T.P ESPECIAL

FUNDAMENTOS DE LA CIENCIA DE DATOS

Grupo 15

Iturralde santiago, iturraldeariel61@gmail.com
Laborde Joaquin, labjoaquin005@gmail.com

1. INTRODUCCIÓN:

Contexto del problema:

El cólico equino es una de las principales causas de emergencias veterinarias en caballos y puede implicar tanto problemas gastrointestinales simples como lesiones graves que requieren cirugía. La detección temprana y el diagnóstico correcto son fundamentales para aumentar las probabilidades de supervivencia. En este contexto, la ciencia de datos puede resultar una herramienta valiosa para analizar grandes volúmenes de casos clínicos, identificar patrones y asistir en la toma de decisiones veterinarias.

El conjunto de datos que nos tocó (Horse Colic) disponible en el repositorio UCI Machine Learning, contiene 368 instancias (300 de entrenamiento y 68 de prueba) y un total de 28 atributos, que incluyen tanto variables clínicas (temperatura, pulso, dolor, estado abdominal) como variables de resultado (si el caballo sobrevivió, si fue necesaria cirugía, tipo de lesión, entre otros). Este dataset combina variables discretas, continuas y nominales lo cual lo hace versátil para realizar distintos tipos de análisis, también presenta particularidades como por ejemplo el 30% de los valores están ausentes y múltiples variables de resultado posibles, lo que permite enfocar el análisis en distintos objetivos.

En cuanto a su utilidad, podemos ver que sirve para poder estudiar factores asociados a la supervivencia y la necesidad de alguna cirugía, ayudando a la construcción de modelos que puedan asistir a veterinarios en la práctica clínica.

Organización del trabajo:

El análisis de este dataset se organiza en las siguientes secciones:

1. Análisis exploratorio inicial: se describe la estructura de la base, se investiga el origen de los datos, se caracterizan las variables y se estudian distribuciones, outliers y valores faltantes.
2. Limpieza de datos: se implementan acciones de preprocesamiento necesarias para trabajar con la base, detallando y justificando cada corrección aplicada.
3. Planteo de hipótesis: se formulan seis hipótesis de interés, justificadas en el contexto y relacionadas con los atributos del dataset. Se incluyen hipótesis univariadas, bivariadas y multivariadas.
4. Validación de hipótesis: se aplican los métodos estadísticos y de prueba vistos en clase para evaluar cada hipótesis, discutiendo los resultados obtenidos.
5. Conclusión: se presentan los principales hallazgos del análisis y su relevancia en el contexto del problema.

2.1. ANÁLISIS EXPLORATORIO DE LOS DATOS:

Estructura y preparación del dataset:

- Para facilitar el análisis, se descargaron y descomprimieron los archivos originales .data y .test desde la fuente oficial. Se asignaron los nombres de columnas según la documentación, y se reemplazaron los valores desconocidos por NaN de manera que puedan ser tratados adecuadamente. Posteriormente, ambos conjuntos se unieron en un solo Data Frame para simplificar el análisis exploratorio.

Inspección inicial de los datos:

- Se realizó una revisión rápida de las primeras filas del dataset para verificar que las columnas correspondieran a los atributos esperados y se confirmaron los tipos de datos. Además, se observaron valores faltantes en varias columnas cuando ejecutamos raw_dataset.info(), lo que es consistente con la documentación que indica que aproximadamente el 30% de los valores están ausentes.

Observaciones iniciales:

- La presencia de valores faltantes es significativa en varias columnas, lo que requerirá imputación o estrategias de manejo de datos ausentes antes de realizar modelos predictivos.
- Las variables categóricas (cirugía, resultado, lesión_quirúrgica, dolor, etc.) permitirán definir objetivos de predicción claros y explorar asociaciones clínicas.
- La exploración inicial muestra que hay mezcla de variables numéricas, discretas y nominales, por lo que los gráficos y análisis de correlación deberán adaptarse a cada tipo de dato

ANALISIS DE VALORES FALTANTES:

Para este análisis, calculamos la cantidad de valores faltantes(Nan) de cada variables y observamos que las variables con mayor cantidad de valores nulos son ph_reflujo_nasogástrico(299 nulos),proteína_total_abdominocentesis(235 nulos) y apariencia_abdominocentesis(194 nulos). Esto sugiere que las mediciones relacionadas con los exámenes del abdomen o procedimientos invasivos no se realizaron en todos los casos, posiblemente por limitaciones clínicas o porque no todos los caballos requirieron esos estudios.

