo nada del modelo funciona si es que en la etapa de *evaluation* no cuentas con lo necesario para respaldar que esto es lo correcto, para esto presentamos matrices de confusión para que sepan que nuestro modelo puede llegar a detectar que de 10 vacas inquietas que hay, el modelo puede detectar 7 y pierde 3, para esta última semana de trabajo terminamos el *deployment,* para poder entregar una aplicación de escritorio la cual sea capaz de utilizar nuestros modelos escogidos para dar una predicción, entendemos que esto tiene que correr en los dispositivos de los socios por lo que todo está subido en AWS S3, para que no sea necesario de que este proyecto cuente con pedazos hardcodeados dando un modelo muerto que no sea reproducible. CRISP-DM me enseñó al igual que en semestre anteriores muchas otras metodologías, que se tienen que seguir las fases y que no se pueden hacer cosas sin haber terminado algo antes porque eso desencadena al caos y a un proyecto de alto riesgo de ser terminado.

Para mi participación en el equipo, con toda honestidad tengo que decir que no tuve una participación muy amplía dentro del código, ya que solamente participé en la creación del modelo base de comportamiento implementando un modelo de regresión logística. Sin embargo esto no significa que no haya trabajado o apoyado a mis compañeros en las actividades del equipo, mi mayor contribución fue en la documentación de todo el proyecto, al inicio me sentía extraño ya que veía a todos programando los modelos del proyecto y tenía la espina de no estar aportando lo suficiente, pero conforme fui avanzando descubri que para un proyecto sea exitoso en su totalidad este debe de contar con una documentación clara y concisa. Invertí varias horas de mi tiempo detallando lo que hacían los modelos, interpretando gráficas, diciendo porque escogimos lo que escogimos, explicando con datos reales el porqué es que vamos como vamos. Aprendí que en un equipo, no todos tienen que escribir todo el código. Algunos son excelentes programadores de algoritmos complejos, descubrí que una de mis fortalezas es el hacer que el trabajo completo sea entendible para las personas que recibirán toda esta información.

Cuando empezó este semestre el machine learning para mí era una caja misteriosa, con un contenido que no podía entender todavía, hoy por hoy puedo decir que entiendo lo que contiene esa caja, no me considero un experto ni mucho menos porque sé que hay un universo entero de información allá afuera que no he conocido, pero ahora sé como aprender, como es que funcionan los algoritmos desde cero y que es lo necesario para seguir aprendiendo, este semestre me dio las herramientas necesarias que estaba buscando al inicio, saber cuál y saber diseñar un modelo capaz de ayudar a la empresa familiar en la que trabajo, facilitando la toma de decisiones y agilizando procesos que podrían tardarnos varios meses, en dado caso de tener otro proyecto de minería de datos sin duda alguna aplicaré CRISP-DM, ya sin verlo como una imposición académica, sino como una herramienta para simplificar mi trabajo cuando se me complica. También aprendí que para este tipo de proyectos no todos tienen que escribir código para aportar valor, aunque el sentimiento sea inadecuado o por el estilo, hay valor en poder contribuir a la estructura del proyecto, documentar las decisiones que hicimos, construir un puente entre código y documentación, que cada quien aporta desde

su fortaleza y que una documentación clara y estructurada sólida son igual de críticas que un algoritmo sofisticado.

Pero más allá de las tecnologías y metodologías, ahora veo la IA de una forma diferentes, al inicio creía que más capas, más épocas, más features, más parámetros, un mejor modelo y el accuracy era todo lo que importante, que todo tendría que tener el rendimiento que espero desde la primera iteración. Ahora sé que la IA cuenta con una matemática muy grande detrás, que se tiene que tener paciencia entrenando cada modelo, que tener más datos valen más que tener más complejidad, que la métrica correcta depende del contexto del socio con el que estamos trabajando. Ahora sé que fallar rápido, aprender e iterar es exactamente el proceso correcto.

Evidencia:

Plan Business Unsertandi...

Fase del proyecto	Entregable	Actividad	Personas	Estado
		Background	Mauricio Anguiano Juárez	Revisado
		Problemas a los que se enfrentan	Mauricio Anguiano Juárez	Revisado
		Identificar qué reportes tienen actualmente	Mauricio Anguiano Juárez	Revisado
		Inventario de recursos	Mauricio Anguiano Juárez	Revisado
		Riesgos y contingencias	Mauricio Anguiano Juárez	Revisado
		Corregir objetivos	Mauricio Anguiano Juárez	Revisado
		Terminar de identificar problemas antes de los objetivos del negocio	Mauricio Anguiano Juárez	Revisado

Plan Business Unsertandi...

