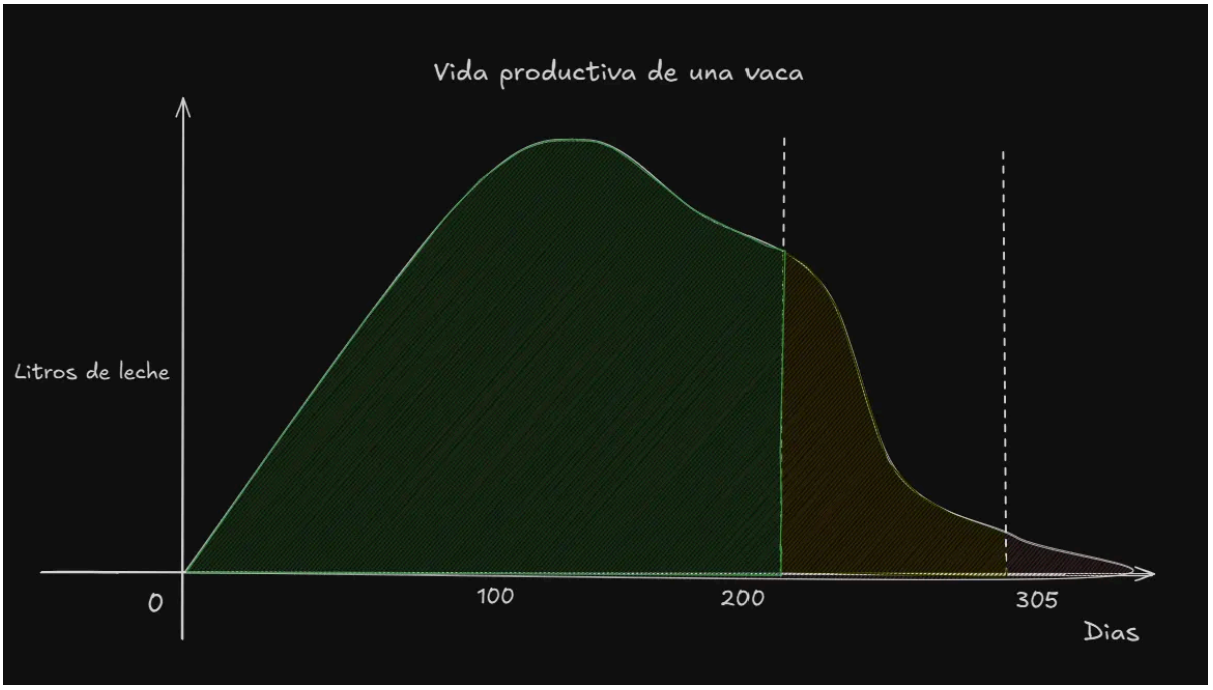
Modeling Technique (Tercera iteración)

Resumen

En la fase de modelado correspondiente a la tercera iteración del proyecto, se implementó un enfoque distinto al inicial, que se centraba en la estimación de la fecha de secado mediante series de tiempo o en cambios en la producción de leche.

Tras la revisión de problemas y sugerencias con la Dra. Guadalupe López Rendón, se redefinió la estrategia para incorporar información integral de todo el hato. El objetivo es replicar y automatizar los criterios que la Dra. utiliza para clasificar a cada vaca Holstein en las categorías de MANTENERSE EN PRODUCCIÓN, MONITOREO o SECADO.

En esta tercera iteración se implementaron cinco modelos distintos con el propósito de comparar su desempeño en la clasificación. Se incluyeron modelos de Machine Learning como Regresión Logística, Random Forest y XGBoost, así como enfoques más complejos de Deep Learning, específicamente una red MLP y TabNet.



1.1. Propuesta de la tercera iteración

Tomando en cuenta los resultados de las primeras 2 iteraciones se decidió abordar el problema desde otra perspectiva, es decir buscar otra manera en la que se pudiera

conseguir el mismo objetivo. Para la validación del mejor modelo para el problema se crearon diferentes propuestas de modelos que nos permitirán comparar los rendimientos y evaluar cual de ellos es el más adecuado para los objetivos originales.

Regresión Logística	Se seleccionó un modelo de clasificación basado en una técnica estadística para predecir la probabilidad de que ocurra un evento llamado `Logistic Regression` , este modelo se seleccionó como baseline, punto de referencia, para comparar el rendimiento de los diferentes modelos a evaluar.
Random Forest	Se seleccionó un modelo de clasificación supervisada basado en Random Forest, utilizando la implementación de `RandomForestClassifier` de scikit-learn. Esta técnica fue elegida debido a: Su capacidad para manejar relaciones no lineales entre variables productivas y reproductivas. Su buen desempeño en problemas de clasificación multiclase como el estado productivo (”En Producción”, “En Monitoreo”, “Previo a Secado”). Su capacidad para operar sin necesidad de una normalización estricta.
XGBoost	Se propone `XGBoost` utilizar como modelo principal debido a su capacidad para manejar relaciones no lineales, su robustez ante datos tabulares con variabilidad estructural y su excelente desempeño en contextos de desbalance moderado. Este permite capturar interacciones complejas entre variables fisiológicas, reproductivas y productivas del hato, lo que lo convierte en una alternativa altamente competitiva frente a modelos más simples.
MLP	Para corroborar los modelos de ML anteriores, se selecciono una arquitectura de `Deep Learning` con capas densas (MLP por sus siglas en ingles) para este problema multiclase. Esto ultimo se eligio al ser una arquitectura lo suficientemente robusta y sencilla para poder clasificar los tres estados de nuestras vacas.
TabNet	Se selección un modelo basado en la arquitectura ya existente de `TabNet`, en problemas de clasificación es muy común utilizar este tipo de estructuras de datos tabulares y capaz de aprender representaciones jerárquicas mediante atención secuencial sobre las features, este modelo busca no sólo mejorar la exactitud global, sino también ofrecer predicciones más útiles en la práctica **veterinaria**, donde detectar vacas “En Monitoreo” y “Previo a Secado” de manera oportuna es más crítico que maximizar únicamente la precisión en la clase mayoritaria “En Producción”.

