Análisis exploratorio y visualización de datos de color de capa en caballos peruanos de paso

Autores: Julexi Zambrano Barre, Jessica Morales-Carrillo, Luis Cedeño-Valarezo, Carlos Larrea Izurieta.

Escuela Superior Politécnica Agropecuaria de Manabí Manuel Félix López, Calceta, Manabí, Ecuador, Grupo de Investigación SISCOM, Carrera de Computación.

Escuela Superior Politécnica Agropecuaria de Manabí Manuel Félix López, Calceta, Manabí, Ecuador, Grupo de Investigación SISCOM, Carrera de Medicina Veterinaria.

Correo: julexi.zambrano@espam.edu.ec, jmorales@espam.edu.ec,lcedeno@espam.edu.ec, clarrea@espam.edu.ec

DOI: https://doi.org/10.56124/encriptar.v7i14.004

Resumen

Esta investigación presenta un estudio exhaustivo sobre los patrones genéticos que determinan el color del pelaje de caballos. Se utilizó la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), el estudio se llevó a cabo en varias fases, desde la comprensión del problema hasta el despliegue de los resultados. Los datos analizados provienen de la Asociación Nacional Ecuatoriana de Criadores y Propietarios de Caballo de Paso (A.N.E.C.P.C.P.), abarcando registros desde 1933 hasta 2020, aunque para el análisis de cruce de pelajes se filtraron los datos desde 1985 hasta 2020 lo cual redujo la data a 6131 registros. Los resultados obtenidos indican predominancia de pelajes ALAZÁN y CASTAÑO, en comparación con otros colores menos frecuentes, indica una fuerte influencia de los genes de Extension (E) y Agouti (A), que determinan la mayoría de los colores observados de la data analizada.

Palabras-clave: Patrones genéticos; color del pelaje; caballos peruanos de paso; metodología CRISP-DM.

Exploratory analysis and visualization of coat color data in Peruvian Paso horses

Abstract

This research presents an exhaustive study on the genetic patterns that determine horse coat color. The CRISP-DM (Cross Industry Standard Process for Data Mining) methodology was used; the study was carried out in several phases, from understanding the problem to displaying the results. The data analyzed comes from the Ecuadorian National Association of Paso Horse Breeders and Owners (A.N.E.C.P.C.P.), covering records from 1933 to 2020, although for the analysis of crossbreeding the data was filtered from 1985 to 2020, which reduced the data to 6131 records. The results obtained indicate a predominance of CHURCH and CHESTNUT coats, compared to other less frequent colors, indicating a strong influence of the Extension (E) and Agouti (A) genes, which determine the majority of the colors observed in the data analyzed.
Keywords: Genetic patterns; coat color; Peruvian Paso horses; CRISP-DM methodology.

1. INTRODUCCIÓN

En la actualidad la expansión de datos complejos provenientes de diversos ámbitos impulsa la necesidad de transformar este volumen en información útil. La minería de datos, parte del proceso de descubrimiento de conocimiento en bases de datos, se enfoca en identificar patrones en grandes conjuntos de información (Medina et al., 2020; López, 2021). la minería de datos utiliza una variedad de técnicas, como regresión lineal, redes neuronales, árboles de decisión y análisis de asociación (Rojas, 2020; Flores et al., 2019). Dentro de las diversas herramientas empleadas en la minería de datos, la Inteligencia Artificial (IA) ha demostrado ser altamente efectiva, logrando resolver una gran cantidad de problemas y desafíos (Tortosa, 2020). Asimismo, García et al. (2020) mencionan que la IA ha sido fundamental en este campo, revolucionando el análisis de datos y generando un impacto

significativo en la educación. El rápido avance de este progreso condujo a lo que conocemos como Aprendizaje Automático o "Machine Learning" (Gómez William, 2023).

El Aprendizaje Automático, parte clave de la IA, se centra en cómo las máquinas pueden aprender de los datos disponibles (Gómez, 2020; Alfaro & Ospina, 2021). Esta disciplina se integra en la minería de datos para identificar patrones ocultos en conjuntos extensos de información (Sánchez, 2021; Ramírez, 2021). Las técnicas de visualización se han adaptado para explorar conjuntos multidimensionales y han resultado útiles en la creación de modelos de aprendizaje automático, lo que amplía la cantidad de información obtenida en análisis exploratorio (Valencia et al., 2020; Mohedano, 2021).

El Análisis Exploratorio de Datos (AED) es una etapa inicial en el estudio de datos, donde se aplican técnicas específicas para estudiar y extraer información (Fuentes, 2022). Se caracteriza por utilizar herramientas visuales, gráficas y semi-gráficas para explorar los datos antes de iniciar un análisis más profundo (Troya, 2019).

El caballo peruano de paso, descendiente de caballos introducidos durante la Conquista, ha ganado reconocimiento internacional por su singularidad en belleza y funcionalidad (Montalván & Rojas, 2019). La selección en criadores de caballos ha impactado la diversidad de colores en estas criaturas, influenciada por genes específicos que determinan tonalidades y patrones en su pelaje (Bolaños, 2020). El matiz en el color de los equinos no surge por casualidad; La información relativa al color que será transmitida a la descendencia está meticulosamente registrada en los genes. La totalidad de estos genes se conoce como GENOTIPO, y la manifestación de dichos genes en el individuo se denomina FENOTIPO. Cada gen está compuesto por dos unidades llamadas ALELOS: uno proviene del padre y el otro de la madre. Estos alelos, a su vez, pueden clasificarse como dominantes o recesivos.

La investigación tiene como objetivo principal examinar y comprender los diversos patrones genéticos, tanto en términos de genotipo como de fenotipo, que determinan la coloración de estos ejemplares. Este análisis proporcionará una base sólida para futuras investigaciones en el campo de la genética equina y ofrecerá información valiosa para mejorar la crianza y gestión de los caballos.