Fase del proyecto	Entregable	Actividad	Personas	Estado
		Politica de privacidad y acceso	Mauricio Anguiano Juárez	Revisado
		Documento de politica de datos	Mauricio Anguiano Juárez	Revisado

Plan Data preparation

Fase del proyecto	Entregable	Actividad	Personas	Estado
		Descripcion del dataset	Mauricio Anguiano Juárez	Revisado
		Lista de datos que vamos a utilizar y por qué	Mauricio Anguiano Juárez	Revisado
		Documentar si se aplicó alguna normalización a los datos	Mauricio Anguiano Juárez	Revisado
		Documentar el formato que se le dió a cada dato	Mauricio Anguiano Juárez	Revisado
		Documentacion de datos calculados	Mauricio Anguiano Juárez Diego Isaac Fuentes Juvera	Revisado
		Documentar la justificación de la calificación de las vacas	Mauricio Anguiano Juárez	Revisado

Plan Modeling

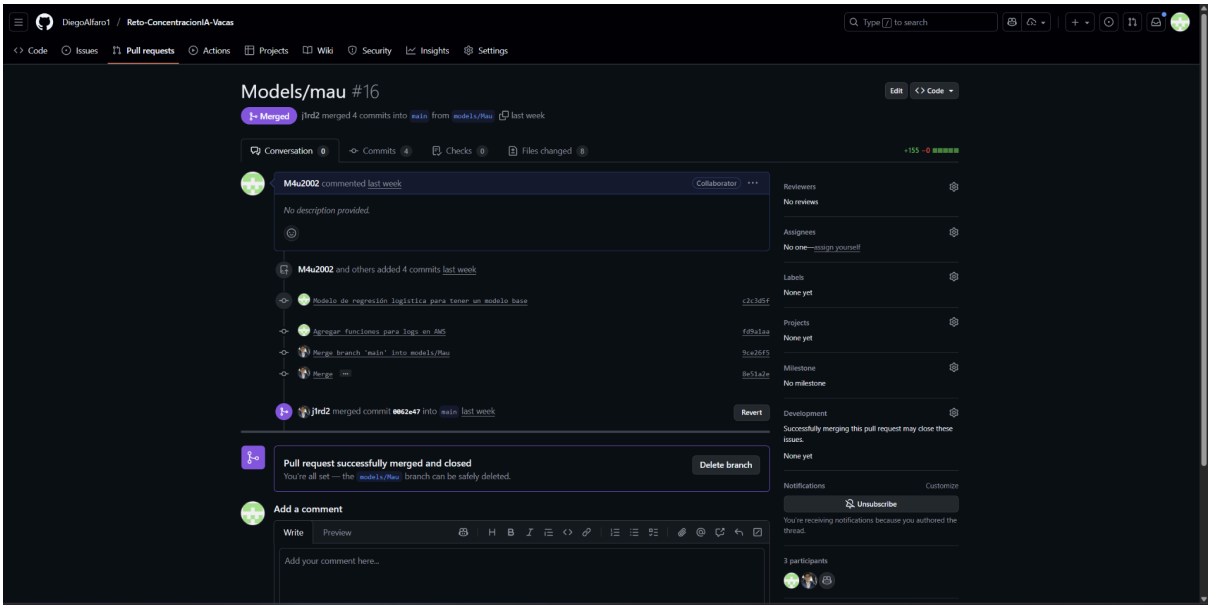
Fase del proyecto	Entregable	Actividad	Personas	Estado
	Modelo Comportamiento V1.0	Descripcion del modelo	Mauricio Anguiano Juárez	Revisado
	Modelo Salud V1.0	Descripcion del modelo	Mauricio Anguiano Juárez	Revisado
	Modelo Comportamienti base	Descripcion del modelo	Mauricio Anguiano Juárez	Revisado
	Modelo Comportamienti base	Configuracion de parametros	Mauricio Anguiano Juárez	Revisado
	Modelo Comportamienti base	Codificacion del modelo	Mauricio Anguiano Juárez	Revisado

Evaluation

Fase del proyecto	Entregable	Actividad	Personas	Estado
		Evaluacion de resultados de modelado	Mauricio Anguiano Juárez Daniel Queijeiro Albo	Listo para revisión
		Proceso de revision	Mauricio Anguiano Juárez Diego Alfaro	Listo para revisión
		Lista de trabajo futuro	Mauricio Anguiano Juárez	Listo para revisión
		Decisiones	Mauricio Anguiano Juárez	Listo para revisión

Deployment

Fase del proyecto	Entregable	Actividad	Personas	Estado
		Definir que métricas se mostrarán en la interfaz	Diego Isaac Fuentes Juvera Mauricio Anguiano Juárez	Listo para revisión
		Generar diseño de la aplicación en Figma	Diego Isaac Fuentes Juvera Mauricio Anguiano Juárez	Listo para revisión
		Generar manuales de usuario de la aplicación	Mauricio Anguiano Juárez Diego Isaac Fuentes Juvera	Listo para revisión



https://github.com/DiegoAlfaro1/Reto-ConcentracionIA-Vacas/pull/16