Modelo de clasificación para la evaluación de Toros

Bárbara Paola Alcántara Vega

Resumen:
El objetivo principal del presente reporte es desglosar el análisis detrás del modelo desarrollado para clasificar el nivel de aceptación para 17 características de conformación lineal en toros dentro de la base de datos de SelectSires para la raza Holstein. El modelo propuesto incluye limitaciones actuales y propuestas de mejora, ya que actualmente el modelo cuenta con una certeza media de 85% para el entrenamiento y prueba de cada una de las clases.

Introducción:
La selección de semen de toro es una práctica común de los establos lecheros para asegurar una progenie con características que maximicen la productividad mientras mantienen la salud del hato lechero. Est esta la razón por la cual empresas como SelectSires ha desarrollado un catálogo extensivo con datos detallados sobre las características físicas y genéticas de cada toro, en el cual los productores lácteos pueden seleccionar los toros que más beneficien la producción en base a características ya presentes dentro de su hato, sin embargo, muchos productores no cuentan con el tiempo o el conocimiento para hacer un análisis extensivo de las características de los toros que podrían elegir, y terminan eligiendo toros que aunque buenos, no maximizan el rendimiento productivo del hato, y en el peor de los casos, si sólo se elige uno o dos toros para todo el hato, terminan con problemas de consanguinidad.

Normalmente, la evaluación y selección de ganado requiere de un experto certificado para la raza con la que se pretende trabajar, pero con los avances recientes en machine learning, es posible automatizar el proceso de evaluación no solo mediante puntaje, sino también con imágenes, permitiendo a los productores realizar una evaluación apropiada de su ganado y de su selección de toros sin contar con experiencia previa de la raza en cuanto a características de conformación lineal.

En este contexto, la selección de toros con machine learning permite clasificar a los toros basándose en su puntaje; muy bajo, bajo, alto y muy alto. Considerando varios atributos como designaciones y lineamientos para adaptarse a las necesidades de cada productor, como la optimización en el uso de alimento forrajero, alimentación con pasto, selección de género para la progenie, y más.

Las aplicaciones de la Inteligencia artificial y el machine learning están volviéndose cada vez más accesibles, permitiendo que los productores tomen decisiones informadas y competitivas con mayor facilidad, no solo permitiendo una mejor selección de material genético, sino también un vistazo a los factores que más les convienen para mejorar su hato lechero en términos de producción y de salud.

Descripción del data set
La base de datos de sementales de toro, la cual sirve como base del proyecto, se deriva del acceso público del catlálogo de sementales que contiene características organizadas y clasificadas por la empresa SelectSires, para la selección e identificación de sus sementales para el comprador e incluso para la toma de decisiones cuando se trata de conseguir nueva

genética para el consumidor. La base de datos se relaciona directamente con los atributos que califican a un semental para diferentes necesidades.

Features

El data set incluye 161 columnas, cada una representando una característica importante del semental promocionado. Sin embargo, con el objetivo de simplificar el análisis, se escogieron 17 características principales que se utilizan para evaluar el ganado de forma universal y sus identificadores principales:

Identificadores principales

- **NAAB:** Número de serie para el semental
- **Lineups/designations:** Clase a la que pertenece el semental

Características principales

- **PTAT:** Profundidad corporal
- **STA:** Estatura
- **STR:** Fuerza
- **DFM:** Forma lechera
- **RUA:** Ángulo de la cadera
- **RLS:** Ángulo de las patas traseras con vista lateral
- **RTP:** Colocación de pezones traseros
- **FTL:** Largo de pezones delanteros
- **RW:** Ancho de cadera
- **RLR:** Ángulo de las patas traseras con vista trasera
- **FTA:** Ángulo de las pezuñas
- **FUA:** Inserción delantera de la ubre
- **RUH:** Altura de la ubre trasera
- **RUW:** Ancho de la ubre trasera
- **UCL:** Ligamento medio suspendido
- **UDP:** Profundidad de la ubre
- **FTP:** Colocación de pezones delanteros

Adicionalmente, es importante tomar en cuenta que cada característica cuenta con un porcentaje de heredabilidad y con un nivel de importancia en la salud promedio y producción de leche.

