

**Pontificia Universidad Católica del Ecuador**

**Facultad De Ingeniería**



**TEMA:**

MODELO PREDICTIVO APLICANDO ALGORITMOS DE MACHINE LEARNING  
PARA LA PRODUCCIÓN LECHERA EN LA HACIENDA EL PRADO, DEL INSTITUTO  
AGROPECUARIO SUPERIOR ANDINO (IASA)

**AUTOR:**

JORGE ISAAC ORDOÑEZ MERINO

TRABAJO PREVIA A LA OBTENCIÓN DEL TÍTULO DE MAGISTER EN SISTEMAS DE  
INFORMACIÓN MENCIÓN DATA SCIENCE

**Quito, Enero – 2023**

## DEDICATORIA

---

Con afecto, respeto y los valores más nobles de mi sentir dedico este proyecto de investigación a mi Padre, quien fue el artífice y ejemplo de superación personal y profesional, eres para mí el mejor maestro de todos los tiempos, siempre estaré agradecido por tus enseñanzas.

## **AGRADECIMIENTO**

---

Agradezco a Dios por brindarme la oportunidad de vivir, experimentar, soñar y cumplir las metas deseadas, la fe y la perseverancia me han ayudado en este camino que aun empieza y está lejos de terminar; a mis maestros por saber inculcar los conocimientos y experiencias que serán los KPI en los retos del futuro y en especial a mi tutor quien me brindó la oportunidad de demostrar mi talento.

## RESUMEN

---

El presente trabajo tiene como fin el desarrollo de un aplicativo web aplicando algoritmos de Machine Learning para la predicción de la producción lechera, de tal manera que pueda ser utilizado en programas de mejoramiento genético y en la toma de decisiones de los administradores de la ganadería del Instituto Agropecuario Superior Andino IASA. La ganadería mencionada se ubica en la Sierra Norte del Ecuador, tiene un sistema de producción de pastoreo, y cuenta con ejemplares de las razas Holstein y las del producto de la cruce Montbéliarde x Holstein. El sistema en pastoreo es influenciado por variables exógenas al sistema, que son determinadas por factores agroecológicos y de manejo de praderas, cuya influencia tiene variación local. Se utilizó la metodología CRISP-DM para sistematizar la información clave y el flujo de actividades para el desarrollo y posterior evaluación del modelo predictivo. El modelo se desarrolló con Python con el uso de las librerías pandas y scikit-learn. El algoritmo seleccionado para el desarrollo del modelo fue el de regresión lineal múltiple, con 17 variables predictoras, que fueron tomadas entre los factores productivos, reproductivos y ambientales. El modelo obtenido es útil para la predicción de la producción lechera, siendo validado por los indicadores  $R^2$ -ajustado de 0.80 y razón entre la raíz del error cuadrático y la media de producción total de 0.11. Con el modelo obtenido se desarrolló el aplicativo web utilizando la plataforma Stramlit, de tal manera que el acceso al mismo sea público.

## ÍNDICE

---

ÍNDICE DE GRÁFICOS .....	VI
ÍNDICE DE TABLAS .....	VIII
CAPÍTULO I: INTRODUCCIÓN .....	1
1. INTRODUCCIÓN .....	1
1.1 Antecedentes.....	1
1.2 Planteamiento del problema .....	2
1.3 Justificación .....	3
1.4 Objetivos.....	4
1.4.1 Objetivo General.....	4
1.4.2 Objetivos Específicos.....	4
CAPÍTULO II: FUNDAMENTACIÓN TEÓRICA .....	5
2. MARCO TEÓRICO Y CONCEPTUAL .....	5
2.1 El proceso de producción lechera.....	5
2.2 Sistemas de producción lechera .....	5
2.2.1 Sistema de producción en pastoreo.....	7
2.2.2 Sistema de producción en estabulación.....	7
2.2.3 Sistema de producción en semi estabulación .....	8
2.3 Factores determinantes en la producción lechera.....	8

2.3.1	Factores genéticos .....	9
2.3.2	Factores ambientales.....	9
2.4	Mejoramiento genético en el ganado lechero.....	10
2.4.1	Breeding value y las evaluaciones genéticas tradicionales .....	10
2.4.2	Evaluación genómica.....	11
2.5	Parámetros de la producción lechera.....	11
2.5.1	Parámetros productivos .....	12
2.5.2	Parámetros ambientales .....	12
2.5.3	Parámetros reproductivos .....	14
2.6	Machine Learning .....	15
2.6.1	Aprendizaje Supervisado .....	16
2.6.2	Aprendizaje no supervisado.....	17
2.7	El modelo de regresión lineal.....	17
2.7.1	Fundamentos estadísticos y matemáticos .....	17
2.7.2	Fundamentos de programación en R.....	17
2.7.3	Fundamentos de programación en Phyton .....	18
2.8	Cross- validation .....	19
2.9	Publicación web de modelos de Machine Learning.....	19
CAPÍTULO III: METODOLOGÍA Y TÉCNICAS .....		21
3.	Desarrollo del modelo de predicción de la producción lechera mediante CRISP-DM	
	21	
3.1	Comprensión del negocio .....	21

3.1.1	Perspectiva del negocio.....	21
3.1.2	Objetivos del negocio.....	22
3.1.3	Evaluación de la situación .....	22
3.1.4	Determinación de los objetivos de la modelación.....	24
3.1.5	Plan de proyecto.....	24
3.2	Comprensión de los datos .....	25
3.2.1	Recolección inicial de los datos .....	25
3.2.2	Descripción de los datos.....	25
3.2.3	Exploración de los datos.....	25
3.3	Preparación de los datos .....	26
3.3.1	Selección de datos .....	27
3.3.2	Limpieza de datos.....	32
3.3.3	Construcción de nuevos datos.....	33
3.3.4	Formateo de los datos .....	33
3.4	Modelado.....	34
3.4.1	Selección de técnica.....	34
3.4.2	Construcción del modelo .....	34
3.4.3	Evaluación del modelo.....	36
3.5	Evaluación .....	37
3.5.1	Evaluación de los resultados .....	37
3.5.2	Determinación de pasos a seguir.....	38

3.6 Implementación .....	38
CAPÍTULO IV: RESULTADOS Y DISCUSIÓN .....	40
4.1 Desarrollo y prueba de aplicativo web .....	40
4.2 Discusión .....	44
CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES .....	47
5.1 Conclusiones .....	47
5.2 Recomendaciones .....	47
BIBLIOGRAFÍA .....	49
ANEXOS .....	52

## ÍNDICE DE GRÁFICOS

<b>Gráfico 1.</b> Factores que intervienen en la producción lechera. Autor: Tentor, 2022 .....	6
<b>Gráfico 2</b> Ciclo reproductivo del ganado bovino en condiciones de manejo reproductivo ideales. Autor: Buchanan,2022.....	15
<b>Gráfico 3</b> Toma de captura de una muestra de la base de datos. Autor: Jorge Ordoñez, 2022 .....	26
<b>Gráfico 4</b> Salida de Jupiter notebook sobre valores nulos del conjunto de datos. Autor: Jorge Ordoñez, 2022.....	32
<b>Gráfico 5</b> Diagrama de calor de variables numéricas. Autor: Jorge Ordoñez, 2022 .....	35
<b>Gráfico 6.</b> Variables del modelo. Autor: Jorge Ordoñez, 2022 .....	36
<b>Gráfico 7</b> Salida de Jupiter notebook del modelo lineal. Autor: Jorge Ordoñez, 2022 .....	36
<b>Gráfico 8.</b> Evaluación estadística del modelo. Autor: Jorge Ordoñez, 2022 .....	37
<b>Gráfico 9</b> Métricas de evaluación del modelo. Autor: Jorge Ordoñez, 2022 .....	37
<b>Gráfico 10.</b> Gráfico de ajuste lineal del modelo entre valores predichos y reales. Autor: Jorge Ordoñez, 2022.....	38
<b>Gráfico 11.</b> Módulo de subida de archivos para actualización de registros en Git Hub Autor: Jorge Ordoñez, 2022.....	39
<b>Gráfico 12.</b> Captura de pantalla del encabezado del aplicativo web. Autor: Jorge Ordoñez, 2022 .....	40
<b>Gráfico 13.</b> Captura de pantalla de variables del factor genético. Autor: Jorge Ordoñez, 2022 .....	41
<b>Gráfico 14.</b> Captura de pantalla de sección de selección de variables de producción del aplicativo web. Autor: Jorge Ordoñez, 2022 .....	41
<b>Gráfico 15.</b> Captura de pantalla de sección de selección de variables de reproducción del aplicativo web. Autor: Jorge Ordoñez, 2022 .....	42

<b>Gráfico 16.</b> Captura de pantalla de sección de selección de variables agroecológicas del aplicativo web. Autor: Jorge Ordoñez, 2022 .....	42
<b>Gráfico 17.</b> Captura de pantalla de sección de resultado del modelo predictivo del aplicativo web. Autor: Jorge Ordoñez, 2022 .....	43
<b>Gráfico 18.</b> Captura de pantalla del código de la aplicación (back-end). Autor: Jorge Ordoñez, 2022.....	44

## ÍNDICE DE TABLAS

<b>Tabla 1</b> Personal del proyecto. Autor: Jorge Ordoñez, 2022 .....	22
<b>Tabla 2.</b> Equipos requeridos en el proyecto. Autor: Jorge Ordoñez,2022 .....	23
<b>Tabla 3</b> Riesgo y contingencias del proyecto. Autor: Jorge Ordoñez,2022 .....	23
<b>Tabla 4</b> Costos del proyecto. Autor: Jorge Ordoñez,2022 .....	23
<b>Tabla 5</b> Plan del proyecto. Autor: Jorge Ordoñez,2022 .....	24
<b>Tabla 6.</b> Variables para el modelo predictivo de la producción. Autor: Jorge Ordoñez,2022	27
<b>Tabla 7.</b> Descripción de la variable raza. Autor: Jorge Ordoñez, 2022 .....	28
<b>Tabla 8</b> Identificación del tipo de dato de las variables. Autor: Jorge Ordoñez,2022 .....	31
<b>Tabla 9</b> Identificación de nuevas variables. Autor: Jorge Ordoñez, 2022 .....	34

## CAPÍTULO I: INTRODUCCIÓN

---

### 1. INTRODUCCIÓN

#### 1.1 Antecedentes

En Ecuador la región con mayor producción lechera es la Sierra, donde se tiene un promedio por vaca de 7.7 litros al día (INEC, 2021). Sin embargo, esta producción es muy inferior a la de países con condiciones climáticas y sistemas productivos similares como Uruguay y Nueva Zelanda, que tienen producciones promedio de 18 y 25 litros por vaca al día respectivamente (DairyZn, 2020). Los factores detrás del éxito en los sistemas productivos de ambos países radican en la toma de decisiones en base a datos.

Los factores que influyen en la producción lechera se pueden atribuir a dos grandes grupos: la genética y el medio ambiente, siendo el más importante el componente ambiental al que se le atribuye un 70% de la variación observada. Adicionalmente, los factores relacionados con la genética tienen una serie de interacciones genéticas y ambientales, lo que hace que cada sistema de producción tenga expresiones diferentes del factor genético (Uribe, 2012).

Actualmente, se usan modelos predictivos con el fin de realizar mejoramiento genético en cada país, sin embargo, en Ecuador el avance en el tema es bastante limitado, lo que hace que la mayoría de las granjas comerciales carezca de un programa de selección animal, a esto se suma la falta de planificación y control, e incluso un desconocimiento de la utilidad del uso de registros. Por todo lo anterior, en nuestro medio existe la necesidad latente, en el sector ganadero, del uso de modelos predictivos locales que permitan mejorar la toma de decisiones en función de los datos para mejorar la productividad.

## **1.1 Planteamiento del problema**

La producción de leche de manera sostenible es el principal objetivo de una explotación de ganado lechero. A través de los principios del mejoramiento genético animal se ha venido mejorando la productividad por vaca a nivel mundial, a través del uso sistemático de registros de producción, diseños experimentales y herramientas de la biotecnología animal (Bourdon, 1999).