2. MATERIALES Y MÉTODOS

Para el desarrollo de esta investigación se empleó la metodología CRISP-DM (Cross Industry Estándar Process for Data Mining: Procedimiento Industrial Estándar para realizar Minería de Datos) como guía para la ejecucion del proyecto. Esta metodología comprende seis fases y abarca desde la comprensión inicial del problema hasta el despliegue de los resultados (Cardona, 2021; Huber et al., 2019), Basado en los componentes de la metodología, se utilizaron los principios de CRISP-DM con adaptaciones específicas para el presente proyecto. A continuación, se detalla el proceso de desarrollo de cómo se abordó cada una de las fases.

2.1 Comprensión del negocio. Constituye el punto de partida esencial para entender los objetivos y requisitos del proyecto desde la perspectiva del negocio.

2.2 Estudio y comprensión de los datos. Aquí se recopilan y reúnen los datos de la fuente original, se describen los datos, se realiza una pre-exploración y se verifica su calidad.

2.3 Preparación de los datos. Implica limpiar, transformar y preprocesar los datos para asegurar su idoneidad para el análisis.

2.4. Modelado. En la fase de modelado se definen el tipo de análisis que se quiere realizar y se definen las variables a ser analizadas.

2.5. Evaluación. La fase de evaluación permite analizar los resultados obtenidos y definir los principales resultados alcanzados.

2.6 Despliegue. La fase de despliegue se centra en la presentación del informe que contiene el análisis de los resultados obtenidos.

Consideraciones éticas. - Para proteger a la asociación, se garantizó que los procedimientos de recolección y manejo de datos cumplieran con las regulaciones de privacidad y los estándares éticos pertinentes. Cabe destacar que, para preservar la privacidad y confidencialidad, se omitieron ciertos datos en el informe.

3. RESULTADOS Y DISCUSIÓN

3.1 Comprensión del negocio

La problemática radica en la ausencia de un análisis detallado sobre las capas de los caballos peruanos de paso, lo que impide la identificación de patrones genéticos que

determinan el color de la capa, prevalencia de colores específicos, y posiblemente, la comprensión de factores geográficos o genéticos asociados.

Por esta razón, en este trabajo se incluye un análisis para caracterizar fenotípicamente a cada animal y asociar los genes implicados en el color del pelaje. Este conjunto de datos proviene de un registro genealógico de caballos peruanos de paso, con el objetivo de estudiar el comportamiento de diferentes patrones genéticos, tanto fenotípicos como genotípicos, que determinan el color de la capa e identificar con precisión los colores. La data fue proporcionada por la Asociación Nacional Ecuatoriana de Criadores y Propietarios de Caballo de Paso (A.N.E.C.P.C.P.), la misma que fue fundada en 1985 y se registraron los primeros caballos peruanos de paso nacidos en Ecuador. La base de datos abarca registros desde 1933 hasta 2020 y contiene un total de 8426 registros de animales.

3.2 Estudio y Comprensión de los datos

Una vez adquiridos los datos, se llevó a cabo una exploración exhaustiva de la base de datos utilizando la plataforma Google Colab con Python y las librerias pandas y matplotlib.pyplot. Este proceso incluyó la evaluación de la calidad de los datos y la comprensión de su estructura y contenido. Se realizó una breve descripción de los datos, que abarca los nombres de las variables, los tipos de datos y una descripción concisa de cada variable. Tal como se detalla en la tabla 1 la data original esta distribuida en trece columnas. De estas variables, una es numérica, mientras que las restantes son categóricas.

Tabla 1. *Variables de la dataset original.*

VARIABLES	DESCRIPCIÓN	VALORES
No.	Identificador único	Numérico
Nombre	Nombre de cada animal	Categórico
Reg.Aso	Registro de asociación	Categórico
Sexo	Género del equino	Categórico
Pelaje	Color de capa (fenotipo)	Categórico
F.N.	Fecha de nacimiento	Categórico
Padre	Parentesco familiar	Categórico
Reg.Padre	Registro del padre	Categórico
Madre	Parentesco familiar	Categórico
Reg.Madre	Registro de la madre	Categórico

Criador	Criador del equino	Categórico
Propietario	Propietario del equino	Categórico
Origen	Lugar de procedencia	Categórico

Fuente: La autora (2024)

3.3 Preparación de los datos

Para garantizar la calidad de los datos, se implementaron varias estrategias de depuración:

a) Se realizó una revisión exhaustiva para detectar y manejar valores nulos y faltantes.

b) Se identificaron y corrigieron los registros duplicados para evitar redundancias,

c) Se realizó una limpieza de datos para corregir errores y eliminar inconsistencias,

d) Se crearon indicadores para identificar preferencias de tipos de pelaje y analizar proporciones de sexo,

e) Se transformaron los datos para codificar numéricamente las variables y normalizarlas según sea necesario.

f) Se realizo un filtrado de datos para asegurar la precisión y relevancia del análisis.

Tal como se oberva en la tabla 2, se descartaron los campos que comprometían la integridad de la asociación, cuyas variables eran registro de padre y madre, registro de asociación, también el criador y propietario del equino. En cuanto a la transformación de datos, se modificaron las variables nombre y padre, ya que contenían caracteres categóricos; a cada variable se asignó un valor numérico para facilitar su integración en la minería de datos. Una vez finalizada la limpieza y transformación de los datos, se almacenaron en una nueva base de datos, lo cual redujo el número de variables a utilizar en el análisis.