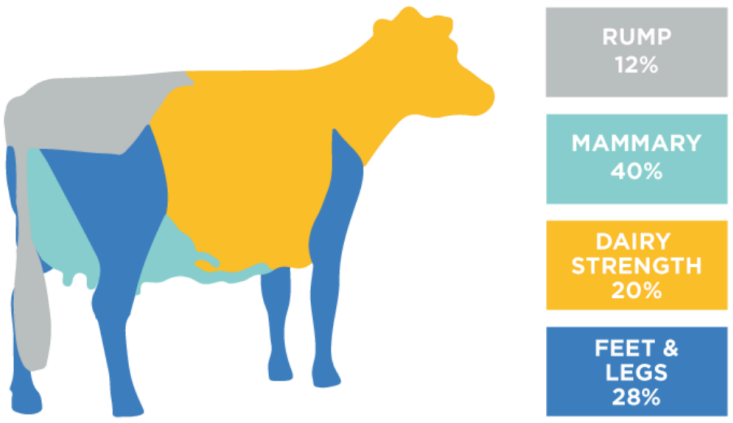

Figura 1. Importancia de cada grupo de categorías en la salud y producción de leche (*Holstein Canada: Services - Classification*, 2015)

Data cleaning:
La base de datos contiene 161 features, y no todos los datos que contienen son numéricos. Por lo que lo primero que se tuvo que hacer es permutar las columnas que no proporcionaban información relevante o clara, para poder trabajar con datos numéricos que puedan ser utilizados en el aprendizaje y entrenamiento del modelo. Además, aunque las instancias de la base de datos ya se encuentran de forma aleatoria, es necesario una aleatorización entre cada entrenamiento y prueba con el objetivo de que el modelo no se ajuste incorrectamente a los datos y pueda predecir puntajes sin importar el orden de los datos proporcionados.

(si lo logro) (claro que lo logras, ya lo hiciste en excel el año pasado, solo tradúcelo a código)
Además, se llevó a cabo una normalización de los datos mediante la siguiente ecuación

$$X_{\text{norm}} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

Para mantener el rango de cada característica en una escala estándar, ya que cada característica cuenta con datos cuyo rango es variable con respecto a los demás. Esta estabilidad numérica asegura que cada característica se encuentra dentro de una escala comparable y así se evita que los rangos más altos dominen en el proceso de entrenamiento.

Análisis de correlación y selección de características

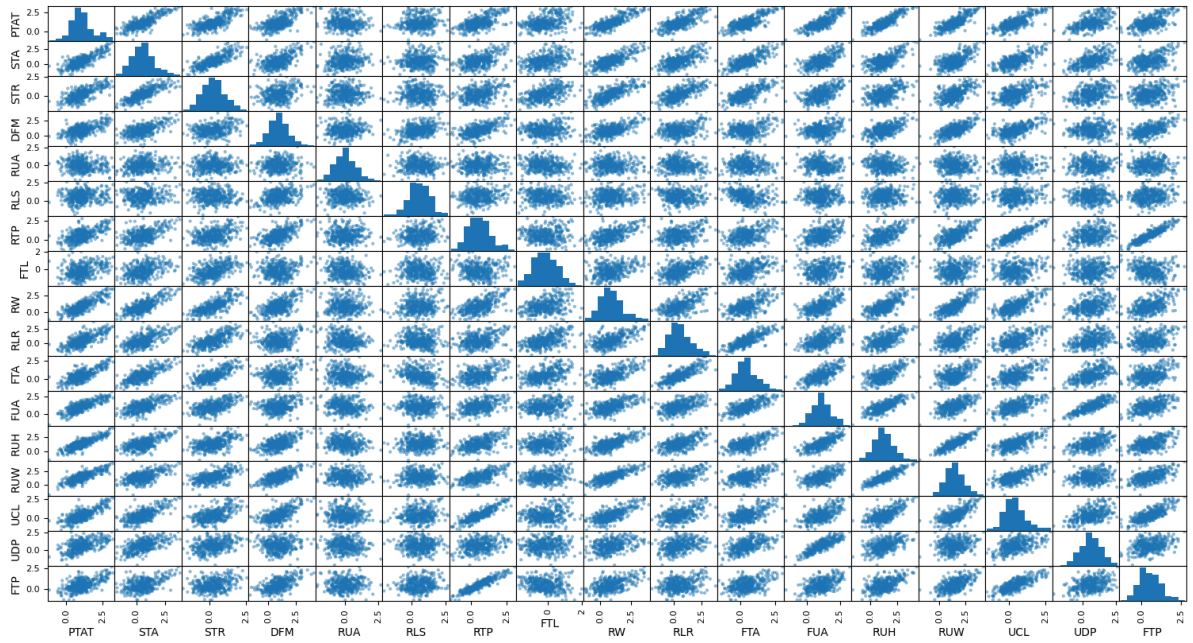

Figura 1. Scatterplot de las 17 características seleccionadas.