El enfoque utilizado en el mejoramiento animal, ciencia que busca mejorar la producción en función de la selección animal teniendo como base el intercambio genético, ha permitido obtener modelos de producción animal para predecir la mejora en producción o valor de cría en función de la genética de los progenitores, sin embargo, estos modelos tienen en cuenta únicamente las condiciones óptimas de producción en los países en los que fueron desarrollados, donde los animales pueden expresar todo su potencial genético, lo que no es común en los sistemas de pastoreo. Adicionalmente, las técnicas predictivas desarrolladas en este contexto requieren modelos experimentales bastante costosos que involucran varios años, y miles de animales para su validación (Wignans & Gengler, 2011).

En una granja comercial la predicción no solo puede usarse para la mejora genética de la producción sino también para la administración, lo que convierte a la predicción en una herramienta para tomar decisiones operativas en el corto plazo y establecer programas de mejora continua en el largo plazo.

Para lograr la máxima eficiencia operativa en la producción lechera es imperativo tener un modelo que permita relacionar la producción con las variables de importancia desde un punto de vista técnico y administrativo, no necesariamente ligadas al mejoramiento animal, de tal manera que se pueda optimizar la producción a través de un sistema de planificación y control.

## **1.2 Justificación**

Los propietarios o administradores de las granjas lecheras buscan obtener utilidades como en cualquier actividad empresarial. De manera particular, la producción lechera depende de la eficiencia operativa, que está en función de la eficiencia productiva de las vacas lecheras, y los factores que afectan la producción no son iguales ni tienen la misma influencia entre regiones y sistemas de producción. En este contexto es necesario contar con herramientas de predicción de la producción para lograr la optimización de los recursos a través de la administración basada en datos.

Al contar con un modelo local, usando técnicas de Machine Learning, que involucre variables de: sistema de producción, material genético, sistema de manejo, condiciones ambientales, sanidad y nutrición animal, entre otras; los administradores de la hacienda el Prado del Instituto Agropecuario Superior Andino IASA podrán conocer la influencia de estos factores en su sistema productivo particular.

Adicionalmente el modelo predictivo permitirá detectar problemas sanitarios, reproductivos y nutricionales de manera rápida y precisa en el corto plazo y en el largo plazo perfeccionar los programas de selección animal. Todo lo anterior enmarcado en los objetivos de la hacienda, y la eficiencia económica.

Un modelo de Machine Learning generado con datos locales se convierte en una alternativa de bajo costo, altamente eficiente y muy útil para los administradores a la hora de predecir la producción dentro de un hato lechero. A través del modelo se podrá establecer estándares de producción y llegar a gestionar de manera óptima el sistema productivo.

## **1.3 Objetivos**

### **1.3.1 Objetivo General**

Desarrollar un aplicativo web aplicando algoritmos de Machine Learning, para estimar la producción lechera por vaca en la hacienda el Prado, del Instituto Agropecuario Superior Andino IASA. Teniendo como data inicial los registros productivos de los semovientes, en los periodos comprendidos entre 2013 y 2021.

### **1.3.2 Objetivos Específicos**

- Realizar la revisión de la literatura que integre el uso de un modelo predictivo aplicando algoritmos de Machine Learning y la producción lechera.
- Analizar y evaluar la información de los modelos prototipo.
- Seleccionar el modelo predictivo adecuado en función del nivel de precisión de los resultados.
- Diseñar el aplicativo web, acoplado el modelo predictivo seleccionado para su uso público.

## CAPÍTULO II: FUNDAMENTACIÓN TEÓRICA

---

### 2. MARCO TEÓRICO Y CONCEPTUAL

#### 2.1 El proceso de producción lechera

El ganado lechero inicia la producción de leche luego del parto, este tipo de ganado es especializado y produce grandes cantidades de leche, más de lo que necesita para criar a su ternero, el manejo es completamente diferente a lo que ocurre con otros animales de interés zootécnico. El ganado lechero necesita comer alimento con alto contenido nutricional, para que estos nutrientes pasen al torrente sanguíneo y lleguen a la ubre. La producción láctea ocurre por impulso sensorial, y por este impulso el cerebro libera una hormona que actúa en la glándula mamaria lo que ocasiona el flujo de la leche. Una parte importante del proceso es el ambiente, cuando el animal está relajado y en confort liberará toda la leche almacenada. El flujo de leche se da únicamente cuando las vacas paren y dura generalmente entre 10 y 12 meses. Pero al final de este periodo la producción baja sustancialmente. Por eso de manera óptima se espera que la vaca se preñe prontamente y su lactancia dure alrededor de 305 días.

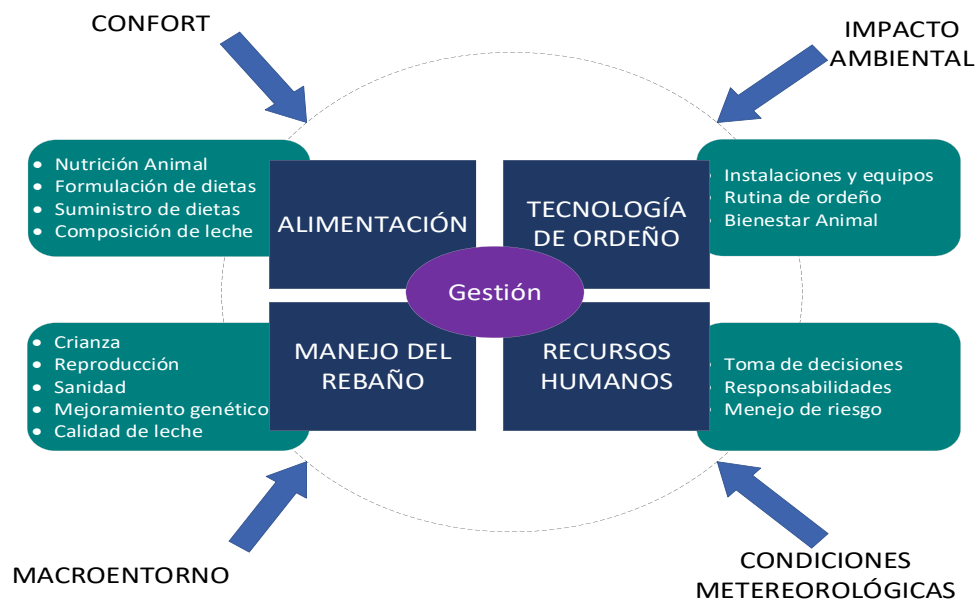
#### 2.2 Sistemas de producción lechera

Desde el punto de vista agropecuario, un sistema es un conjunto de factores que permite obtener un producto de origen animal o vegetal. Este conjunto de factores, por su acción e interacción originan un proceso dinámico y complejo. Adicionalmente, los sistemas agropecuarios tienen un resultado económico (McDaniel, 2011).

Una de las características más importantes de los sistemas agropecuarios es que deben ser sustentables, lo que se logra siempre y cuando el sistema sea rentable y estable en el tiempo. Entre los muchos factores que intervienen en el sistema de producción de leche podemos englobar la mayoría de ellos en cinco ejes fundamentales en la producción que son la

alimentación, el manejo del rebaño, la tecnología de ordeño, los recursos humanos y la gestión, estos ejes adicionalmente son condicionados por diferentes elementos exógenos al sistema como legislaciones relacionadas con bienestar animal e impacto ambiental y condiciones del macroentorno como fuerzas sociales, políticas económicas sectoriales, mercado mundial de la leche entre otras, un diagrama de los componentes del sistema lo podemos observar en la figura 1. Entre los ejes la alimentación es el que determina el tipo de sistema de manera amplia. Un punto importante de la elección del sistema óptimo es que este debe ser adecuado para los recursos y condiciones locales. Y por esta razón cada producción representa un sistema de producción único, desde un punto de vista específico. Hay que tomar en cuenta que los animales deben adaptarse al sistema de producción, y es por eso que el productor debe seleccionar inicialmente los animales que tengan mejor adaptación al sistema de producción que va a utilizar en la producción (Tentor, 2020).

**Gráfico 1.** Factores que intervienen en la producción lechera. Autor: Tentor, 2022



### **2.2.1 Sistema de producción en pastoreo**

El sistema de producción en pastoreo se caracteriza por manejar la alimentación del rebaño únicamente con pasto, el cual es cosechado por los animales en los potreros. Este sistema representa el más eficiente desde el punto de vista económico en la producción lechera, debido a que el costo de la alimentación llega a ser el más bajo. El éxito de este sistema depende del manejo del pasto como cultivo, de tal manera, que los animales puedan obtener alimento de calidad todo el año. Al tener un buen manejo del sistema se pueden llegar a producciones de entre 3,000 a 5,000 kg de pasto como cultivo por lactancia. Sin embargo, la alimentación basada únicamente en pasto imposibilita la expresión de todo el potencial genético de las vacas, sobre todo de las mejores productoras. El factor alimentación influye en los otros factores de la producción, por lo que al tener un sistema de pastoreo los productores, se requieren además menos cuidados sanitarios, menos inversión de capital y menos dependencia de mano de obra (Moran, 2005).

### **2.2.2 Sistema de producción en estabulación**

Este sistema se caracteriza por el confinamiento de los animales en establos y una alimentación basada en raciones preparadas en base a concentrados o balanceados. Bajo este sistema de producción, las vacas llegan a tener un mayor consumo de alimentos, por lo cual llegan a tener una producción superior a las vacas en el sistema de pastoreo que oscila entre el 5% y el 300% adicional, dependiendo del potencial genético de producción de las vacas. A pesar de ser un sistema más avanzado de producción, tiene como limitantes: los altos costos de inversión de capital, altos costos de alimentación, mayor dependencia de mano de obra y mayores cuidados sanitarios. Este sistema es eficiente desde el punto de vista energético de los animales, debido a que estos no deben moverse mucho para buscar el alimento, y la mayoría de la energía llega a aprovecharse en producción de leche. La producción estimada en estos sistemas oscila entre 7,000 y 10,000 kg de leche por lactancia.

Este sistema es muy común en países de cuatro estaciones y que además tengan costos bajos en los commodities necesarios para la producción de concentrados y balanceados como el maíz y la soja (Tentor, 2020).

### **2.2.3 Sistema de producción en semi estabulación**

La mayoría de los sistemas de producción no pertenece a ninguno de los grupos anteriores en su totalidad, por lo que se puede concluir que es una combinación entre ambos dependiendo de las características agroecológicas de cada ubicación geográfica. Todos estos sistemas intermedios, se conocen como los sistemas semi estabulados. El uso de los sistemas estabulados implica el uso del pastoreo en la alimentación complementado con el uso de concentrados, esto permite optimizar la producción lechera en cuanto a costos de producción, sin embargo, esto implica un conocimiento profundo tanto de la producción y nutrición animal cómo del cultivo de pastos. Los sistemas semi estabulados con un manejo deficiente en realidad son los menos productivos (Moran, 2005).

## **2.3 Factores determinantes en la producción lechera**

Dentro de los factores más importantes en la producción lechera tenemos la alimentación, seguido de los del manejo del rebaño que incluyen la crianza, la sanidad animal, la reproducción y el mejoramiento genético. Sin embargo, al estandarizar algunos factores y fijar un sistema productivo, el factor que determina el aumento de la producción es el mejoramiento genético animal, es así como desde esta perspectiva se pueden clasificar los factores de producción como ambientales y genéticos, siendo estos los relacionados con el mejoramiento animal cuando se han alcanzado niveles aceptables u óptimos en los factores ambientales.

### **2.3.1 Factores genéticos**

A través del mejoramiento genético se obtienen animales más productivos que sus ancestros, siendo un factor determinante, siempre y cuando se manejen de manera al menos aceptable los factores ambientales. Para obtener éxito en el mejoramiento genético es necesario tener objetivos claros, lo cual dependerá del sistema de producción que se haya implementado, además del material genético, sistema de registros y biotecnologías disponibles (Wiggans & Gengler, 2011).