Tabla 2. *Variables de la dataset depurada.*

VARIABLES	DESCRIPCIÓN	VALORES
No.	Identificador único	Numérico
Sexo	Género del equino	Categórico
Pelaje	Color de capa (fenotipo)	Categórico
F.N.	Fecha de nacimiento	Datetame
Padre	Parentesco familiar	Numérico
Madre	Parentesco familiar	Numérico
Origen	Lugar de procedencia	Categórico

Fuente: La autora (2024)

Con los datos depurados, se generaron varios puntos clave: la frecuencia de tipos de fenotipo para identificar preferencias, los indicadores geográficos para comprender la distribución regional de los animales. el análisis de cruce de pelaje basado en la información de los padres y madres, la asociación de fenotipos con su probable genotipo. La tabla 3 muestra una visión clara de la distribución de pelajes, destacando tanto las frecuencias como los porcentajes de cada tipo de fenotipo.

Tabla 3. *Distribución de pelajes.*

FENOTIPOS	FRECUENCIA	PORCENTAJE
ALAZÁN	4365	51,80%
CASTAÑO	1391	16,51%
NO IDENTIFICADOS	1061	12,59%
ZAINO	618	7,33%
PALOMINO	418	4,96%
TORDILLO	169	2,01%
NEGRO	162	1,92%
BAYO	124	1,47%
MORO	43	0,51%
ROSILLO	34	0,40%
RUANO	10	0,12%
PERLINO	9	0,11%
OVERO	7	0,08%
CERVUNO	6	0,07%
SABINO	3	0,04%
CREMELLO	3	0,04%
BLANCO	2	0,02%
LOBUNO	1	0,01%
TOTAL	8426	100%

Fuente: La autora (2024)

Al analizar los datos, observamos que hay un gran desbalanceo de datos, donde el color de pelaje ALAZÁN es más frecuente, con 4365 casos que representan el 51.80% de la población total. Este dato indica que más de la mitad de los caballos tienen este color de pelaje, lo que sugiere una fuerte predisposición genética hacia este color de pelaje en la

población estudiada. El siguiente pelaje más frecuente es el CASTAÑO, que cuenta con 1391 casos, representando el 16.51% de la población. Aunque significativamente menos común que el ALAZÁN, sigue siendo un pelaje notablemente prevalente.

Además, se destaca un grupo considerable de caballos cuyo color de pelaje no ha sido identificado, sumando 1061 casos, lo que constituye el 12.59% de la población. Otros pelajes moderadamente comunes incluyen el ZAINO, con 618 casos (7.33%), y el PALOMINO, que representa el 4.96% con 418 casos. El TORDILLO, aunque menos frecuente, tiene una presencia con 169 casos, representando el 2.01%. Entre los pelajes menos comunes, encontramos el pelaje NEGRO con 162 casos (1.92%) y el BAYO con 124 casos (1.47%). Pelajes como el MORO y el ROSILLO, aunque menos frecuentes, todavía tienen cierta representación con 43 y 34 casos respectivamente, representando el 0.51% y el 0.40% de la población. Los pelajes RUANO, PERLINO, OVERO, CERVUNO, SABINO, y CREMELLO son bastante raros, con frecuencias que oscilan entre 10 y 3 casos, cada uno representando menos del 0.12% de la población. Finalmente, los pelajes más raros en esta población son el BLANCO con solo 2 casos (0.02%), y el pelaje LOBUNO con solo 1 caso (0.01%).

En términos generales, la frecuencia total de caballos es mayor en las hembras en comparación con los machos, representando el 59.44% y el 40.56% respectivamente. Esto evidencia una predominancia del sexo femenino en el registro. A continuación, en la tabla 4 se presenta la distribución de pelaje por sexo para analizar comportamientos relacionados.

Tabla 4. *Distribución de pelaje por sexo.*

FENOTIPOS	HEMBRA	MACHO
ALAZAN	29,77%	22,02%
BAYO	0,83%	0,64%
CASTAÑO	10,04%	6,47%
NEGRO	1,16%	0,76%
NO IDENTIFICADOS	8,06%	4,53%
OTROS	0,93%	0,50%
PALOMINO	2,85%	2,11%
TORDILLO	1,53%	0,47%
ZAINO	4,27%	3,06%

Fuente: La autora (2024)

El pelaje ALAZÁN es el más frecuente tanto en hembras como en machos. El pelaje CASTAÑO también es común en ambas categorías de sexo. Un aspecto destacado de la gráfica es la alta frecuencia de pelajes no identificados, especialmente en las hembras. Pelajes moderadamente comunes como el ZAINO, NEGRO, PALOMINO y TORDILLO también tienen una presencia considerable en ambos sexos. Los pelajes agrupados en la categoría "OTROS" (que incluyen MORO, ROSILLO, RUANO, PERLINO, OVERO, CERVUNO, SABINO, CREMELLO, BLANCO, y LOBUNO), tienen una representación menor en ambas categorías de sexo. La baja frecuencia de estos pelajes indica que son raros dentro de la población equina estudiada, pero su presencia añade diversidad genética. A continuación, la tabla 5 muestra la distribución de frecuencias de sexo por origen.

Tabla 5. *Distribución de sexo por origen.*

ORIGEN	HEMBRA	MACHO
ECU	32,76%	26,16%
PER	26,35%	14,28%
OTHER	0,33%	0,12%

Fuente: La autora (2024)

Al comparar entre orígenes, se observa que Ecuador (ECU) es el origen con la mayor cantidad de registros, liderando con el 58,92% de los caballos, y mostrando una clara predominancia de hembras. En este país se registran 2760 hembras lo que equivale a 32,76% y 2205 machos lo que equivale a 26,17%. Por otro lado, Perú (PER) es el segundo origen en términos de frecuencia, con el 40,63% de los caballos, también con una predominancia de hembras, con 2220 hembras lo que equivale a 26,35% y 1203 machos, que equivale a 14,28%. En contraste, la frecuencia de caballos en la categoría "OTHER" que incluye registros provenientes de Panamá (PAN), Guatemala (GUA), Estados Unidos (USA), y Honduras (HON, es muy baja, casi imperceptible en la tabla, con un total de 0,45%, lo que indica una menor representación de estos orígenes. A continuación, en la tabla 6 se presenta la distribución de frecuencias de pelajes en caballos según su origen.