1.2. Proceso de la tercera iteración

Tras los resultados poco satisfactorios de la iteración 1 (regresión) y la complejidad de la iteración 2 (LSTM), se replanteó el problema como una tarea de clasificación del estado reproductivo de cada vaca Holstein.

El nuevo objetivo se enfocó en clasificar correctamente si una vaca se encuentra:

- En Producción
- En Monitoreo (30 días antes del secado)
- Previo a Secado (7 días antes del secado)

Para esto se reconstruyó un dataset de global hato usando:

- Datos globales del hato
- Reglas fisiológicas de producción
- Sistema de data augmentation que garantiza la generalización
- Balanceo entre clases (cerca de 90, 660 registros finales)

1.3 Evolución desde la primera iteración

En la iteración 1 se usaron modelos de regresión para predecir DEL como variable continua. Random Forest explicó sólo 71% de varianza y no cumplió el objetivo de negocio. Se abandonó el enfoque por no captar bien los patrones.

En la iteración 2 se replanteó el problema como serie temporal en el último tercio de la lactancia. Se capturó temporalidad y tendencia, pero el sistema resultó difícil de validar y dependiente de curvas completas por vaca. Además, no contábamos con datos suficientes para el entrenamiento.

Finalmente, después de una asesoría final con los profesores y expertos en el área, se vio que el enfoque dado en la segunda iteración era inviable. El manejo de series de tiempo para ver los decrementos en producción era una tarea ardua y compleja por lo que se decidió cambiar este enfoque.

A partir de ello, la selección de estos modelos y este nuevo enfoque se basa completamente en las recomendaciones y sugerencias que **la Dra. Guadalupe López Rendón nos dio.**

Donde se redefinió la estrategia para incorporar información integral de todo el hato yendo en una segunda ocasión al CAETEC. Teniendo como objetivo principal replicar y automatizar los criterios que la Dra. utiliza para clasificar a cada vaca Holstein en las categorías de MANTENERSE EN PRODUCCIÓN, MONITOREO o SECADO.

1.4. Criterios de selección

Los criterios de selección de los modelos que se usaron durante esta iteración están basados en arquitecturas que puedan servir para este problema de clasificación multiclase que planteamos. Donde se consideraron cosas como:

- Sencillez vs complejidad

- Tamaño del dataset y dataset final después de Data Augmentation
- Resultados de las métricas de evaluación

Regresión Logística	Este modelo se determinó como el modelo base debido a que al ser de los modelos más sencillos de implementar para predicciones se podrá determinar si el problema que se está abordando requiere arquitecturas más complejas o se puede resolver con menores recursos.
Random Forest	El modelo puede clasificar correctamente el estado productivo para apoyar decisiones de secado. Random Forest aprovecha muy bien features como producción 7 días, DEL, días para el parto, etc.
XGBoost	XGBoost se elige porque ofrece un buen balance entre rendimiento global y detección de clases críticas, mostrando un F1 macro sólido y alto recall donde más importa. Es estable, rápido de entrenar y fácil de interpretar, lo que permite validar sus predicciones con el criterio veterinario y usarlo operativamente sin complicaciones.
MLP	Al tener un mayor número de instancias el uso de modelos de DL puede buscar relaciones que modelos tradicionales de ML no pueden. De esta forma, nos decantamos por esta arquitectura que puede clasificar correctamente el estado productivo para apoyar decisiones de secado.
TabNET	A diferencia de los modelos anteriores, lo que hace TabNet permite aprender representaciones jerárquicas mediante máscaras de atención que seleccionen dinámicamente las características más relevantes para cada decisión o que resulta especialmente adecuado para datos tabulares complejos como los del sistema lechero. Esta propuesta integra un preprocesamiento más refinado, con imputación específica por tipo de variable, codificación consistente de atributos categóricos y escalado normalizado de variables numéricas.

1.5 Supuestos generales

Se asume que todos los registros del dataset tanto reales como aumentados cumplen con los rangos fisiológicos establecidos para cada estado productivo.

El modelo desarrollado en la tercera iteración tiene como finalidad automatizar la clasificación del estado productivo de cada vaca, es decir, determinar de manera objetiva si una vaca se encuentra en Producción, En Monitoreo o Previo al Secado sin depender de supervisión manual ni interpretación por parte del personal.

La automatización proporciona alertas tempranas, cosa que no era posible con métodos anteriores (Iteración 1: regresión, Iteración 2: LSTM demasiado dependiente de series completas).