Este análisis es clave para decidir si imputar o eliminar estas variables en las siguientes etapas del preprocesamiento.

Fig. 1

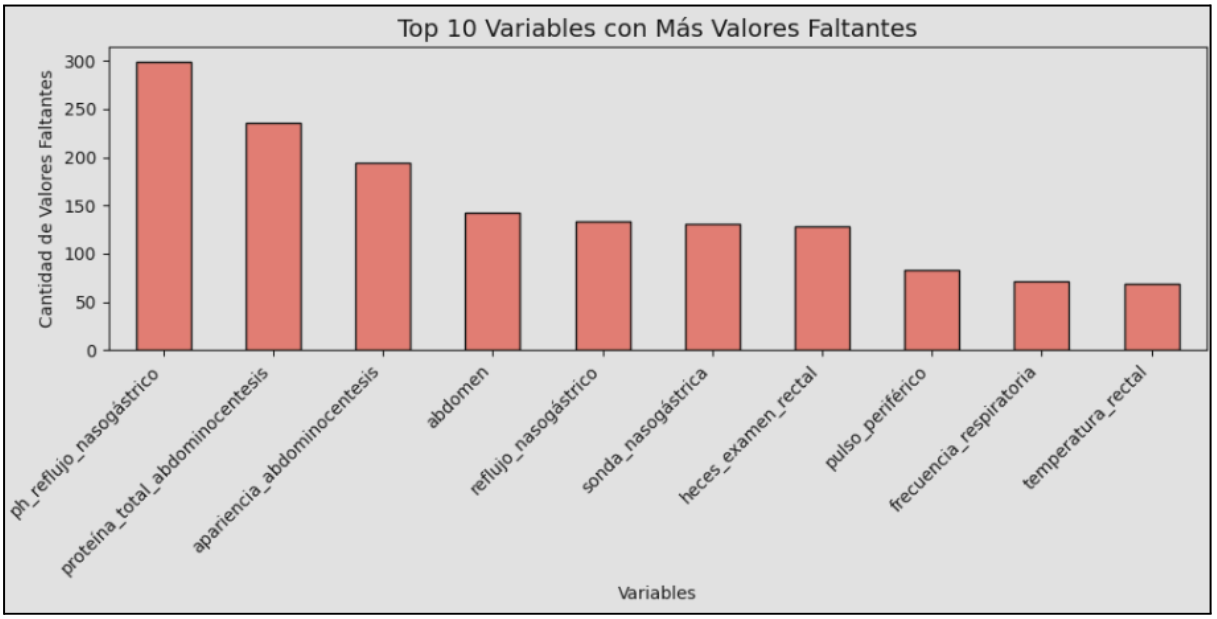


Fig. 1: Gráfico de barras que muestra las 10 variables con mayor cantidad de valores faltantes en el conjunto de datos.

ANÁLISIS DE LA DISTRIBUCIÓN DE VARIABLES:

En cuanto al análisis de cómo se distribuye cada variable, lo que hicimos en primera instancia fue observar a qué tipo de clasificación (Numéricas, Categóricas) pertenece cada variable, como resultado obtuvimos 21 categóricas y 7 numéricas.

CATEGÓRICA NOMINAL	CATEGÓRICA ORDINAL	NUMÉRICAS CONTINUAS	NUMÉRICAS DISCRETAS
cirugia, Edad, Numero_hospital membranas_mucosas abdomen, resultado, lesion_quirurgica, dolor, reflujo_nasogastrico heces_examen_rectal sitio_lesion, tipo_lesion, subtipo_lesion, cp_data,	temperatura_extremidades, pulso_periférico, tiempo de llenado_capilar, peristalsis, distencion_abdominal, sonda_nasogastrica, apariencia de la abdominocentesis	temperatura rectal, pulso, ph reflujo nasogástrico, vol cel empaquetado, proteína total, proteína total abdominocentesis	frecuencia_respiratoria

Para cada tipo de clasificación de las variables existen diferentes gráficos para visualizar su distribución

1. Numéricas:
    - Histogramas para poder ver la forma de su distribución
    - Boxplots para mostrar la mediana, los cuartiles y los outliers
2. Categóricas:
    - Gráficos de barras para poder ver la frecuencia de cada categoría

Para no analizar las 28 variables existentes en el dataset, estudiamos las que tienen una mayor importancia, en este caso el dataset nos indica que se intentan predecir si la lesión de los caballos es quirúrgica o no, el resultado, es decir si el caballo sobrevivió, falleció o fue eutanasiado, los tipos de lesión y lugar de lesión. Esto tiene sentido ya que por ejemplo para la variable resultado (probablemente la más importante) casi todas las hipótesis deberían intentar predecirla.