Gracias a un scatterplot de cada característica de conformación lineal, podemos observar que la relación entre datos involucrados con ubres y patas están estrechamente relacionados. Correlación que tiene sentido al ser dos características dependientes para el desarrollo de una buena salud mamaria en la raza Holstein. Esta es la razón por la cual los ganaderos tienden a enfocarse en estas características, sin embargo, aunque ayude mucho en términos de productividad, únicamente enfocarnos en ubres y patas podría resultar en otros problemas, como falta de fortaleza y la accidental omisión del tendón suspensorio, dos características muy importantes para la salud de los animales y una mayor cantidad de días productivos. Esta es la razón por la cual no se mantuvieron características que describen las ppm de sólidos en leche, el pedigree, o dólares por litro de leche en promedio, ya que además de que la raza Holstein es bien conocida como productora de leche fluida y no de productos lácteos, el enfoque en salud y productividad es una prioridad en la mejora de producción lechera actual a nivel nacional en México.

Data set split

Dividir la base de datos en tres partes permite un entrenamiento, validación y prueba de los resultados. Para que esto funcione, es importante que las proporciones de esta división sean del 60%, 20% y 20% respectivamente para cada parte. Todo eso con el objetivo de entrenar el modelo con la mayor cantidad de datos disponibles posible y consecuentemente, evaluar el modelo con respecto a su rendimiento como si se le presentaran datos nuevos, simulando de esta manera, una selección en un escenario real.

Modelo base

El proyecto emplea un modelo de regresión lineal construido desde sus elementos básicos, sin utilizar librerías existentes de machine learning con el objetivo de profundizar el entendimiento del funcionamiento de estos procesos durante el análisis de los datos. Este modelo es utilizado principalmente como herramienta de selección binaria, pero también puede utilizarse para una clasificación multiclase.

Debido a la cantidad de clases que necesitan predecirse, el ciclo debe repetirse por cada una de las 12 clases. En cada ciclo, el modelo emplea la estrategia llamada One-vs-Rest

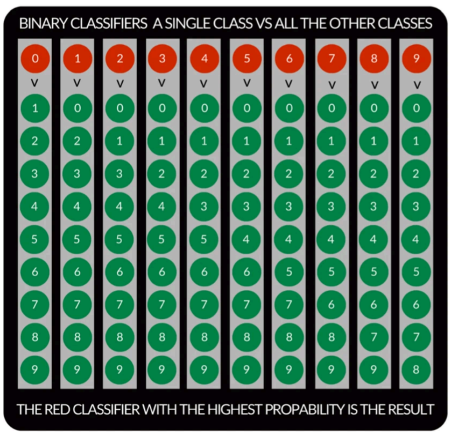

Figura 2. One-vs-Rest funcionamiento básico (Psaltakis et al., 2024)

Este enfoque implica entrenar un clasificador binario para cada clase, donde el clasificador de una clase específica se entrena para distinguir esa clase de todas las demás clases combinadas (*OneVsRestClassifier*, 2025). De este modo, en vez de intentar predecir las 12 clases al mismo tiempo, se enfoca en una sola clase para tratar al resto de las 11 como una

clase por sí misma, con el objetivo de determinar si el dato pertenece a esa clase mediante la aplicación de la función de hipótesis. De este modo, tras haber repetido el proceso para tres clases, encuentra el índice de la clase con la mayor probabilidad de corresponder con el dato analizado.

Función de hipótesis

La función de hipótesis es utilizada para predecir la probabilidad de que un dato de entrada proveniente de los datos, pertenece a una clase o combinación de clases en específico. En este caso, se utilizó la función sigmoide.

$$h\theta(x) = \frac{1}{1+e^{-\theta Tx1}}$$

Debido a que su dominio se mantiene en todos los números reales, y su rango es de (0,1), por lo que incluso si se recibe un número positivo o negativo de gran magnitud, el resultado siempre será entre 0 y 1. Es por esta misma razón que el método de One-vs-Rest es apto para la selección, ya que solo puede determinar entre 0 y 1.