### **2.3.2 Factores ambientales**

Dentro de los factores ambientales, el más importante sin lugar a duda es la alimentación, debido a que es la base para que exista una producción de origen animal. La alimentación en los sistemas en pastoreo implica el hecho de que los animales deben moverse a cosechar su alimento y los pastos deben tener un manejo agronómico ideal en cuanto a su cultivo, para que los animales puedan obtener nutrientes de manera óptima, lo cual también dependerá de la especie de pasto y la ubicación geográfica del tambo. En muchos tambos adicionalmente se suministran raciones marginales de concentrados para suplir los requerimientos de los animales productivos (Moran, 2005).

Otro factor muy importante dentro de los ambientales es la sanidad animal, debido a que las enfermedades que se pueden presentar en un rebaño llegan a disminuir considerablemente la producción, inhibir la fertilidad de las vacas, inhibir el consumo de alimentos, entre otros daños, por lo que un plan de manejo sanitario, que incluya elementos de prevención y control oportunos, es trascendental en una explotación lechera. La crianza o levante de animales es un factor relacionado con el manejo de animales jóvenes para que puedan entrar en el proceso productivo, en el caso de la ganadería de leche esto comienza con el cuidado y levante de terneras hasta que lleguen a ser vacas que replacen eventualmente a las actuales productoras. Finalmente,

entre los factores determinantes ambientales determinantes tenemos la reproducción. La dinámica de la producción lechera exige que las vacas tengan partos recurrentemente en lapsos que no superen los 2 años. Esto se debe a que, de no hacerlo, eventualmente las vacas no entrarán en producción lechera, y esto repercutirá sobre la producción total del tambo. Para lograr éxito en el manejo reproductivo se requiere de una correcta planificación y uso de biotecnologías de mejoramiento de la reproducción animal como lo son la inseminación artificial, el trasplante de embriones y la sincronización de celos (Tentor, 2020).

## **2.4 Mejoramiento genético en el ganado lechero**

El mejoramiento genético en el ganado lechero ha permitido el desarrollo de la industria lechera en todo el mundo, la selección de mejores ejemplares y las técnicas desarrolladas en el área de la biotecnología han permitido un progreso muy grande desde la década de 1950, cuando la producción promedio en países como Estados Unidos estaba en 4,000 litros por lactancia por vaca en condiciones de confinamiento hasta la actualidad, cuando la producción promedio en dicho país es 10,000 litros por lactancia por vaca en dichas condiciones (DairyZn, 2020).

### **2.4.1 Breeding value y las evaluaciones genéticas tradicionales**

El mejoramiento genético debe ser medido, para que pueda ser evaluado, aunque en la mayoría de los tambos se lo hace de manera empírica, cuando se tiene un adecuado control de registros se pueden usar modelos matemáticos para estimar la mejora obtenida a través del programa de mejoramiento genético. La manera cuantitativa de expresar la mejora genética es a través del “breeding value” o valor de cría, este indicador refleja de alguna manera la suma del efecto de muchos genes que interactúan para expresar una característica animal de importancia como el peso o la producción lechera. A pesar de ello, la limitación a su uso es que únicamente permite estimar una probabilidad de mejora genética sobre la progenie de los animales evaluados

cuando se tienen factores ambientales iguales a los que se usaron para establecer el modelo (Bourdon, 1999).

El valor de cría se obtiene con el uso de modelos de regresión denominados BLUP (mejor predicción lineal insesgada, por sus siglas en inglés), en los que se estima la variación poblacional a través de métodos RELM (máxima verosimilitud restringida, por sus siglas en inglés) para lo que es necesario plantear modelos experimentales estrictos para que los modelos de inferencia no presenten sesgos. Un trabajo de tal magnitud además de ser costoso implica grandes esfuerzos para obtener buenas estimaciones, como lo que se ha logrado en países como Estados Unidos (Piñehira, Riverós, & Felmer, 2009).

#### **2.4.2 Evaluación genómica**

La evaluación genética tradicional requiere de mucho tiempo, costos y un adecuado manejo experimental. Adicionalmente, la recombinación de genes, que es parte de un proceso biológico, hace que por métodos tradicionales exista una gran variación en las predicciones. Actualmente con el avance de la biología molecular y genómica, es posible conocer a ciencia cierta, si la progenie de un animal ha heredado o no los genes que codifican para características deseables de producción, mediante evaluaciones genómicas cuando estos todavía son jóvenes y no han entrado en la etapa productiva. A pesar de esto, los aún altos costos de esta tecnología en el medio local imposibilitan a la mayoría de los productores su utilización; sobre todo porque estos no manejan registros ni usan modelos de predicción de ningún tipo, que les permita dinamizar su programa de mejoramiento genético (Piñehira, Riverós, & Felmer, 2009).

#### **2.5 Parámetros de la producción lechera**

El manejo de una explotación lechera puede ser evaluado por medio de parámetros productivos, reproductivos, nutricionales y sanitarios. Estos parámetros sirven de indicadores cuantitativos de la eficiencia y eficacia de un tambo lechero. Además de permitir mantener un

sistema de control de la calidad dentro de un cuadro de mando integral en el sistema productivo, también son herramientas de planificación operativa para los productores.

### **2.5.1 Parámetros productivos**

Dentro de los parámetros productivos, en la producción lechera tenemos los relacionados a la eficiencia en la producción de leche cruda, que es el producto final de un tambo lechero. Se consideran los siguientes parámetros de importancia:

**Producción de leche diaria:** Es la cantidad total de leche que produce una vaca en producción durante un día, esto equivale a la sumatoria de la producción de leche de la vaca en cada ordeño en un día, su registro es esencial en cualquier tambo lechero.

**Producción de leche por lactancia:** Es la cantidad de leche que una vaca produce en una lactancia, es decir desde que pare una cría hasta que su producción láctea cesa, debido a que el periodo de lactancia de las vacas es variable en función de varios factores, se toma en cuenta como estándar la lactancia ajustada a 305 días, de tal manera que la producción pueda ser comparada entre animales y entre lactancias (Buchanan, 2011).

### **2.5.2 Parámetros ambientales**

Los parámetros ambientales que nos permiten evaluar la nutrición animal son muy diferentes dependiendo del sistema de producción, esto se debe a que la diferencia fundamental entre sistemas está relacionada con la alimentación. En un sistema de producción en pastoreo los más comunes son:

**Peso:** El peso del animal es un parámetro fundamental para evaluar la nutrición, esto dependerá de la disponibilidad de alimento, raza, estado fisiológico y estado sanitario del animal. Una vaca en producción puede tener un peso entre 450 y 700 kg (Moran, 2005).

**Consumo de materia seca:** Es la cantidad de alimento seco que la vaca consume. Al referirse a alimento seco, debe entenderse que es el alimento al cual le ha sido extraído el agua, dado que cada alimento tiene diferentes cantidades de agua, que no aporta a los requerimientos nutricionales. El consumo de materia seca estará en función del peso del animal. Los valores normales oscilan entre 20 y 28 kg por día (Cambell, 1966).

**Producción de materia seca:** Es la cantidad de materia seca de pasto que una pradera produce, lo cual está determinado por el tipo de manejo pastoril, especies de pastos presentes en la pradera, niveles de fertilidad de la pradera, niveles de precipitaciones y riego entre otros. Los valores normales están entre 18 y 25 toneladas por hectárea por año (Cambell, 1966).

**Carga Animal:** Se refiere al número de animales que pastan por unidad de superficie. Este parámetro es afectado directamente por la producción de materia seca de la pradera y el manejo pastoril. Los valores normales oscilan en el ganado lechero oscilan entre 2 y 4 unidades bovinas adultas por hectárea (Cambell, 1966).

El monitoreo sanitario, es parte del bienestar animal relacionado con el ambiente, y es muy importante para detectar problemas de salud que pueden afectar la vida y producción del rebaño, estos son evaluados generalmente en todo el rebaño, aunque existen parámetros de ocurrencia de enfermedades por lapso temporal o por eventos como la lactancia que pueden ser evaluados de manera individual. Entre los más importantes tenemos:

**Mortalidad anual:** Mide la frecuencia de muertes en el rebaño en un año en relación con el número total de vacas en el rebaño. Un alto índice de mortalidad indica un mal manejo sanitario. Lo normal es que este parámetro este por debajo del 10% (Merck Sharp & Dohme, 2020).

**Recuento de células somáticas:** Este parámetro sirve como un indicador de la salud de la glándula mamaria, donde se dan los procesos fisiológicos que originan la producción láctea, además sirve como parámetro de calidad de la leche. Su medición se hace por el conteo de

leucocitos y células epiteliales por mililitro de leche. Los valores aceptables deben ser inferiores 750,000 células por mililitro de leche (Merck Sharp & Dohme, 2020).

**Tasa de abortos:** Es la proporción entre las gestaciones que terminaron en aborto y el número de gestaciones en un periodo de tiempo, un porcentaje elevado indica la posible presencia de enfermedades de tipo infeccioso como Brucelosis o Leptospira, o enfermedades de tipo sistémico. Este indicador debe estar por debajo de del 3% (Merck Sharp & Dohme, 2020).

### 2.5.3 Parámetros reproductivos

La eficiencia reproductiva es fundamental para mantener un volumen constante de producción de leche en el tambo, debido a que para que las vacas productivas son aquellas que han parido una cría. El parto es el evento fisiológico gatilla la producción láctea. En este contexto, es deseable que las vacas permanezcan dentro del grupo productivo un intervalo de tiempo aceptable, se preñen en poco tiempo y que tengan partos continuamente. Todo esto puede ser medido y evaluado con parámetros productivos específicos, entre los más importantes tenemos:

**Promedio de edad al primer parto:** Marca el inicio de la vida reproductiva de las vacas, usualmente se mide en meses, su valor está determinado por la raza de los animales, la eficiencia nutricional y un adecuado manejo sanitario, los valores óptimos del parámetro están entre 24 y 30 meses de edad.

**Intervalo entre partos:** Corresponde al periodo de tiempo entre un parto y el siguiente, la duración de este intervalo de tiempo está determinada por factores sanitarios, nutricionales e incluso genéticos, así como de un buen manejo las tecnologías referentes a la reproducción animal en el tambo. El evento que optimiza el tiempo en este parámetro es la pronta aparición de estro posparto. El valor optimo del parámetro esta entre 12 y 14 meses (Bustillo & Melo, 2020).

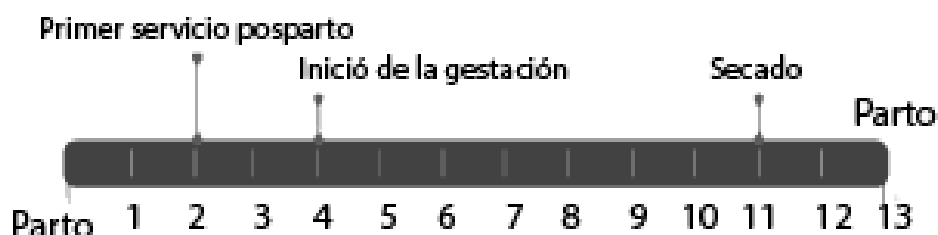
**Días al primer servicio después del parto:** Es el periodo de tiempo entre el parto y el primer servicio (intento de preñez), lo normal en este indicador oscila entre 50 y 60 días.

**Servicios por concepción:** Hace referencia a la cantidad de servicios (intentos de preñez) que tiene una vaca para quedar gestantes, ya sea por monta natural o por inseminación artificial. Un número aceptable en todo el rebaño oscila entre 1.7 y 2.5.

**Días abiertos:** Es la cantidad de días entre el parto hasta el inicio de la siguiente gestación, lo normal es que este indicador oscile entre 85 y 120 días.

**Duración del periodo seco:** Corresponde al intervalo de tiempo en que la vaca ha cesado su producción de leche, hasta el momento en que inicia nuevamente la producción producto de un nuevo parto, los valores normales del parámetro están entre 50 y 60 días.