Tabla 6. *Distribución de pelaje por origen.*

FENOTIPOS	ECU	OTHER	PER
ALAZAN	3015	4	1344
BAYO	97	0	27

CASTAÑO	946	3	442
NEGRO	87	0	75
NO IDENTIFICADOS	1	28	1032
OTROS	48	1	71
PALOMINO	318	1	99
TORDILLO	57	0	112
ZAINO	396	1	221

Fuente: La autora (2024)

El pelaje ALAZÁN es el más frecuente en todos los orígenes, indicando su predominancia en la población equina. En Ecuador, el pelaje ALAZÁN registra un total de 3,015 caballos, representando una mayoría clara. Perú sigue con 1,344 caballos de pelaje ALAZÁN, y en la categoría "OTHER" se registran 4 caballos con este pelaje. El pelaje CASTAÑO también es común en todos los orígenes, aunque en menor medida que el ALAZÁN. Ecuador muestra 946 caballos de pelaje CASTAÑO, mientras que Perú registra 442. En la categoría "OTHER", hay 3 caballos de este pelaje. Los pelajes ZAINO, PALOMINO, NEGRO y BAYO son más comunes en Ecuador, mientras que los pelajes NO IDENTIFICADOS, TORDILLO y OTROS son más prevalentes en Perú, sugiriendo mayor diversidad o menos precisión en la identificación de pelajes en Perú.

3.4 Modelado

Para el análisis de cruce de pelajes, se decidió trabajar con registros de los años 1985 a 2020, debido a un gran desbalanceo en el conjunto de datos. El proceso incluyó un riguroso filtrado, seleccionando únicamente los registros con información completa, es decir que solo se consideraron aquellos registros con identificación de ambos padres y color de pelaje especificado. Se descartaron todos los registros incompletos, como aquellos con ambos padres identificados, pero sin registro de pelaje, o con un solo padre o madre sin identificación de pelaje, así como aquellos con registro de color de pelaje, pero con solo un progenitor identificado. Este filtrado redujo el total a 6,131 registros. Para realizarlo, se ejecutó el código mostrado en la figura 1.a.

Figura 1. Código de filtrado de datos

```python
# Filtrar los datos entre 1985 y 2020
filtered_data = data[(data['F.N.'] >= '1985-01-01') & (data['F.N.'] <= '2020-12-31')]

# Filtrar los registros con ambos padres y pelaje identificado
filtered_data_identified = filtered_data[
    (filtered_data['numero_madre'] != 0) &
    (filtered_data['numero_padre'] != 0) &
    (filtered_data['Pelaje'] != 'NO IDENTIFICADOS')
]
```

Figura 1.a. *Código para filtrado de datos.*

```python
# Filtrar las combinaciones con un total de descendencia de 100 o más
result_df_filtered = result_df[result_df['Total_Descendencia'] >= 100]

# Mostrar la tabla de resultados filtrada
from IPython.display import display
display(result_df_filtered)
```

Figura 1.b. *Código para selección de descendencia.*

Fuente: La autora (2024)

De los datos filtrados, se seleccionaron únicamente las combinaciones de pelajes que contaban con más de 100 descendientes. Esta selección permitió enfocarse en combinaciones con suficiente representatividad estadística para un análisis robusto de las probabilidades de cruces con el código de la figura 1.b.

Los detalles específicos de estas combinaciones de cruce se presentan en la Tabla 7.

Tabla 7. *Cruces de pelaje.*

PADRE		ALAZÁN	ALAZÁN	ALAZÁN	ALAZÁN	CASTAÑO	CASTAÑO	PALOMINO	ZAINO	ZAINO
MADRE		ALAZÁN	CASTAÑO	PALOMINO	ZAINO	ALAZÁN	CASTAÑO	ALAZÁN	ALAZÁN	CASTAÑO
DESCENDENCIA	ALAZAN	2318	293	148	80	319	49	131	164	16
	BAYO	3	1	3	0	0	0	1	2	2
	CASTAÑO	54	225	4	86	272	123	1	103	50
	NEGRO	2	8	0	6	11	1	0	7	6
	OTROS	2	2	0	0	0	1	2	0	1
	PALOMINO	23	2	154	2	2	1	109	2	1
	TORDILLO	1	2	0	0	0	0	1	2	0
	ZAINO	15	36	0	55	66	29	2	82	36
TOTAL		2418	569	309	229	670	204	247	362	112

Fuente: La autora (2024)

Considerando que el sexo del progenitor no influye en el color del pelaje de los descendientes, se unificaron los cruces de pelaje, es decir, no importa si el macho es ALAZÁN y la hembra es CASTAÑO o viceversa. Esta misma lógica se aplicó a todos los cruces con el mismo color de pelaje, independientemente del sexo del progenitor, manteniendo cada una de sus probabilidades de colores de pelaje en los descendientes. Esta unificación se realizó ejecutando el código mostrado de la figura 2.

Figura 2. *Código para filtrado de datos.*

```
# Recalcular las probabilidades después de combinar los datos, solo para columnas numéricas
numeric_columns = unified_result.select_dtypes(include='number').columns
total_offspring_unified = unified_result['Total_Descendencia']  # Sumar la descendencia total
probabilities_unified = unified_result[numeric_columns].div(total_offspring_unified, axis=0)  # Calcular las probabilidades

# Redondear las probabilidades a dos decimales
probabilities_unified = probabilities_unified.round(2)

# Crear un DataFrame para almacenar los resultados unificados con las probabilidades recalculadas
unified_result_final = unified_result[['Madre', 'Padre', 'Total_Descendencia']].copy()
for col in probabilities_unified.columns:
    unified_result_final[f'Prob_{col}'] = probabilities_unified[col]  # Añadir las columnas de probabilidad al DataFrame final

# Mostrar el resultado unificado con las probabilidades recalculadas
display(unified_result_final)
```

Fuente: La autora (2024)

A continuación, en la tabla 8 se presentan los cruces de pelajes unificados y sus respectivas probabilidades.