Función de costo

La función de costo cuantifica la diferencia entre las probabilidades predichas y la clase actual a la que pertenecen, por lo que si minimizamos esta diferencia durante el entrenamiento, permitimos que el modelo se ajuste a sus parámetros para mejorar la certeza. Ya que contamos con 12 clases, la función de costo se utiliza para cada una de ellas de forma independiente.

$$J(\theta) - \frac{1}{m}\sum_{i=1}^{m}[y^{(i)}log(h\theta(x^{(i)})) + (1 - y^{(i)})log(1 - h\theta(x^{(i)})$$

Gradiente descendente

La gradiente descendente permite que el modelo aprenda de los datos mediante la minimización de la función de costo. Lo que quiere decir que primero se calcula la gradiente de la función de costo con respecto a los parámetros (theta); posteriormente, se actualiza cada parámetro en la dirección que reduce el costo, incrementando así la precisión de las predicciones.

$$\Theta_j := \Theta_j - \frac{\alpha}{m}\sum_{i=1}^{m}[(h_\Theta(x_i) - y)x_i]$$

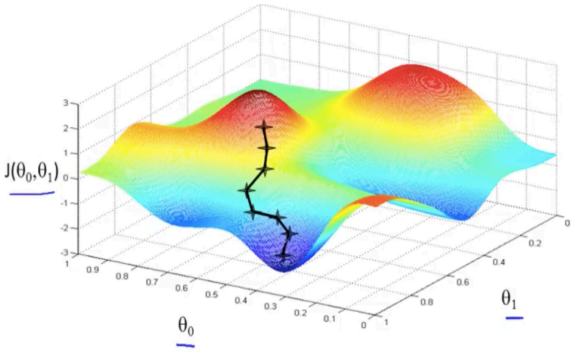

Figura 3. Resumen ilustrativo de la gradiente descendente (DBCerigo, 2018).

Back propagation (si es que logro implementarlo correctamente)

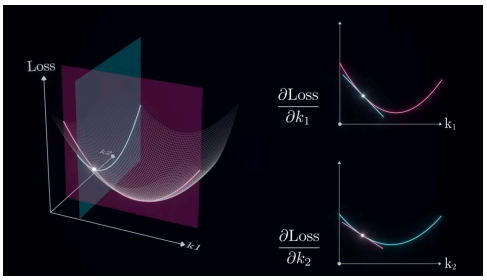

Figura x. Resumen ilustrativo del BackPropagation (Artem Kirsanov, 2024).

Ando viendo que no lo necesito… funciona bien con una sola neurona y regresión logística. Un modelo más complejo podría beneficiarse mucho más de un back propagation que esta demostración de conocimientos teóricos y prácticos.

Entrenamiento, validación y prueba

El modelo aprende mediante el ajuste de parámetros mediante el gradiente descendente, minimizando el error en las predicciones por cada época. La estrategia del One-vs-Rest es utilizada para entrenar el modelo de forma separada a cada clase, mientras el costo es monitoreado para observar el progreso.

Además, el comportamiento y avance del modelo también se evalúa con un set de validación separado de los datos de entrenamiento. Esto con el objetivo de registrar que tan bien generaliza el modelo a datos nunca antes vistos. De este modo, monitorear la certeza de validación permite que el modelo se ajuste mientras se previene el overfitting durante el entrenamiento.

Posterior al entrenamiento, el modelo es probado con datos completamente nuevos (el set de prueba), para evaluar su habilidad de generalización, mientras que la prueba final de certeza nos indica que tan probable es que el modelo se comporte correctamente en un escenario real.

Resultados del modelo base (retroalimentación o avance)

Después de entrenar el modelo y probarlo, los resultados muestran que el modelo alcanzó una certeza promedio del 85% en el entrenamiento.
Además, se puede observar una reducción del error a la mitad tras 300 épocas,