**Gráfico 2** Ciclo reproductivo del ganado bovino en condiciones de manejo reproductivo ideales. Autor: Buchanan,2022



## 2.6 Machine Learning

Actualmente vivimos en un mundo que genera datos en gran cantidad, como hemos revisado hasta este momento los sistemas de producción animal como los tambos lecheros no son la excepción. Los datos que se están generando incluso en este momento tienen relaciones entre sí, existen patrones en los datos, no hay procesos completamente aleatorios, también en los tambos lecheros la producción de leche de cada vaca no se da al azar, depende de los factores de producción.

En este contexto a través de la ayuda de la programación, se pueden generar algoritmos o inferirlos a partir de los datos para poder identificar de manera aproximada el proceso que origina una salida de datos concluyente, esto es Machine Learning. En el caso de los tambos lecheros, se busca identificar ciertos patrones que nos permitan entender el proceso que determina la producción de leche en un sistema específico y usar esa información para poder hacer predicciones acerca de la producción de leche.

Machine Learning permite elaborar un modelo de datos, definido por ciertos datos de entrada y a través de la programación optimizar los parámetros del modelo y usarlos para predecir salidas de datos futuros. Machine Learning utiliza la teoría estadística para la elaboración de modelos, esto es porque se está realizando una inferencia a través de una muestra; y además usa la programación computacional para optimizar los parámetros del modelo mediante algoritmos eficientes tanto para la optimización del modelo como para su uso en la predicción (Alpaydin, 2014).

### **2.6.1 Aprendizaje Supervisado**

En Machine Learning tenemos diferentes métodos de ajuste de los parámetros de un modelo, lo que se conoce como aprendizaje. En el aprendizaje supervisado, para cada ejemplo tenemos datos de entrada y los datos de salida deseados. Además, asumimos que existe un “supervisor” que puede proveer de los datos de salida deseados para un determinado dato de entrada. El aprendizaje se refiere al ajuste de parámetros de tal manera que el modelo pueda realizar las predicciones más precisas de los datos, en otras palabras, el aprendizaje se refiere al mejoramiento del parámetro de acuerdo con un criterio de desempeño, que puede ser la predicción, clasificación entre otras tareas. Un modelo de aprendizaje supervisado busca aprender de un conjunto de datos de entrada y predecir futuros datos de salida de manera precisa (Flach, 2012).

## **2.6.2 Aprendizaje no supervisado**

En este tipo de aprendizaje no existe un supervisor, es decir no se cuentan con datos de entrada y salida apareados, solo existen datos de entrada. Para aprender se trata de encontrar regularidades en los datos de entrada. En estos casos se busca una estructura dentro de los datos de entrada, en la que se puedan encontrar patrones en los datos que son más comunes, con esto se puede identificar qué es lo que ocurre con más frecuencia, desde el punto de vista estadístico este proceso es la estimación de la densidad. Existen varios métodos de estimación de la densidad como “clustering”, y asociación (Alpaydin, 2014).

## **2.7 El modelo de regresión lineal**

En una regresión, el dato de salida es de tipo numérico por lo que se busca aprender una función numérica. En Machine Learning, la función es desconocida, pero a través de un conjunto de datos de entrenamiento (ejemplos) se la puede calcular.

### **2.7.1 Fundamentos estadísticos y matemáticos**

En una regresión se busca una función  $f(x)$  que pase por los puntos formados por la variable de salida y la variable de entrada. En la regresión, adicionalmente encontramos ruido, entendido como irregularidad indeseable en los datos, que se añade a los datos de salida de la función desconocida  $y = f(x) + \epsilon$ . El ruido en la regresión se genera por las variables que tienen un efecto en los datos de salida, pero no se las conoce. El cálculo de la función se hace mediante de un método de estimación de parámetros poblacionales, como el de mínimos cuadrados ordinarios que busca la distancia más corta entre los valores reales y predichos (Alpaydin, 2014).

### **2.7.2 Fundamentos de programación en R**

Para aplicar regresiones el lenguaje R tiene la función `lm` como se la caracteriza a continuación:

```
lm(formula, data, subset, weights, na.action,  
  
method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,  
  
singular.ok = TRUE, contrasts = NULL, offset, ...)
```

Los valores más importantes de la función son:

**Formula:** es un objeto de la clase fórmula para indicar la variable dependiente y las variables independientes. Por ejemplo, si formula =  $y \sim x1 + x2$  lo que se indica es que la variable dependiente es y, las variables independientes serían x1 y x2.

**Data:** Se refiere al marco de datos donde se buscarán las variables (Hernández & Mazo, 2020).

### 2.7.3 Fundamentos de programación en Python

Para realizar regresiones en Python, bajo el esquema de Machine Learning se pueden utilizar varias librerías, entre las principales NumPy, Pandas y Scikit-learn.

Los siguientes códigos nos permiten ingresar los parámetros de la regresión:

```
from sklearn.linear_model import LinearRegression  
  
regressor=LinearRegression()  
  
regressor=regressor.fit(X_train, y_train )
```

Con este código habiendo cargado los datos correspondientes se puede obtener la regresión, adicionalmente el paquete Scikit-learn permite dividir los datos de aprendizaje y prueba.

## **2.8 Cross- validation**

Los modelos de predicción dependen de parámetros estructurales, que se refieren a los coeficientes y parámetros de ajuste, que controlan la penalización en los datos para lograr las mejores predicciones posibles. El método de cross validación permite obtener los mejores parámetros. Un modelo predictivo es considerado bueno cuánto es capaz de predecir muestras no vistas con precisión. Los diferentes tipos de errores, en predicción son el error training y el error de prueba, el primero es la pérdida media sobre la muestra original, y el segundo es el error de predicción sobre muestras independientes. Los modelos pueden diferir en función del sesgo y la varianza permitidas, los modelos con menos grados de libertad tienen menos varianza, pero más sesgo, pero los modelos con muchos grados de libertad tienen más varianza y menos sesgo (Venturini, 2016).

Los modelos más complejos tienden a aumentar el error de prueba debido el problema de overfitting, en tales casos se modela el ruido más no la tendencia natural de los datos. Para optimizar el balance entre sesgo varianza el método más utilizado es la Cross validación. Este método permite utilizar muestras independientes para reducir el error de prueba, de tal manera que el modelo con menor error de prueba provee el mejor predictor. La ventaja de este método es que permite utilizar el mismo conjunto de datos de muestra para realizar la validación, el algoritmo permite dividir el conjunto de datos en 2, conservando su distribución, sin tener que recolectar más datos.

La cross validación provee un estimador del error de prueba para cada modelo, por lo que la selección del modelo se optimiza (Yang, 2007).

## **2.9 Publicación web de modelos de Machine Learning**

Para el desarrollo de aplicativos webs existen diferentes plataformas que permiten desarrollo de aplicativos con código de R y Python respectivamente. Una alternativa es Shiny aps, que es

un paquete de R que permite la creación de aplicaciones interactivas con R. Una aplicación Shiny tiene 2 funciones o archivos, el primero ui en donde se describe la interfaz gráfica de la aplicación, y otro llamado server donde se escribe la programación de Machine Learning. Con este modelo de publicación se crea una aplicación interactiva que permite manipular datos sin manipular el código.

Otra alternativa que permite construir aplicaciones web es Streamlit, con esta plataforma se pueden crear aplicaciones usando el lenguaje Python.

## CAPÍTULO III: METODOLOGÍA Y TÉCNICAS

---

### **3. Desarrollo del modelo de predicción de la producción lechera mediante CRISP-DM**

En este capítulo se describen las acciones tomadas para el desarrollo del modelo de Machine Learning., para ello se detallan cada una de las fases de la metodología CRISP-DM, que es la que se utilizó para esta investigación.

#### **3.1 Comprensión del negocio**

El objetivo de esta fase es conocer los factores que determinan el resultado del proyecto, se formulan objetivos y requerimientos desde una perspectiva del negocio, en esta fase se ligan los factores de negocio con los factores técnicos para conseguir el resultado deseado (Rodríguez, Piñeiro, & Del Llano, 2014).

##### **3.1.1 Perspectiva del negocio**

La granja lechera del Instituto Agropecuario Superior Andino (IASA), tiene por objetivo maximizar la utilidad económica en su giro de negocio que es la venta de leche cruda; para ello debe maximizar la producción lechera a la vez que minimiza costos. Esta granja trabaja bajo un sistema de producción en pastoreo, y su modelo de trabajo ha sido estandarizado por más de 20 años. Actualmente, este tambo lechero está buscando una herramienta de control de operaciones y selección de animales, por lo que un modelo de Machine Learning alimentado de parámetros de las vacas lecheras del tambo se convierte en una herramienta de predicción ideal para subsanar estas necesidades.

### 3.1.2 Objetivos del negocio

Desarrollar un modelo predictivo utilizando algoritmos de Machine Learning, para estimar la producción lechera por vaca en la hacienda el Prado, del Instituto Agropecuario Superior Andino IASA. Teniendo como data inicial los registros de parámetros por vaca, en los periodos comprendidos entre 2013 y 2017.

### 3.1.3 Evaluación de la situación

Actualmente el control de operaciones en el IASA se realiza a través del uso de métricas empíricas basadas en registros físicos y electrónicos, que se elaboran mensualmente en diferentes formatos. En los primeros, se registra información concerniente a partos de vacas, entradas y salidas de vacas en las diferentes cuentas de la ganadería, enfermedades, entre otras. Adicionalmente, el tambo cuenta con el Software Ganadero SG, donde se ingresan datos referentes a reproducción animal y producción. Con la información generada los administradores evalúan el proceso operativo, sin embargo, este sistema carece de estándares comunes que permitan detectar problemas de mermas de producción y corregir rápidamente sus causas. Aunque el tambo cuenta con un equipo de trabajo de profesionales en diferentes áreas, muchas veces es difícil detectar problemas y formular soluciones rápidamente.

#### *i. Recursos necesarios*

**Tabla 1** Personal del proyecto. Autor: Jorge Ordoñez, 2022

<b>Cargo</b>	<b>Título</b>	<b>Experiencia</b>	<b>Cantidad</b>
Jefe del área de ganadería	Ing. Zootecnista Msc. Reproducción Animal	20 años	1
Asistente de operaciones de ganadería	Ing. Agropecuario	5 años	1
Analista de sistemas	Ing. Sistemas	3 años	1

**ii. Equipos**

**Tabla 2.** Equipos requeridos en el proyecto. Autor: Jorge Ordoñez,2022

Descripción	Características	Cantidad
Laptop o Pc de escritorio	Intel i7 Gen10 o superior 1TB Disco 16GB Ram	2

**iii. Riesgos y contingencias**

**Tabla 3** Riesgo y contingencias del proyecto. Autor: Jorge Ordoñez,2022

Riesgo	Contingencia
Cambios Administrativos	Elaborar un contrato que incluya cláusulas de herencia de responsabilidades con cualquier dirigencia administrativa que estuviere a cargo
Registros de datos insuficientes	Revisar registros físicos y transcribirlos de haberlos.
Eficacia de modelo	No incluir datos complejos

**iv. Costos**

**Tabla 4** Costos del proyecto. Autor: Jorge Ordoñez,2022

Rubro	Valor
Materiales	\$1,000
Subtotal sin IVA	\$1,000

Total	\$1,120
-------	---------

### 3.1.4 Determinación de los objetivos de la modelación

- Realizar la revisión de la literatura que integre el uso de un modelo predictivo aplicando algoritmos de Machine Learning y la producción lechera.
- Analizar y evaluar la información de los modelos prototipo.
- Seleccionar el modelo predictivo adecuado en función del nivel de precisión de los resultados.