Tabla 8. *Datos unificados.*

PADRE	ALAZÁN	ALAZÁN	ALAZÁN	ALAZÁN	CASTAÑO	CASTAÑO
MADRE	**ALAZÁN**	**CASTAÑO**	**PALOMINO**	**ZAINO**	**CASTAÑO**	**ZAINO**
TOTAL	2418	1239	556	591	204	112
Prob_ALAZAN	96%	49%	50%	41%	24%	14%
Prob_BAYO	0%	0%	1%	0%	0%	2%
Prob_CASTAÑO	2%	40%	1%	32%	60%	45%
Prob_NEGRO	0%	2%	0%	2%	0%	5%
Prob_OTROS	0%	0%	0%	0%	0%	1%
Prob_PALOMINO	1%	0%	47%	1%	0%	1%
Prob_TORDILLO	0%	0%	0%	0%	0%	0%
Prob_ZAINO	1%	8%	0%	23%	14%	32%

Fuente: La autora (2024)

3.5 Evaluación

En la figura 3 se detallan los cruces de pelaje más frecuentes que se encontraron en el registro, así como el porcentaje de probabilidades.

Figura 3. *Cruces de pelaje.*

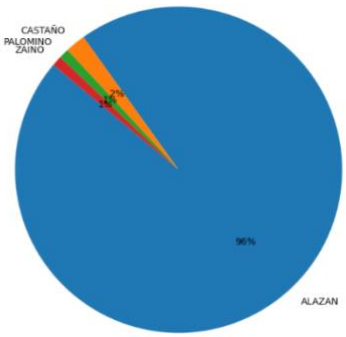

Figura 3.a. ALAZÁN X ALAZÁN

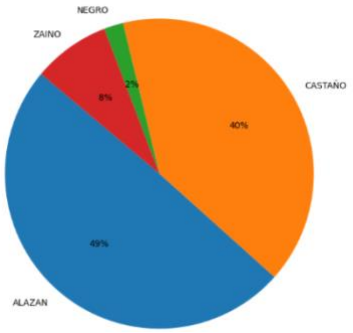

Figura 3.b. ALAZÁN X CASTAÑO.

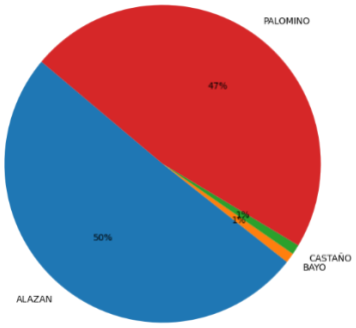

Figura 3.c. ALAZÁN X PALOMINO.

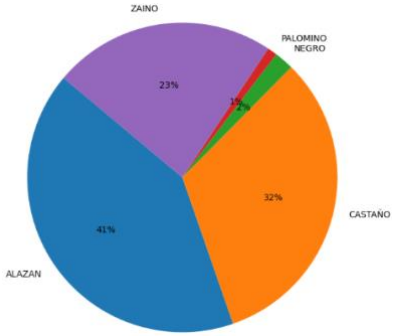

Figura 3.d. ALAZÁN X ZAINO.

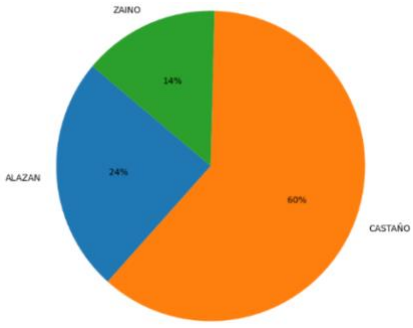

Figura 3.e. CASTAÑO X CASTAÑO.

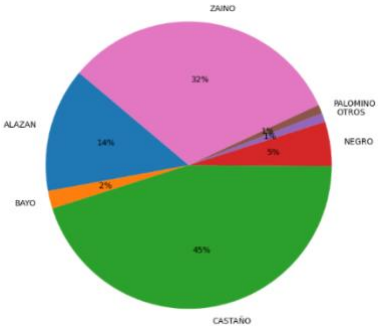

Figura 3.f. CASTAÑO X ZAINO.

Fuente: La autora (2024)

Tal como se observa en la figura 3.a, la combinación de progenitores ALAZÁN muestra una clara dominancia del fenotipo ALAZÁN, con una probabilidad del 96% de que los descendientes presenten este pelaje. Según bibliografía consultada (Losinno, 2009) el cruce referido debe dar solo ALAZÁN por lo que se infiere que podría tratarse de errores de tipeo en la data original, la misma que debería ser revisada de acuerdo con este informe.

ALAZÁN x CASTAÑO: En la figura 3.b, la probabilidad de obtener descendientes con pelaje ALAZÁN es del 49%, mientras que la probabilidad de obtener descendientes con pelaje CASTAÑO es del 40%. Esta distribución relativamente equilibrada indica una influencia significativa de ambos fenotipos parentales. Los demás fenotipos son prácticamente inexistentes, resaltando la dominancia de ALAZÁN y CASTAÑO en esta combinación.

ALAZÁN x PALOMINO: En la figura 3.c, las probabilidades de este cruce están predominantemente divididas entre los fenotipos ALAZÁN (50%) y PALOMINO (47%). Esta casi paridad sugiere una fuerte herencia codominante entre estos dos fenotipos, haciendo que otros pelajes sean raros entre los descendientes.

ALAZÁN x ZAINO: La combinación mostrada en la figura 3.d presenta una mayor diversidad en las probabilidades de fenotipos. Los descendientes tienen un 41% de probabilidad de ser ALAZÁN, un 32% de ser CASTAÑO y un 23% de ser ZAINO. La notable distribución entre estos tres fenotipos indica una influencia genética significativa y más equilibrada de los progenitores, promoviendo una mayor variabilidad fenotípica.