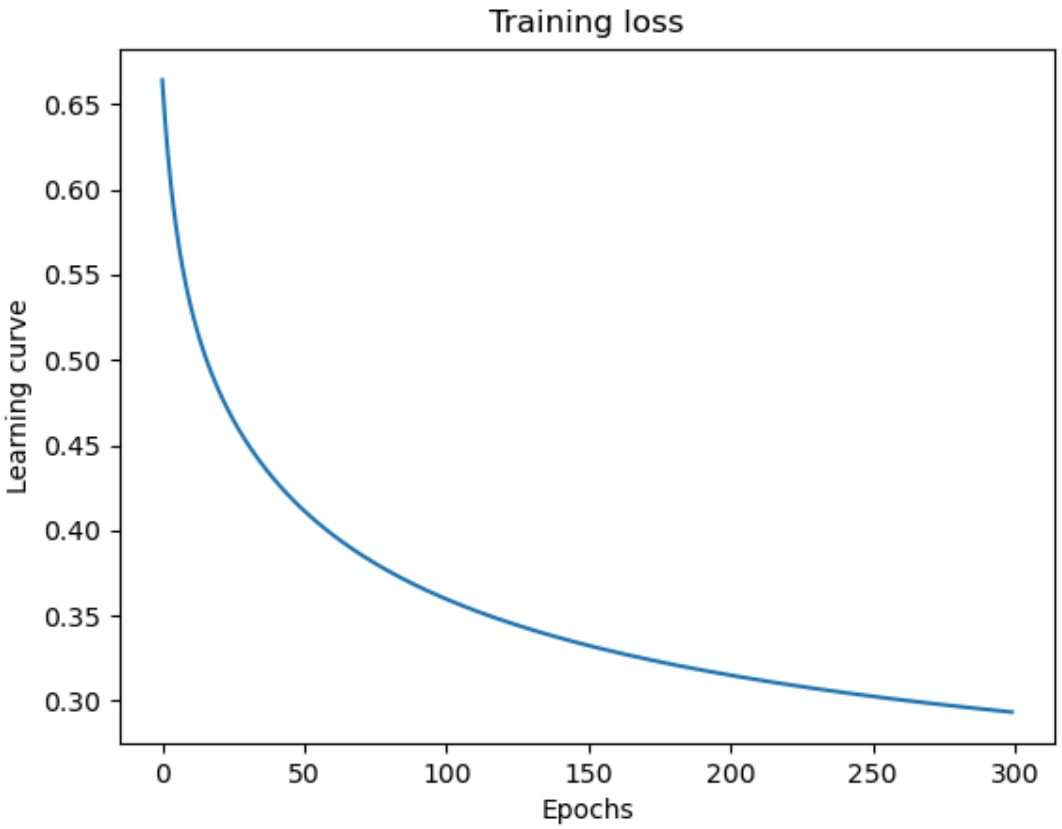

Figura 4. Gráfica de pérdida promedio.

El error inicia relativamente alto alrededor de 0.66, y a medida que aumentan las épocas, la pérdida desciende hasta estabilizarse cerca de 0.29.
Ya que la curva no cuenta con oscilaciones ni aumentos bruscos, el modelo aprende y reduce su error con cada iteración sin presentar signos evidentes de overfitting (sin aumento en la pérdida/error). La convergencia es progresiva con una mejora marginal alrededor de las 250-300 épocas, lo que significa que entrenar durante más épocas aportaría una mejora mínima del error.

En cuanto a los resultados de certeza para cada una de las clases:

Feature	Accuracy
PTAT	0.898
STA	0.874
STR	0.882
DFM	0.840
RUA	0.715
RLS	0.785
RTP	0.914

FTL	0.777
RW	0.835
RLR	0.825
FTA	0.869
FUA	0.898
RUH	0.869
RUW	0.887
UCL	0.882
UDP	0.911
FTP	0.882
Mean accuracy:	**0.855**

La gráfica de entrenamiento en conjunto con nuestras certezas observadas, nos indica que no hay presencia de overfitting. Afortunadamente ninguna de las características rebasa el 95%, lo cual indicaría un nivel sospechoso de certeza, aunque sí es importante notar que el largo de los pezones delanteros (FTL) no cuenta con un nivel de certeza muy alto. Esto tiene sentido, ya que el largo de los pezones no suele influir en las clases, a excepción de RobotPRO, en donde el largo de los pezones facilita una ordeña robotizada, pero no la determina, lo que significa que tener pezones largos o cortos no impide por completo una ordeña robotizada.

Referencias

OneVsRestClassifier. (2025). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.multiclass.OneVsRestClassifier.html

Psaltakis, G., Rogdakis, K., Loizos, M., & Kymakis, E. (2024). One-vs-One, One-vs-Rest, and a novel Outcome-Driven One-vs-One binary classifiers enabled by optoelectronic memristors towards overcoming hardware limitations in multiclass classification. *Discover Materials*, *4*(1). https://doi.org/10.1007/s43939-024-00077-7

Select Sires. (2024). Default. https://www.selectsires.com/home

3Blue1Brown. (2017, October 16). *Gradient descent, how neural networks learn | Deep Learning Chapter 2* [Video]. YouTube. https://www.youtube.com/watch?v=IHZwWFHWa-w

DBCerigo. (2018). *talks-presentations/deep_learning_without_hype/On Why Gradient Descent Is Needed - Without The Hype.ipynb at master · DBCerigo/talks-presentations*. GitHub. https://github.com/DBCerigo/talks-presentations/blob/master/deep_learning_without_hype/On%20Why%20Gradient%20Descent%20Is%20Needed%20-%20Without%20The%20Hype.ipynb

Holstein Association USA, the world's largest dairy cattle breed association. (2024). https://www.holsteinusa.com/

Holstein Canada: Services - classification. (2015). Holstein Canada. https://www.holstein.ca/Public/en/Services/Classification/Breakdown_of_Traits