### 3.1.5 Plan de proyecto

**Tabla 5** Plan del proyecto. Autor: Jorge Ordoñez,2022

Tarea	Fecha de Inicio	Fecha de Fin	Entregable
Comprensión del Negocio	6-Jun	5-Aug	Autorización de datos
Comprensión de los de Datos	8-Aug	19-Aug	Entrega de datos
Preparación de Datos	22-Aug	31-Aug	Base de datos final
Modelado	1-Sep	14-Sep	Modelo de datos
Evaluación	17-Sep	30-Sep	Ajuste de modelo
Implementación	3-Oct	14-Oct	Acta de entrega recepción

## **3.2 Comprensión de los datos**

En esta fase el enfoque es la recopilación de los datos desde su fuente pura (Software Ganadero SG, datos en estado físico), mantener el formato de origen, y después de ello explorar los datos usando tablas y gráficas para evaluar su calidad y viabilidad para el proyecto.

### **3.2.1 Recolección inicial de los datos**

La digitalización de un grupo de datos se la realizó través del gestor de base de datos del programa Excel, de tal forma que se pudo organizar la información, y posteriormente se realizara la unión con la información del software que maneja la administración de la ganadería. Entre los registros que sirvieron para obtener la base de datos estuvieron los registros físicos de: libro de la ganadería, registros productivos y registros reproductivos.

### **3.2.2 Descripción de los datos**

Para alimentar el modelo de predicción, se requieren variables que corresponden a información de las vacas en producción en función del ciclo de producción, conocido como lactancia. Esta información se obtiene de los registros mensuales de producción, reproducción y la dinámica de entradas y salidas de animales en producción que se conoce como libro de la ganadería, además de los registros de la estación meteorológica más aledaña. Las anteriores bases mencionadas fueron generadas en formato csv.

### **3.2.3 Exploración de los datos**

Los datos recopilados corresponden a 277 registros de 27 variables, estos han sido recopilados en formato xls y posteriormente convertidos a formato csv. La recopilación de datos de la ganadería dentro de la hacienda el Prado comenzó aproximadamente en el año 1998, sin embargo, mucha información se perdió por no hacer uso de esta, los registros con los que se cuenta actualmente constan desde el año 2012. Actualmente, los registros son tomados según

se van generando de manera física y con el Software Ganadero SG. Sin embargo, no se ha podido sintetizar toda la información a partir del año 2018, por no tener información de todas las variables de manera íntegra, por lo que la muestra tomada comprende los registros generados entre los años 2013 y 2017 tiempo en el que se generaron los 277 registros analizados. Con el tamaño actual del rebaño se generan aproximadamente entre 6 y 8 registros mensuales que se generan por cada parto.

Una muestra de la base de datos inicial, donde se exponen registros del año 2013 con la información de las fuentes citadas se puede ver a continuación:

**Gráfico 3** Toma de captura de una muestra de la base de datos. Autor: Jorge Ordoñez, 2022

	id	arete	raza	fnacimiento	fnacimiento.1	fconcepcion	fparto	fsecado
1	2	240	holstein	9/10/2002		5 9/20/2012	7/1/2013	9/18/2014
2	5	301	f1	1/15/2003		8 9/20/2012	6/26/2013	8/9/2013
3	6	313	holstein	4/20/2003		9 1/10/2013	10/26/2013	8/30/2014
4	6	313	holstein	4/20/2003		10 2/14/2014	11/25/2014	7/21/2015
5	7	319	holstein	6/29/2003		4 2/16/2013	12/10/2013	8/6/2014
6	12	408	holstein	4/6/2004		5 8/12/2012	6/2/2013	8/9/2013
	servicios	natimuerto	aborto	produccion	pre_2antes	pre_1antes	pre_mesparto	
1	3		1	0	3447	75.1	0.6	8.3
2	3		1	0	695	106.2	75.1	0.6
3	2		0	0	3946	57.5	7.3	127.1
4	1		0	0	2799	28.5	115.8	140.2
5	2		0	0	2424	127.1	86.4	93.3
6	2		0	0	781	106.2	75.1	0.6
	pre_1pos	pre_2pos	rad_2antes	rad_1antes	rad_mesparto	rad_1pos	rad_2pos	
1	57.5	7.3	225.0000	124.0928	224.00000	225.0000	124.15929	
2	8.3	57.5	224.0000	163.3679	124.09277	224.0000	225.00000	
3	86.4	93.3	151.7963	124.1593	78.84142	151.7963	145.49376	
4	156.7	34.8	115.4269	204.9985	153.47510	115.4269	160.82230	
5	135.3	100.2	117.3201	151.7963	145.49376	117.3201	96.11147	
6	8.3	57.5	224.0000	163.3679	124.09277	224.0000	225.00000	
	hums_2antes	hums_1antes	hums_mesparto	hums_1pos	hums_2pos			
1	34.10798	30.15693	17.81038	10.94347	9.83416			
2	30.40626	34.10798	30.15693	17.81038	10.94347			
3	10.94347	9.83416	19.36059	31.03356	30.31932			
4	9.83416	19.36059	31.03356	30.31932	25.41054			
5	19.36059	31.03356	30.31932	25.41054	33.87778			
6	30.40626	34.10798	30.15693	17.81038	10.94347			

### 3.3 Preparación de los datos

La preparación de los datos incluye las actividades necesarias para adecuar el conjunto de datos al modelo, entre las actividades principales tenemos: limpieza de datos, clasificación de

datos, generación de variables, integraciones y cambios de formato. También se usan herramientas de visualización de datos para establecer relaciones.

### 3.3.1 Selección de datos

La síntesis de la información fue muy importante para organizar los datos encontrados, debido a que se tuvo que dar estructura a los datos, el conjunto de datos se presenta de tal manera que cada registro corresponda al número de lactancia de una vaca específica, en total se recopiló información de 277 lactancias, de cada observación se tuvo un conjunto inicial de 27 variables. Las variables seleccionadas se detallan a continuación:

**Tabla 6.** Variables para el modelo predictivo de la producción. Autor: Jorge Ordoñez, 2022

Número de variable	Variable	Grupo variable
1	Id	Identificación
2	Arete	Identificación
3	Raza	Identificación
4	Fecha de nacimiento	Identificación
5	Número de lactancia	Producción
6	Fecha de concepción	Reproducción
7	Fecha de parto	Reproducción
8	Fecha de secado	Reproducción
9	Servicios por concepción	Reproducción
10	Natimuerto	Reproducción
11	Aborto	Reproducción
12	Producción	Producción
13	Precipitación 2 meses antes del parto	Agrometeorológica
14	Precipitación 1 mes antes del parto	Agrometeorológica
15	Precipitación en el mes de parto	Agrometeorológica
16	Precipitación 1 mes después del parto	Agrometeorológica
17	Precipitación 2 meses después del parto	Agrometeorológica
18	Radiación Solar 2 meses antes del parto	Agrometeorológica
19	Radiación Solar 1 mes antes del parto	Agrometeorológica
20	Radiación Solar en el mes del parto	Agrometeorológica
21	Radiación Solar 1 mes después del parto	Agrometeorológica
22	Radiación Solar 2 meses después del parto	Agrometeorológica
23	Humedad del suelo 2 meses antes del parto	Agrometeorológica

24	Humedad del suelo 1 mes antes del parto	Agrometeorológica
25	Humedad del suelo en el mes del parto	Agrometeorológica
26	Humedad del suelo 1 mes después del parto	Agrometeorológica
27	Humedad del suelo 2 meses después del parto	Agrometeorológica

Las variables seleccionadas, son las que reflejan de mejor manera la influencia de cada factor en el proceso productivo. Estas variables están divididas en tres grupos, correspondientes a los factores que son más fáciles de cuantificar, y con los que se encontró información completa que son: producción, reproducción y factores agrometeorológicos (por su influencia directa en el sistema de producción de pastoreo); adicionalmente se contemplan variables de identificación para cada animal.

La descripción de las variables descritas en la tabla 2 es la siguiente:

1. Id: registro único en la base de datos para cada vaca ingresada.

1. Arete: registro único utilizado en el tambo lechero, indica año de nacimiento, y el número de parto del año en que el animal nació.

2. Raza: indica la raza del animal, esta variable identifica la influencia genética en la producción, en el tambo del IASA existe un sistema de cruzamiento que busca obtener animales de raza Montbéliarde, con ejemplares iniciales de raza Holstein, los valores posibles en esta variable son los siguientes:

**Tabla 7.** Descripción de la variable raza. Autor: Jorge Ordoñez, 2022

<b>Raza</b>	<b>Descripción</b>
Holstein	Animal de raza Holstein pura
F1	Animal 50% Montbéliarde, 50% Holstein
F2	Animal 75% Montbéliarde, 25% Holstein
F3	Animal 87.5% Montbéliarde, 12.5% Holstein

3. Fecha de nacimiento: indica el día, mes y año de nacimiento del animal
4. Número de lactancia: indica el número de parto al que corresponde ese ciclo productivo en particular.
5. Fecha de concepción: indica el día, mes y año en el que se dio la concepción que originó el parto y la lactancia correspondientes.
6. Fecha de parto: corresponde al día, mes y año en el que se dio el parto de esa lactancia en particular.
7. Fecha de secado: indica el día, mes y año en el que se dejó de ordeñar a la vaca y se dio por terminado el ciclo productivo correspondiente a esa lactancia en particular.
8. Servicios por concepción: indica cuantas veces la vaca tuvo que ser montada o inseminada hasta lograr la preñez que dio origen a esa lactancia en particular.
9. Natimuerto: indica si la cría nació muerta en esa lactancia en particular.
10. Aborto: indica si la vaca abortó, es decir, si la vaca finalizó la preñez antes del nacimiento de la cría en una lactancia en particular.
11. Producción: corresponde a la cantidad de leche en litros que la vaca produjo desde que parió hasta que fue secada o hasta que cumplió 305 días produciendo leche después del parto.
12. Precipitación 2 meses antes del parto: corresponde al valor de precipitación mensual 2 meses antes del parto, esta variable influye sobre el crecimiento de los pastos, que es la fuente de alimentación de las vacas en pastoreo. 2 meses antes del parto la vaca almacena gran cantidad de nutrientes para poder alcanzar una máxima producción. Los valores de precipitación se expresan en milímetros.
13. Precipitación 1 mes antes del parto: corresponde al valor de precipitación mensual 1 mes antes del parto.
14. Precipitación en el mes del parto: indica el valor de precipitación mensual en el mes del parto.

15. Precipitación 1 mes después del parto: corresponde al valor de precipitación 1 mes después del parto; según el perfil de producción lechero entre los 60 y 90 días la vaca alcanzara su producción máxima que es un factor determinante del total de producción en la lactancia.
16. Precipitación 2 meses después del parto: corresponde al valor de precipitación 2 meses después del parto.
17. Radiación Solar 2 meses antes del parto: corresponde a la cantidad de radiación solar que recibe un área específica en el lapso de un mes, el dato corresponde a 2 meses previo al parto, su valor influye sobre el nivel de fotosíntesis en las plantas, en este caso en los patos, se mide en watts por metro cuadrado.
18. Radiación Solar 1 mes antes del parto: corresponde a la cantidad de radiación solar que recibe un área específica en el lapso de un mes, el dato corresponde al mes previo al parto.
19. Radiación Solar en el mes del parto: corresponde a la cantidad de radiación solar que recibe un área específica en el lapso de un mes, el dato corresponde al mes del parto.
20. Radiación Solar 1 mes después del parto: corresponde a la cantidad de radiación solar que recibe un área específica en el lapso de un mes, el dato corresponde al mes después del parto.
21. Radiación Solar 2 meses después del parto: corresponde a la cantidad de radiación solar que recibe un área específica en el lapso de un mes, el dato corresponde a 2 meses después del parto.
22. Humedad del suelo 2 meses antes del parto: corresponde al valor de humedad disponible para la planta como porcentaje, los valores inferiores a 18%, en el tipo de suelo de la hacienda el Prado, indican un proceso de marchitez en las plantas, por lo que los pastos no crecen adecuadamente, mientras que valores superiores a 45% indican procesos de

anegamiento lo que también perjudica el crecimiento de los pastos; el dato particular corresponde a 2 meses antes del parto.

23. Humedad del suelo 1 mes antes del parto: corresponde al porcentaje de humedad disponible para la planta mensualmente, en el mes anterior al parto.