CASTAÑO x CASTAÑO: En la figura 3.e, los descendientes de esta combinación tienen una probabilidad del 60% de presentar pelaje CASTAÑO y del 24% de presentar pelaje ALAZÁN. La elevada probabilidad de fenotipo CASTAÑO indica una fuerte dominancia genética de este pelaje en los cruces homogéneos de CASTAÑO. Los otros fenotipos son prácticamente inexistentes en esta combinación.

CASTAÑO x ZAINO: Las probabilidades de esta combinación mostradas en la figura 3.f, están más distribuidas, con un 45% de descendientes CASTAÑO y un 32% de descendientes ZAINO. Esta combinación sugiere una herencia significativa de ambos fenotipos, con una diversidad fenotípica más amplia en comparación con otras combinaciones.

3.6 Despliegue

De acuerdo con las reglas propuestas por Bartolomé et al. (2008), se asociaron los genotipos con los fenotipos encontrados con mayor frecuencia en los cruces de pelaje de los datos analizados. Los fenotipos principales identificados fueron ALAZÁN, CASTAÑO, PALOMINO y ZAINO, los cuales están determinados por los genes: Extension, Agouti y Crema. Más detalles sobre la herencia de estos genes se encuentran en la tabla 9.

Tabla 9*. Genes Implicados en la determinación del color de la capa.*

GENOTIPOS	FENOTIPOS
Extension(E)	**EE:** Permite la producción de pelos negros (castaño o negro). No puede tener descendencia alazana.
	Ee: Permite la producción de pelos negros (castaño o negro). Puede tener descendencia alazana y no alazana.
	ee: Alazán. Puede tener descendencia alazana y no alazana.
Agouti(A)	**AA: Castaño.** Nunca puede tener descendencia negra.
	Aa: Castaño. Puede tener descendencia negra y castaña.
	aa: Negro. Puede tener descendencia negra y castaña.

Crema (Cr)	**Accion del gen MATP sobre la capa castaña: Capas bayas y perlina:** **cc: Castaño.** Sin efecto. **CrC : Bayo.** Puede tener descendencia de capa diluida. **CrCr: Perlino.** Toda la descendencia será de capa diluida. **Accion del gen MATP sobre la capa alazana: Capas palomino y cremello:** **cc:** Capa alazana. Sin efecto. **CCr:** Palomino o Isabela. Puede tener descendencia de capa diluida. **CrCr:** Cremello. Toda la descendencia será de capa diluida.

Fuente: La autora (2024)

También se utilizaron simuladores de capa genética (Horse Breeding, s.f.; Capas Genéticas - LG PRE ANCCE, s.f.). para corroborar estos resultados. La Tabla 10 presenta los fenotipos asociados al probable genotipo.

Tabla 10. *Genotipos y fenotipos de las capas equinas.*

FENOTIPOS	**GENOTIPOS**		
	Extension	Agouti	Crema
	E; e	A; At; a	Cr; C
ALAZÁN	ee	-	CC
PALOMINO	ee	-	CrC
CASTAÑO	E_	A_	CC
ZAINO	E_	At_	CC

Fuente: La autora (2024)

Este esquema proporciona una visión general de cómo ciertos genes se asocian con los colores de pelajes específicos de los caballos. Es importante tener en cuenta que la genética real puede ser más compleja debido a la interacción de múltiples genes y alelos.

Para obtener un caballo con un pelaje específico, es crucial comprender la combinación genética necesaria. Por ejemplo, para un caballo de pelaje CASTAÑO, se requiere la presencia de al menos un alelo dominante para el gen de Extension (E_) y al menos un alelo dominante para el gen de Agouti (A_). En el caso del pelaje ZAINO, se necesita una combinación de alelos Agouti en AtAt o Ata. El pelaje CASTAÑO (AAt; Aa) domina sobre el ZAINO (Ata). En contraste, para obtener un caballo de pelaje ALAZÁN, ambos alelos del gen de Extension deben ser recesivos (ee). Además, cuando un caballo ALAZÁN presenta el alelo CrC en estado heterocigoto, se produce una dilución parcial del color rojo, dando lugar al pelaje PALOMINO.

En general, el análisis reveló que el ALAZÁN fue la capa más predominante, seguida por el CASTAÑO, por lo cual se puede inferir que hubo una influencia significativa de cruces con los genes de Extension y Agouti, siendo los más comunes en la determinación de estos colores de capa en la población de caballos estudiada probablemente debido a una preferencia por esos colores de pelaje. El código procesado está disponible en el repositorio de GitHub, en el siguiente enlace*: https://github.com/julexii/deploy/blob/main/Analisis_exploratorio.ipynb*

El estudio titulado “Medidas hipométricas e índices zoométricos del Caballo Peruano de Paso criado en Cutervo, Cajamarca” desarrollado por Monteza (2021) presenta como resultado, que el color predominante en los caballos machos fue el CASTAÑO, representando el 41.67% de la muestra, seguido por el ZAINO con un 29.17% y el ALAZÁN con un 20.83%. En cuanto a las yeguas, se observó una mayor variedad de colores. El ALAZÁN fue el más común con un 35.29%, seguido por el CASTAÑO con un 29.41%, luego el ZAINO con un 14.71%, el NEGRO con un 11.76% y el TORDILLO con un 5.88%. Estos resultados muestran una preferencia notable por los pelajes ALAZÁN y CASTAÑO, los cuales también predominan en el presente estudio tanto en caballos machos como en yeguas.

Existe información sobre la población y distribución, sanidad, reproducción/mejoramiento genético, alimentación, manejo e instalaciones, de esta raza caballar, donde se muestra que el color de capa alazán es predominante con un (49.32%), seguido de un (13.12%) del color castaño, también el (9.95%) ocupa un color palomino y así entre los menores porcentajes tenemos al bayo y otros colores donde también predomina el alazán sobre el castaño que es el que tuvo mayor presencia en este estudio (Marquez Davila, 2015).