24. Humedad del suelo en el del parto: corresponde al porcentaje de humedad disponible para la planta mensualmente, en el mes del parto.

25. Humedad del suelo 1 mes después del parto: corresponde al porcentaje de humedad disponible para la planta mensualmente, en el mes posterior al parto.

26. Humedad del suelo 2 meses después del parto: corresponde al porcentaje de humedad disponible para la planta mensualmente, 2 meses después del parto.

Las variables descritas anteriormente, fueron identificadas en la base de datos con la nomenclatura mostrada a continuación y corresponden al siguiente tipo de variable:

**Tabla 8** Identificación del tipo de dato de las variables. Autor: Jorge Ordoñez,2022

Variable	Identificación	Tipo de variable
Id	id	integer
Arete	arete	object
Raza	raza	object
Fecha de nacimiento	fnacimiento	object
Número de lactancia	lactancia	object
Fecha de concepción	fconcepcion	object
Fecha de parto	fparto	object
Fecha de secado	fsecado	object
Servicios por concepción	servicios	float
Natimorto	natimuerto	boolean
Aborto	aborto	boolean
Producción	produccion	float
Precipitación 2 meses antes del parto	pre_2antes	float
Precipitación 1 mes antes del parto	pre_1antes	float
Precipitación en el mes de parto	pre_mesparto	float
Precipitación 1 mes después del parto	pre_1pos	float
Precipitación 2 meses después del parto	pre_2pos	float

Radiación Solar 2 meses antes del parto	rad_2antes	float
Radiación Solar 1 mes antes del parto	rad_1antes	float
Radiación Solar en el mes del parto	rad_mesparto	float
Radiación Solar 1 mes después del parto	rad_1pos	float
Radiación Solar 2 meses después del parto	rad_2pos	float
Humedad del suelo 2 meses antes del parto	hums_2antes	float
Humedad del suelo 1 mes antes del parto	hums_1antes	float
Humedad del suelo en el mes del parto	hums_mesparto	float
Humedad del suelo 1 mes después del parto	hums_1pos	float
Humedad del suelo 2 meses después del parto	hums_2pos	float

### 3.3.2 Limpieza de datos

Debido a que la base de datos se generó con la información física entregada y la información parcial del sistema de software, no se tuvieron valores en blanco, a continuación, perdemos ver en detalle:

**Gráfico 4** Salida de Jupiter notebook sobre valores nulos del conjunto de datos. Autor: Jorge Ordoñez, 2022

```
In [10]: dataset.isnull().sum()
```

```
Out[10]: id          0      pre_1antes      0
         arete       0      pre_mesparto   0
         raza       0      pre_1pos      0
         fnacimiento 0      pre_2pos      0
         lactancia   0      rad_2antes    0
         fconcepcion 0      rad_1antes    0
         fparto      0      rad_mesparto  0
         fsecado     0      rad_1pos      0
         servicios  0      rad_2pos      0
         natimuerto  0      hums_2antes   0
         aborto     0      hums_1antes   0
         produccion  0      hums_mesparto 0
         pre_2antes  0      hums_1pos     0
         dtype: int64
```

### **3.3.3 Construcción de nuevos datos**

A pesar de que los datos obtenidos son consistentes y reflejan procesos dentro del sistema ganadero del tambo, algunas variables debieron ser procesadas por calculo con el objetivo de hacer cada registro comparable y poder modelar y predecir la producción en función de sus factores determinantes. Por esta razón se generaron las siguientes variables:

Identificador vaca - lactancia: Debido a que la variable id refleja una vaca en particular, la cual puede tener varias lactancias, y la variable lactancia refleja una lactancia en particular. Se fusionaron ambas variables para obtener una variable que refleje una lactancia en particular para cada lactancia.

Duración de la lactancia: Las variables de fecha no son de mucha utilidad por sí mismas, para reflejar un parámetro reproductivo se calculó esta variable, que es la diferencia entre la fecha de secado y la fecha de parto, un valor sobre los 305 días indica que la vaca tuvo problemas para ser preñada para una nueva lactancia, y valores inferiores a 305 días indica que la vaca tuvo que ser secada por alguna razón, generalmente patológica.

### **3.3.4 Formateo de los datos**

Para que el modelo no produzca resultados sesgados debido a casos de lactancias incompletas se procedió a crear dos variables para corregir este posible sesgo, al mismo tiempo que no se altere el significado de los datos, como se detalla a continuación: Factor de corrección de la producción: esta variable permite corregir una lactancia incompleta, es decir una lactancia que no llego a los 305 días, para que de esa manera sea útil en el modelo.

Producción por lactancia ajustada: Esta variable incluye la corrección de lactancias con valores inferiores a 305 días a través del uso del factor de corrección de la lactancia, y usa como ajuste el valor de producción de los primeros días de lactancia.

**Tabla 9** Identificación de nuevas variables. Autor: Jorge Ordoñez, 2022

<b>Variable</b>	<b>Identificación</b>	<b>Tipo de variable</b>
Id-lactancia	id-lactancia	object
Duración de la lactancia	dur_lactancia	float
Factor de corrección de la producción	factor_ajuste	float
Producción por lactancia ajustada	produccion1	float

### **3.4 Modelado**

Con el objetivo de realizar modelos predictivos existen varios algoritmos que pueden ser empleados entre los cuales tenemos la regresión lineal, regresión polinomial, regresión con uso de splines, redes neuronales, entre otros.

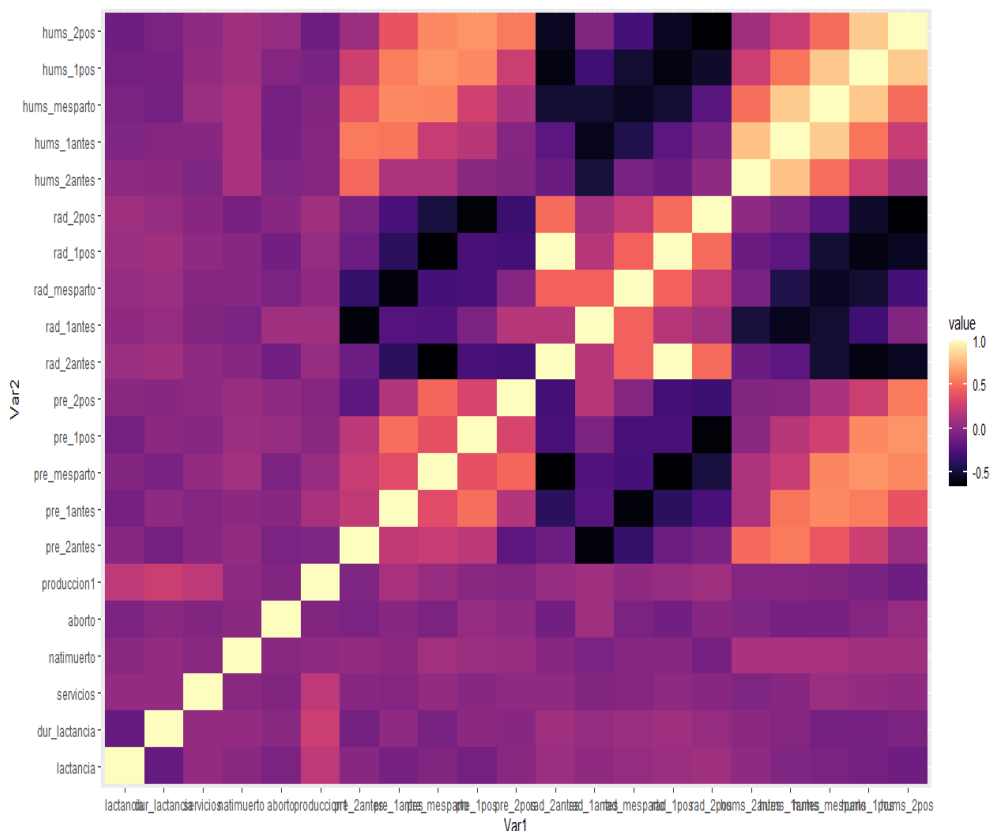
#### **3.4.1 Selección de técnica**

Al tratarse de un modelo de predicción la técnica seleccionada fue la regresión lineal múltiple, debido al ajuste de los datos. Este algoritmo se basa en aprendizaje supervisado. El modelo alcanza a realizar la predicción de un valor de salida basado en las variables independientes que influyen en este valor. Esta tarea se lleva a cabo a través de la relación entre los datos. En el caso de este estudio su selección se da por la naturaleza de los datos, la facilidad de entendimiento del algoritmo y el propósito del proyecto.

#### **3.4.2 Construcción del modelo**

Para la construcción del modelo y selección de las variables a utilizar en el mismo se procedió a realizar un diagrama de calor entre las variables numéricas para preseleccionar el conjunto de variables que permitan tener el mejor modelo.

**Gráfico 5** Diagrama de calor de variables numéricas. Autor: Jorge Ordoñez, 2022



El modelo planteado utilizó las variables independientes X, para la predicción de la variable Y como variable dependiente. En el gráfico anterior se puede evidenciar como la variable independiente producción está correlacionada con las variables: duración de la lactancia, producción 1, servicios por concepción, precipitación 2 meses antes del parto, precipitación 1 mes antes del parto, precipitación en el mes del parto, precipitación 1 mes después del parto, precipitación 2 meses después del parto, humedad del suelo 1 mes antes del parto, humedad del suelo en el mes de parto, humedad del suelo 1 mes después del parto y Humedad del suelo 2 meses después del parto.

**Gráfico 6.** Variables del modelo. Autor: Jorge Ordoñez, 2022

```
x = df[['raza', 'dur_lactancia', 'servicios', 'natimuerto', 'aborto',  
       'pre_2antes', 'pre_1antes', 'pre_mesparto', 'pre_1pos', 'pre_2pos', 'produccion1', 'hums_1antes',  
       'hums_mesparto', 'hums_1pos', 'hums_2pos']]  
y = df['produccion1']
```

Adicionalmente se dicotomizó la variable raza para poder ingresar dicha variable en el modelo.

El modelo utilizado fue el de regresión lineal de Sklearn con un conjunto de entrenamiento del 20% del conjunto total de datos.

**Gráfico 7** Salida de Jupiter notebook del modelo lineal. Autor: Jorge Ordoñez, 2022

```
from sklearn.model_selection import train_test_split  
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

```
from sklearn.linear_model import LinearRegression  
regressor = LinearRegression()  
regressor.fit(X_train, y_train)
```

```
LinearRegression()
```

### 3.4.3 Evaluación del modelo

El modelo evaluado desde su capacidad predictiva a través del coeficiente de determinación  $R^2$  ajustado presenta un valor de 0.798 cómo se puede apreciar en el gráfico 8, por tanto, nos permite dar por sentado que el modelo de regresión lineal de Sklearn es válido para su uso en predicción. Este indicador  $R^2$ , permite inferir el grado de variación que puede ser determinado en la variable dependiente a través del modelo, eliminando el efecto de la cantidad de variables. Los valores de  $R^2$  superiores a 0.75 se consideran significativos en modelos predictivos. (Allwright, 2022).

**Gráfico 8.** Evaluación estadística del modelo. Autor: Jorge Ordoñez, 2022

OLS Regression Results			
Dep. Variable:	produccion	R-squared:	0.812
Model:	OLS	Adj. R-squared:	0.798
Method:	Least Squares	F-statistic:	55.20
Date:	Wed, 16 Nov 2022	Prob (F-statistic):	5.31e-65
Time:	23:51:47	Log-Likelihood:	-1611.7
No. Observations:	221	AIC:	3257.
Df Residuals:	204	BIC:	3315.
Df Model:	16		
Covariance Type:	nonrobust		

### 3.5

#### Evaluación

Esta sección corresponde a la validación de los resultados obtenidos en el modelo con datos del negocio, en este caso el conjunto de datos de prueba, de esta manera se verifica que el modelo sea útil y da solución al problema planteado en la fase inicial.

##### 3.5.1 Evaluación de los resultados

Para evaluar la precisión de la predicción se realizaron una evaluación entre los resultados del modelo entre los conjuntos de entrenamiento y de prueba. Encontrando los siguientes resultados:

**Gráfico 9** Métricas de evaluación del modelo. Autor: Jorge Ordoñez, 2022

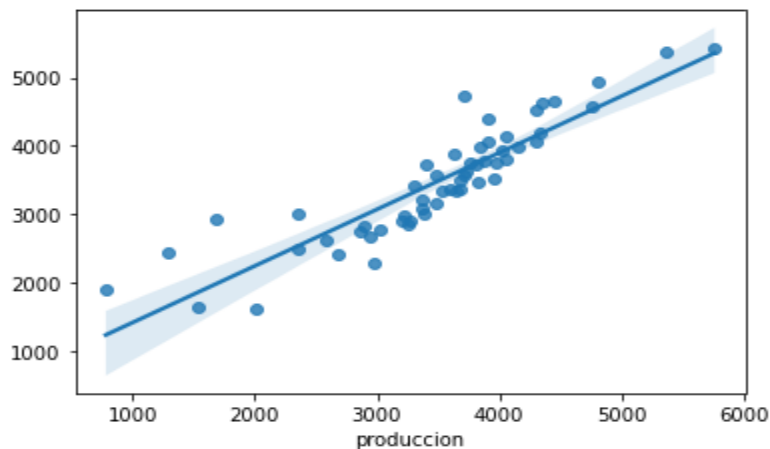
```
from sklearn import metrics
print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, y_pred))
print('Mean Squared Error:', metrics.mean_squared_error(y_test, y_pred))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
```

```
Mean Absolute Error: 293.6814120460367
Mean Squared Error: 159084.299531074
Root Mean Squared Error: 398.8537320009354
```

Al comparar el resultado de la raíz del error cuadrático con la media de producción total, (se puede validar que existe una relación en porcentaje de aproximadamente 11%. Esto significa que el algoritmo es muy preciso, pero aún puede hacer predicciones razonablemente buenas,

que aportan para los fines de la empresa lechera. El siguiente gráfico muestra el nivel de precisión del modelo lineal, en este se puede notar que la mayoría de los datos están cerca de la línea de regresión, lo que indica un buen ajuste sobre todo en valores superiores a 2000.

**Gráfico 10.** Gráfico de ajuste lineal del modelo entre valores predichos y reales. Autor: Jorge Ordoñez, 2022



### 3.5.2 Determinación de pasos a seguir

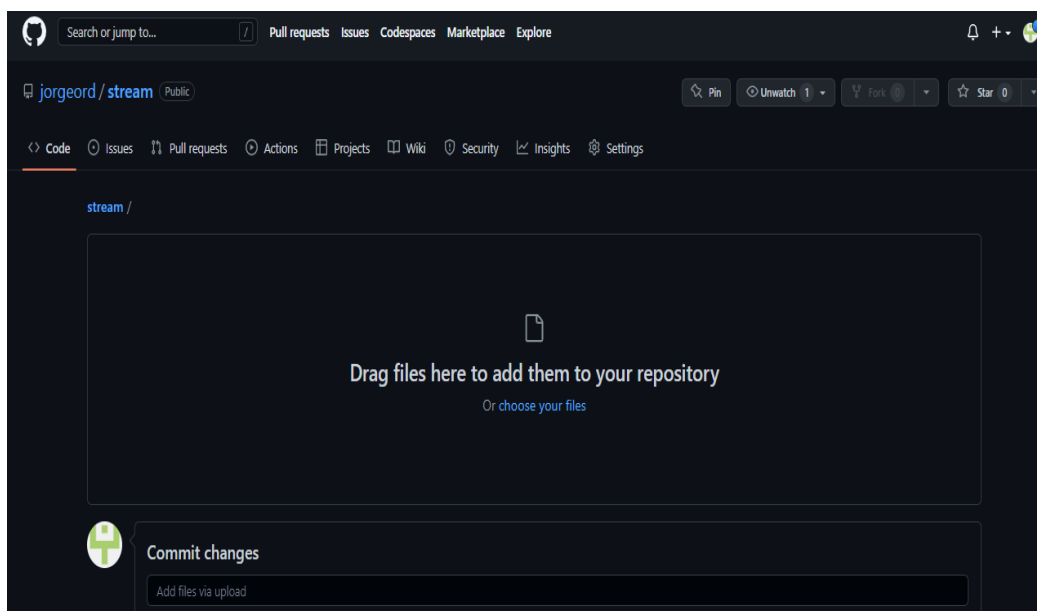
Para que el modelo desarrollado tenga utilidad, se procederá al desarrollo del aplicativo web para que los usuarios finales puedan seguir alimentando el modelo con más datos y puedan hacer uso del mismo en predicción y control de operaciones.

### 3.6 Implementación

Una vez desarrollado el aplicativo web, se procederá con la entrega y capacitación en el uso de este a los administradores y encargados de procesos operativos en la ganadería del IASA, a su vez ellos pondrán a disposición esta información según su criterio interno como institución, dependerá también su correcto funcionamiento, mantenimiento y mejoramiento a partir del acta suscripción del acta entrega recepción de las partes involucradas.

**Gráfico 11.** Módulo de subida de archivos para actualización de registros en Git Hub Autor:

Jorge Ordoñez, 2022



Adicionalmente el encargado de registros podrá alimentar la base de datos generada con nueva información, para lo cual se le proveerá acceso a la cuenta de Git Hub donde esta almacenado el archivo csv que es utilizado en la corrida del modelo, con lo cual los cambios serán ingresados directamente al modelo, lo mencionado se puede evidenciar en gráfico 11.

### 4.1 Desarrollo y prueba de aplicativo web

El aplicativo web fue desarrollado con la plataforma Streamlit que permite el desarrollo de aplicaciones mediante el uso de un “framework” del lenguaje de programación de Python, que además se acopla como una librería dentro de Python. Dentro de la aplicación, en el front-end se establecieron un bloque introductorio donde se exponen los datos más importantes del modelo desarrollado y el origen de los datos; un bloque de ingreso de datos del usuario, en donde se ingresan los datos de las variables a través de los widgets slider y radio; y finalmente un bloque de salida donde se expone el resultado de la predicción.

**Gráfico 12.** Captura de pantalla del encabezado del aplicativo web. Autor: Jorge Ordoñez, 2022

## Aplicativo de la producción lechera por vaca del IASA

Esta aplicación permite predecir la producción por lactancia de vacas en condiciones de pastoreo , bajo un sistema de pastoreo rotacional, bajo condiciones de clima templado como en la Sierra del Ecuador.

La predicción se realiza con el algoritmo de Machine Learning de regresión lineal que ha sido entrenado con datos de producción de la ganadería de la Hacienda el Prado del IASA.

En esta hacienda se ha mantenido el sistema de cruzamiento absorbente Montbéliarde x Holstein que corresponde al componente genético de la ganadería.

Se detalla el encabezado del aplicativo del modelo de predicción del IASA aplicando algoritmos de Machine Learning, así como también los dos tipos de cruces de razas ganaderas de la hacienda.

**Gráfico 13.** Captura de pantalla de variables del factor genético. Autor: Jorge Ordoñez,

2022

## Variables del modelo predictivo

Por favor ingrese los valores requeridos para la estimación:

### Factores genéticos

Escoger entre las razas: ( cruce absorbente entre Holstein y Mombeliarde)

- Holstein    F1 (50% Montbéliarde, 50% Holstein)    F2 (75% Montbéliarde, 25% Holstein)  
 F3 (87.5% Montbéliarde, 12.5% Holstein)

Se visualiza las variables aplicadas del modelo en primera instancia las de factores genéticos mismos que constan los cuatro tipos de razas que se manejan en la ganadería de la hacienda.

**Gráfico 14.** Captura de pantalla de sección de selección de variables de producción del aplicativo web. Autor: Jorge Ordoñez, 2022

### Variables de producción

Duración de la lactancia

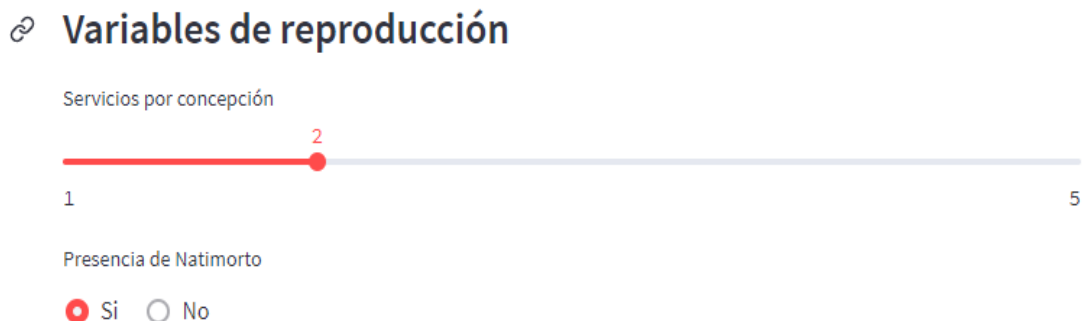


Lactancia en muestra día 10 de la producción ajustada



En esta sección se pueden encontrar dos barras de selección de valores numéricos para las variables de producción del modelo que son: duración de lactancia y producción 1 que corresponde al valor de producción al día 10 de la lactancia.

**Gráfico 15.** Captura de pantalla de sección de selección de variables de reproducción del aplicativo web. Autor: Jorge Ordoñez, 2022



En esta sección podemos encontrar una barra de selección de valores correspondiente a la variable servicios por concepción en la que se indica cuantos intentos de preñez tuvo la vaca para alcanzar esa lactancia. También para la selección de la variable presencia de natimuerto se halla un botón de selección según el acontecimiento o no del caso.

**Gráfico 16.** Captura de pantalla de sección de selección de variables agroecológicas del aplicativo web. Autor: Jorge Ordoñez, 2022



En esta sección se pueden apreciar las barras de selección de las variables agroecológicas que suministran al modelo los valores de precipitación y humedad del suelo.

**Gráfico 17.** Captura de pantalla de sección de resultado del modelo predictivo del aplicativo web. Autor: Jorge Ordoñez, 2022

## Resultado del modelo predictivo

La producción estimada por lactancia es: 3141.29 litros

En esta sección se muestra el valor obtenido en la predicción en kilogramos o litros por lactancia. Todos los valores predeterminados en el aplicativo corresponden a los valores medios o más comunes según la experiencia de los técnicos.

En el back-end se incrusto el código del modelo completamente tomando los coeficientes de la regresión y especificando su interacción con los ingresos que establece el usuario. Todo el código, fue cargado en Git Hub para su uso dentro de la plataforma Streamlit y su deployment. El acceso a la aplicación es público mediante el enlace <https://jorgeord-stream-app1-xd3l7z.streamlit.app/>.

## Gráfico 18. Captura de pantalla del código de la aplicación (back-end). Autor: Jorge

Ordoñez, 2022