Como otro aporte, respecto a los genotipos la investigación titulada “Detección de la diversidad genética del caballo doméstico (Equus caballus) mediante genes asociados al color del pelaje” (Correa et al., 2015) se estudió la estructura genética de siete poblaciones de caballos criollos, siendo el marcador Extensión el de mayor frecuencia mientras los genes Overo y Tobiano presentaron los menores valores.

Yepes et al. (2017) en su investigación titulada “Diversidad Genética del Caballo Criollo (Equus caballus) mediante Genes Asociados al Pelaje en Valencia, Colombia” se evaluó la variabilidad genética en cinco poblaciones de caballos criollos utilizando marcadores de pelaje. Mostrando como resultados ausencia de los marcadores White y Overo, mientras que

los marcadores Extension y Agouti fueron los de mayores frecuencias, posiblemente favorecidas por selección artificial.

Moreno et al. (2018) en su estudio “Variabilidad genética del caballo (Equus caballus) mediante genes del pelaje en Sahagún, Córdoba, Colombia” se realizaron muestreos en siete poblaciones rurales entre noviembre de 2016 y noviembre de 2017. Se caracterizó fenotípicamente cada animal utilizando marcadores autosómicos de codificación morfológica: Extension, Agouti, Cream, Gray, White, Tobiano, Overo y Roan. Los resultados mostraron la ausencia de los marcadores White y Overo, mientras que el marcador Extension fue el más frecuente, probablemente debido a la selección artificial. El marcador Tobiano fue el menos frecuente, lo cual podría estar relacionado con la selección en contra de individuos que portan este marcador.

Los resultados del presente estudio concuerdan con las investigaciones mencionadas anteriormente, donde se encontró que el gen más frecuente es el Extension. Estos hallazgos se basan en la observación de que las variantes alélicas responsables de las coloraciones oscuras son favorecidas en comparación con aquellas que expresan tonalidades claras. Además, se ha demostrado que los animales con capas oscuras tienen una mayor capacidad de absorción de radiación calórica que aquellos con tonalidades claras. Esto sugiere que los caballos oscuros están mejor adaptados a las condiciones del clima tropical, lo que podría explicar la prevalencia de los genes Extension y Agouti en las poblaciones estudiadas.

4 CONCLUSIONES

Se encontró un significativo desbalanceo en los datos, lo que sugiere la necesidad de aplicar técnicas de balanceo de datos para su uso en inteligencia computacional. Este desbalanceo refleja posibles sesgos en la población de caballos estudiada, probablemente influenciados por preferencias de cría hacia ciertos colores de pelajes y prácticas históricas y culturales en la selección de reproductores

Por otro lado, el análisis exploratorio ha permitido identificar patrones y tendencias significativas en la población estudiada. Los resultados muestran que los pelajes ALAZÁN y CASTAÑO son predominantes, lo que indica una fuerte influencia de los genes Extension y Agouti. Esta prevalencia sugiere no solo preferencias genéticas, sino también posibles prácticas históricas y culturales en la cría de caballos.

Este análisis no sólo enriquece nuestra comprensión de la raza, sino que también proporciona información valiosa para criadores y gestores, permitiéndoles tomar decisiones informadas para producir descendencia con características deseables. Además, llevar un registro genético ayuda a prevenir la reproducción entre individuos portadores de enfermedades hereditarias, promueve la autenticidad y el valor de los animales en el mercado, y contribuye a preservar la diversidad genética dentro de la población.

Por otro lado, la genética es una ciencia que posibilita el aumentar las probabilidades de obtener un potro del color deseado, se puede realizar un examen genético tanto de la yegua como del semental. Esto permitirá identificar los genes y variantes presentes, eliminando el margen de error y proporcionando una comprensión precisa de las posibilidades reales.

BIBLIOGRAFÍA

Alfaro, A. J., & Ospina, J. D. (2021). Revisión sistemática de literatura: Técnicas de aprendizaje automático (machine learning). *Cuaderno Activa: REVISTA CIENTÍFICA DE LA FACULTAD DE INGENIERÍA*(13), 114-130. Obtenido de https://ojs.tdea.edu.co/index.php/cuadernoactiva/article/view/849/1366

Bolaños, J. M. (2020). Determinación de capas genéticas. *ExtremaduraPRE: la revista de la Asociación Extremeña de Criadores de Caballos de Pura Raza Española*(37), 54-59. Obtenido de https://dialnet.unirioja.es/servlet/articulo?codigo=7768082

Bartolomé, E., Azor, P., Gómez, M., & F, P. (2008). *La determinación genética del color de la capa en el caballo: Bases y aplicación al caballo de la raza Pottoka*. Obtenido de Pottoka.info: https://www.pottoka.info/files/galeria/Genetica_color_capa_pottoka.pdf

Capas Genéticas - LG PRE ANCCE. (s.f.). Obtenido de Lgancce.com: https://www.lgancce.com/webcapas/inicio

Correa, L., Reyes, C., Pardo, E., & Cavadía, T. (2015). Detección de la diversidad genética del caballo doméstico(Equus caballus) mediante genes asociados al color del pelaje. *Revista MVZ Córdoba, 20*(3), 76-103.

Davila, S. A. (2015). *Diagnóstico de la crianza del Caballo Peruano de Paso en el Valle.* Tesis Ingeniero Zooecnista, HUANCAYO-PERÚ. Obtenido de https://repositorio.uncp.edu.pe/bitstream/handle/20.500.12894/1846/Tesis%20M%C3%A1rquez.pdf?sequence=1

Espinosa, Z. J. (2020). Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública. *Ingeniería. Investigación y Tecnología, 11*(1), 1-17. doi: https://doi.org/10.22201/fi.25940732e.2020.21n1.008

Flores, L. G., Cadena, M. J., Quinatoa, A. E., & Villa, Q. M. (2019). Minería de datos como herramienta estratégica. *Revista Científica Mundo de la Investigación y el Conocimiento, 3*(1), 955- 970. doi:10.26820/recimundo/3.(1).enero.2019.955-970

Fuentes, N. R. (2022). *Una aproximación al análisis exploratorio de datos.* Valladolid: Universidad de Valladolid. Obtenido de Uva.es: https://uvadoc.uva.es/bitstream/handle/10324/53276/TFG-E-1381.pdf?sequence=1&isAllowed=y

García, P. V., Mora, M. A., & Ávila, R. J. (2020). La inteligencia artificial en la educación. *Revista Cientifica: Dominio de las ciencias, 6*(3), 648-666. doi:https://dx.doi.org/10.23857/dc.v6i3.1421

Gómez, C. A. (2020). Aplicación del machine learning en agricultura de precisión. *Revista Cintex*, 14-27. Obtenido de https://revistas.pascualbravo.edu.co/index.php/cintex/article/view/356/327

Gómez William, O. A. (2023). La Inteligencia Artificial y su Incidencia en la Educación: Transformando el Aprendizaje para el Siglo XXI. *REVISTA INTERNACIONAL DE PEDAGOGÍA EINNOVACIÓN EDUCATIVA, 3*(2), 217- 229. doi:https://orcid.org/0000-0002-8178-1253

Horse Breeding. (s.f.). Obtenido de Etalondx.com: https://www.etalondx.com/horse-breeding/

Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model. *Procedia CIRP, 79*, 403-408. doi:https://doi.org/10.1016/j.procir.2019.02.106

López, M. G. (2021). *Pre-procesamiento de datos para aprendizaje de distribución de etiquetas.* Granada: Universidad de Granada. Obtenido de https://digibug.ugr.es/handle/10481/68012

Losinno, L. (2009). *Curso de Producción Equina I.* Departamento de Producción Animal, Facultad de Agronomía y Veterinaria. Universidad Nacional de Río Cuarto. Obtenido de https://www.produccion-animal.com.ar/produccion_equinos/curso_equinos_I/18-Guia_PELAJES_2009.pdf

Medina, Q. F., Castillo, R. W., & Meneses, V. C. (2020). Métricas para el apoyo de la exploración visual de componentes en modelos de minería de datos. *Ingeniare. Revista chilena de ingeniería, 28*(4), 596-611. doi:http://dx.doi.org/10.4067/S0718-33052020000400596

Mohedano, M. Á. (2021). *Análisis de datos mediante visualización de información basada en técnicas de reducción de dimensiones y machine learning.* España: Universidad Rey Juan Marcos. Obtenido de https://www.educacion.gob.es/teseo/imprimirFicheroTesis.do?idFichero=eHv0JdWVblM%3D

Montalván, D. P., & Rojas, R. L. (2019). *“Parámetros hematológicos en equinos (equus caballus) pertenecientes a la asociación de criadores y propietarios del caballo peruano de paso de lambayeque”.* LAMBAYEQUE: UNIVERSIDAD NACIONAL "PEDRO RUIZ GALLO”. Obtenido de Montalván_Damián_Paola_Antonella_y_Rojas_Risco_Liliana_Libertad%20(1)%20(3).pdf

Monteza, C. W. (2021). *Medidas hipométricas e índices zoométricos del Caballo Peruano de Paso criado en Cutervo, Cajamarca.* Lambayeque.

Moreno, C., Causil, L., & Pardo, E. (2018). Variabilidad genética del caballo (Equus caballus) mediante genes del pelaje en Sahagún, Córdoba, Colombia. *Revista De Investigaciones Veterinarias Del Perú, 29*(4), 1295-1302. doi:https://doi.org/10.15381/rivep.v29i4.15188

Ramírez, J. D. (2021). Aprendizaje Automático y Aprendizaje Profundo. *Ingeniare. Revista chilena de ingeniería, 29*(2), 180-181. doi:http://dx.doi.org/10.4067/S0718-33052021000200180

Rojas, E. M. (2020). Machine Learning: análisis de lenguajes de programación y herramientas para desarrollo. *Risti: Revista Ibérica de Sistemas e Tecnologias de Informação*(28), 586 - 599. Obtenido de https://www.proquest.com/openview/c7e24c997199215aa26a39107dd2fe98/1?pq-origsite=gscholar&cbl=1006393

Sánchez, M. M. (2021). *Estudio del funcionamiento de técnicas de minería de datos sobre Conjuntos de Datos relacionados con la Biología.* JAÉN: UNIVERSIDAD DE JAÉN. Obtenido de https://crea.ujaen.es/bitstream/10953.1/14426/1/TFGB%20Moret%20Sanchez%20Martin.pdf

Tortosa, R. E. (2020). *Conjunto de datos para inteligencia artificial.* Primavera: Laboratorio Universidad de Ciencias Aplicadas LTD.

Troya, A. H. (2019). Técnicas estadísticas en el análisis cuantitativo de datos. *Revista Sigma, 15*(1), 28-44. Obtenido de http://coes.udenar.edu.co/revistasigma/articulosXV/1.pdf

Valencia, D. J., Mera, C., & Sepúlveda, L. M. (2020). Visualización de conjuntos de datos de múltiples instancias. *Revista Ibérica de Sistemas y Tecnologías de Información*, 84-99. doi:10.17013/risti.39.84-99

Yepes, W., Pérez, E. P., & Vargas, L. A. (2017). Diversidad Genética del Caballo Criollo (Equuscaballus)mediante Genes Asociados al Pelaje en Valencia, Colombia. *Revista De Investigaciones Veterinarias Del Perú, 28*(3), 562-570. doi:https://doi.org/10.15381/rivep.v28i3.13353