```
65 with datainput:
66     st.header('Variables del modelo predictivo')
67
68     st.markdown('Por favor ingrese los valores requeridos para la estimación:')
69
70
71     st.subheader('Factores genéticos')
72     raza = st.radio('Raza',['Holstein','F1 (50% Montbéliarde, 50% Holstein)','F2 (75% Montbéliarde, 25% Holstein)','F3 (87.5% Montbéliarde, 12.5% Holstein)'])
73
74     st.subheader('Variables de producción')
75     dur_lact = st.slider('Duración de la lactancia', 10, 500, 305, 5)
76     produccion_1= st.slider('Lactancia en muestra día 10',7.0,23.0,13.0,0.5)
77
78     st.subheader('Variables de reproducción')
79     servicios = st.slider('servicios por concepción', 1, 5, 2, 1)
80
81     natimuerto = st.radio('Presencia de Natimuerto',['Si','No'])
82     st.write('<style>div.row-widget.stRadio > div{flex-direction:row;</style>', unsafe_allow_html=True)
83
84     st.subheader('Variables agroecológicas')
85     pre_2antes= st.slider('Precipitación 2 meses antes del parto (mm al mes)', 1,275,78,1)
86     pre_1antes=st.slider('Precipitación 1 mes antes del parto (mm al mes)', 1,275,78,1)
87     pre_mesparto=st.slider('Precipitación en el mes del parto (mm al mes)', 1,275,78,1)
88     pre_1pos=st.slider('Precipitación 1 mes después del parto (mm al mes)', 1,275,78,1)
89     pre_2pos=st.slider('Precipitación 2 meses después del parto (mm al mes)', 1,275,78,1)
90     hums_1antes=st.slider('Humedad del suelo 1 mes antes del parto (% mensual)', 6, 44, 26, 1)
91     hums_mesparto=st.slider('Humedad del suelo en el mes de parto (% mensual)', 6, 44, 26, 1)
92     hums_1pos=st.slider('Humedad del suelo 1 mes después del parto (% mensual)', 6, 44, 26, 1)
93     hums_2pos=st.slider('Humedad del suelo 2 mese después del parto (% mensual)', 6, 44, 26, 1)
94
95     if raza == 'Holstein':
```

## 4.2 Discusión

El uso de modelos de Machine Learning para la predicción de la lactancia se ha limitado a la inferencia de los valores de lactancia ajustada a 305 días en función de los valores de producción de leche diarios, es decir, de la producción diaria, que es expresada por días de toma de muestra establecidos sistemáticamente. Estos modelos sirven para predecir la curva de lactancia y por ende la lactancia total, sin tomar en cuenta los factores que influyen sobre la producción. En este contexto, el algoritmo que mejor poder predictivo tiene es el de redes neuronales, seguido por el de regresión lineal múltiple (Dongre, Gandhi, Singh, & Ruhil, 2012).

Uno de los factores más importantes en el desarrollo de los modelos es la cantidad de datos, debido a la variabilidad de factores subyacentes que se pueden encontrar, en tal sentido se requieren bases de datos con más de 1,000 registros para poder obtener valores coeficiente de determinación sobre el 80%. El incremento de variables en estos modelos también resulta útil para optimizar los modelos, es importante señalar además que los resultados siempre deben ser limitados al alcance de las investigaciones, incluso con grandes bases de datos. Dongre, Gandhi, Singh, & Ruhil en 2012 hacen énfasis en que los algoritmos de Machine Learning utilizados en su investigación se limitan a la inferencia en vacas de la raza Sahiwal en la India. Incluso dentro de un mismo de las regiones de un mismo país y las épocas del año existe diferencia en la estimación de curvas de lactancia como predictores, sobre todo en países de cuatro estaciones (Cole, Null, & De Vries, 2011).

La mayoría de autores sostiene que la influencia de los factores agroecológicos, genéticos, de manejo, sobretodo en sistemas en pastoreo en donde el crecimiento de las praderas está influenciado, y por ende la alimentación animal está influenciado por mucho más factores que en sistemas en confinamiento (Murphy, Mahony, & Shallo, 2014).

Tomando en cuenta las referencias citadas, el desarrollo de un modelo con datos locales es de bastante utilidad, teniendo en cuenta que se alcanzó un coeficiente de determinación de aproximadamente 80% y una razón de raíz del error cuadrático con la media de producción total de 11%.

Adicionalmente, la posibilidad de interacción de los usuarios finales de los datos con el modelo, a través del aplicativo web, cumple la finalidad con la que fue elaborado, debido a que permite ser útil en la toma de decisiones para los administradores del negocio.

A pesar de que el modelo se encuentra limitado a su alcance, en la región Sierra Norte del Ecuador, bajo un sistema de pastoreo rotacional y únicamente para ejemplares de las razas

Holstein y cruce Montbéliarde x Holstein; el modelo puede usarse como referente para regiones con características similares y como fuente de datos exploratorios en estudios posteriores.

## CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES

---

### 5.1 Conclusiones

Se logró desarrollar un aplicativo web en el que se pudo aplicar el algoritmo de Machine Learning de regresión lineal múltiple a los datos de producción lechera de la hacienda el Prado del Instituto Agropecuario Superior Andino (IASA), e identificándose que el modelo fue satisfactorio y puede usarse para la predicción de la producción total de la lactancia en vacas de este tambo, que se ubica en la Sierra Norte del Ecuador, bajo condiciones de pastoreo rotacional y de las razas Holstein y cruce Montbéliarde x Holstein.

Se concluye que el modelo de Machine Learning que utilizó una variable independiente y 17 dependientes que constan de factores reproductivos, productivos, genéticos y ambientales permitió identificar que existe una CORRELACION FUERTE entre las variables ya que se relacionan significativamente por el valor obtenido del R<sup>2</sup> ajustado.

El modelo metodológico utilizado CRISP DM para el desarrollo del modelo predictivo además de resultar ágil permitió sintetizar de manera ordenada y lógica los elementos de exploración de información, planificación, ejecución y evaluación.

### 5.2 Recomendaciones

Es importante hacer evaluaciones periódicas del uso del modelo, así como el ingreso de más datos para entrenamiento y prueba para mejorar la implantación y ajuste del modelo.

Se recomienda la construcción de la curva de lactancia de la predicción a través de inferencia estadística usando modelos establecidos, con el fin de brindar elementos que sean cada vez más útiles para el uso de los datos. Esto es debido a que la curva de lactancia permite una evaluación más periódica y es más familiar para muchos administradores de granjas lecheras, así como

para los colaboradores en el área operativa como ingenieros zootecnistas y veterinarios, esto se puede inferir con modelos cuadráticos y logarítmicos. La traficación del modelo predictivo se puede realizar con librerías de Python como Seaborn, construyendo un dataframe con la información recopilada y las ecuaciones de curva de lactancia.

Se recomienda también el uso de otros algoritmos para la predicción de la producción lechera en futuras investigaciones, sobre todo el de redes neuronales artificiales, que ha sido citado por diferentes autores como el que mejores atributos predictivos tiene.

## BIBLIOGRAFÍA

---

- Allwright, S. (2022). *Stephen Allwright.com*. Obtenido de What is a good R-Squared value? : <https://stephenallwright.com/good-r-squared-value/>
- Alpaydin, E. (2014). *Introducción a Machine Learning*. Londres: MIT Press.
- Bourdon, R. (1999). *Entendiendo el mejorameinto animal*. Harlow: Pearson.
- Buchanan, D. (2011). Animales que producen lacteos. En J. Fuquay, P. Fox, & P. McSweeney, *Enclicopedia de Ciencia Lechera* (págs. 559-568). Londres: Elsiever.
- Bustillo, J., & Melo, J. (2020). *Parámetros reproductivos y eficiencia reproductiva en ganado bovino*. Villavicencio: Universidad Cooperativa de Colombia.
- Cambell, A. (1966). Grazed pasture parameters. I. Pasture dry-matter production and availability in a stocking rate and grazing management experiment with dairy cows. *The Journal of Agricultural Science*, 199-210.
- Cole, J., Null, D., & De Vries, A. (2011). Best prediction of 305-day lactation yields with regional and seasonal effects. *Journal of Dairy Science*, 1601-1604.
- DairyZn. (2020). *New Zeland Dairy Statistics 19-20*. Hamilton: DairyZn.
- Dongre, V., Gandhi, R., Singh, A., & Ruhil, A. (2012). Comparative efficiency of artificial neural networks and multiple linear regression analysis for prediction of first lactation 305-day milk yield in Sahiwal cattle. *Livestock Science*, 192-197.
- Flach, P. (2012). *MACHINE LEARNING: The Art and Science of Algorithms that Make Sense of Data*. Cambridge: Cambridge University Press.

- Hernández, F., & Mazo, M. (2020). *Modelos de Regresión con R*. Bogotá: UNC.
- INEC. (2021). *Encuesta de Superficie y Producción Agropecuaria Continua, 2020*. Quito: INEC.
- McDaniel, B. (2011). Selección: Conceptos. En J. Fuquay, P. Fox, & P. McSweeney, *Enciclopedia de Ciencia Lechera* (Vol. 2, págs. 646-648). Londres: Elsevier.
- Merck Sharp & Dohme. (2020). *MSD Veterinary Manual*. Obtenido de Parameters Useful to Monitor Health and Production of Dairy Herds: <https://www.msdsvetmanual.com/en-au/multimedia/table/parameters-useful-to-monitor-health-and-production-of-dairy-herds>
- Moran, J. (2005). *Producción lechera tropical*. Collingwood: Csiro.
- Murphy, M., Mahony, J., & Shallo, P. (2014). Comparison of modeling techniques for milk-production forecasting. *Journal of Dairy Science*, 3352–3363.
- Piñehira, J., Riverós, J., & Felmer, R. (2009). *Herramientas de última generación para mejoramiento genético animal*. Santiago: INIA.
- Rodríguez, M., Piñeiro, C., & Del Llano, P. (2014). *DEterminación del riesgo de fracaso financiero mediante la utilización de modelos paramétricos de inteligencia artificial, y de infromación de auditoría*. Quito: Estudios Economía.
- Tentor, G. (2020). *Campus FCA-UNC*. Obtenido de Sistemas de producción de leche : <https://www.youtube.com/watch?v=bt2zOhiwWF8>
- Uribe, H. (2012). Objetivos de la Mejora Genetica en Bovinos de Leche. *Informativo INIA Chile*.
- Venturini, S. (2016). *Cross-Validacion for Predictive Analytics*. Obtenido de Milano R: <http://www.milanor.net/blog/cross-validation-for-predictive-analytics-using-r/>

Wignans, G., & Gengler, N. (2011). Selecció: Evaluació y Mètodos. En J. Fuquay, P. Fox, & P. McSweeney, *Enciclopedia de Ciencia Lechera* (Vol. 2, págs. 649-655). Londres: Elsevier.

Yang, Y. (2007). Consistency of cross validation for comparing regression procedures. *The Annals of Statistics*, 2450–2473.

## ANEXOS

Carta de autorización de datos hacienda el prado IASA

### Carta de Autorización de Datos

Sangolquí, 6 de junio de 2022

La hacienda el Prado, del Instituto Agropecuario Superior Andino (IASA), por medio del Ing. Diego Vela Tormen, con CI. No. 1707754535, responsable del Taller de Ganadería, autoriza el uso de datos de la producción lechera del hato bovino, para el proyecto de investigación denominado *"Desarrollo de un modelo predictivo aplicando algoritmos de Machine Learning para la producción lechera en la hacienda el Prado, del Instituto Agropecuario Superior Andino (IASA)"*, al señor Jorge Isaac Ordoñez Merino estudiante de la Maestría de Sistemas de Información – Data Science de la Pontificia Universidad Católica del Ecuador.

Particular que pongo en conocimiento con fines de investigación y estudios para el señor antes mencionado.



Atentamente  
Ing. Diego Vela Tormen  
Responsable del Taller de Ganadería  
Hacienda El Prado - IASA

## Acta entrega recepción del aplicativo web

### Acta entrega recepción del aplicativo web

En Quito, a los 10 días del mes de enero de 2023, con la presencia del Ingeniero Diego Vela Tormen responsable del taller de ganadería del Instituto Agropecuario Superior Andino (IASA) y el Licenciado Jorge Isaac Ordoñez Merino, estudiante de la Maestría de Sistemas de Información Data Science, se procede a la entrega del aplicativo web denominado: "MODELO PREDICTIVO APLICANDO ALGORITMOS DE MACHINE LEARNING PARA LA PRODUCCIÓN LECHERA EN LA HACIENDA EL PRADO, DEL INSTITUTO AGROPECUARIO SUPERIOR ANDINO (IASA)". Mismo que este alojado en el dominio Streamlit, a partir de la suscripción del presente documento será de responsabilidad del IASA el uso correcto del aplicativo; así como también la alimentación de los nuevos datos.

Para validación del presente documento firman abajo los presentes:



Lic. Jorge Isaac Ordoñez Merino

1722807649



Ing. Diego Vela Tormen

1707754535