Otras variables que son importantes son:

- El dolor ya que el dataset nos dice que “en general cuanto más doloroso, más probable es que requiera cirugía”.
- La distensión abdominal ya que un animal con distensión abdominal severa es probable que presente dolor y requiera cirugía.
- El pulso, un pulso elevado puede indicar shock o dolor.
- La temperatura rectal, puede indicar infección (si es alta) o shock tardío (si es baja).
- La temperatura de las extremidades, extremidades frías indican un posible shock y extremidades calientes deben correlacionarse con una temperatura rectal alta.

En primer lugar vamos a analizar las dos variables más importantes, la variable resultado y la variable lesión quirúrgica, vamos a observar cómo se distribuyen, cuantos caballos vivieron y cuántos fallecieron y cuantos necesitan cirugía.

Fig. 2

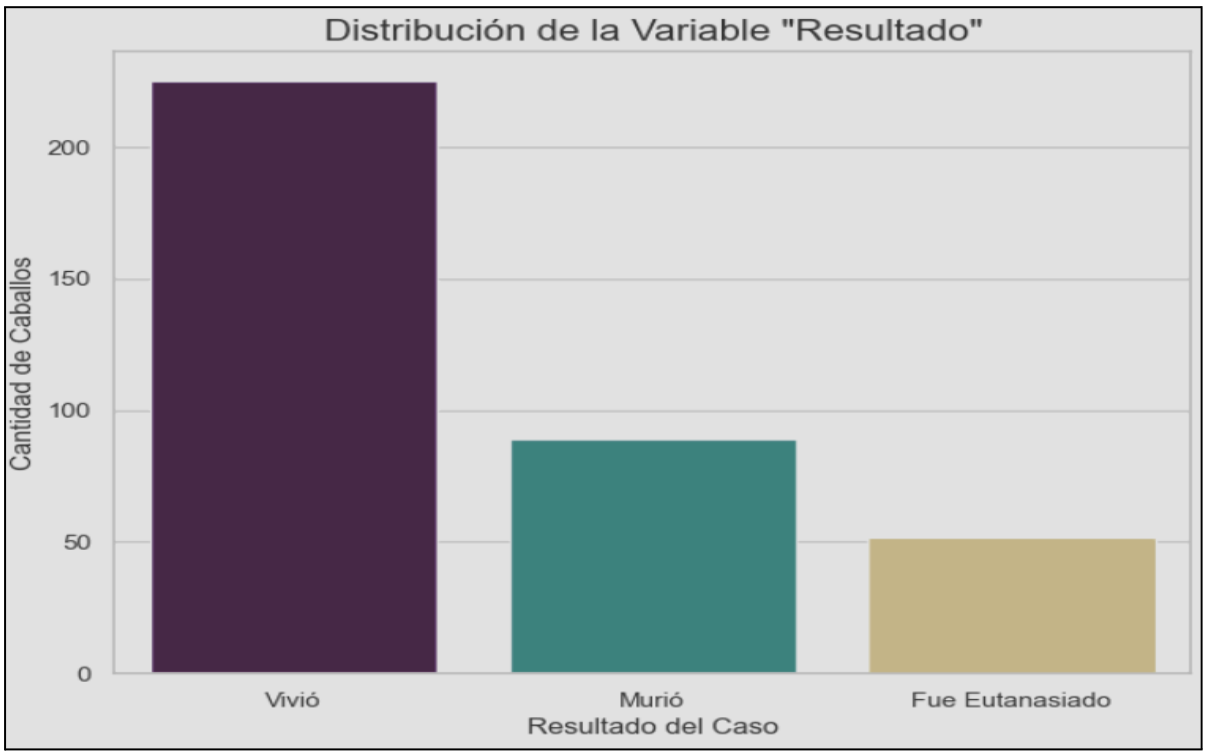


Fig. 2: Gráfico de barras de la variable resultado, se observa la cantidad de animales que sobrevivieron, murieron y fueron eutanasiados.

En este gráfico de la variable resultado observamos principalmente que la categoría “Vivió” es la más frecuente, con más de 200 casos, lo cual es una buena señal ya que la supervivencia es el resultado más común, a pesar de esto vemos que la tasa de muertes es bastante alta (cerca de 90) y los que fueron eutanasiados cerca de 50, por lo que confirmamos que el cólico equino es una condición grave, aunque la mayoría sobrevive, una gran cantidad no lo hace.

Fig. 3

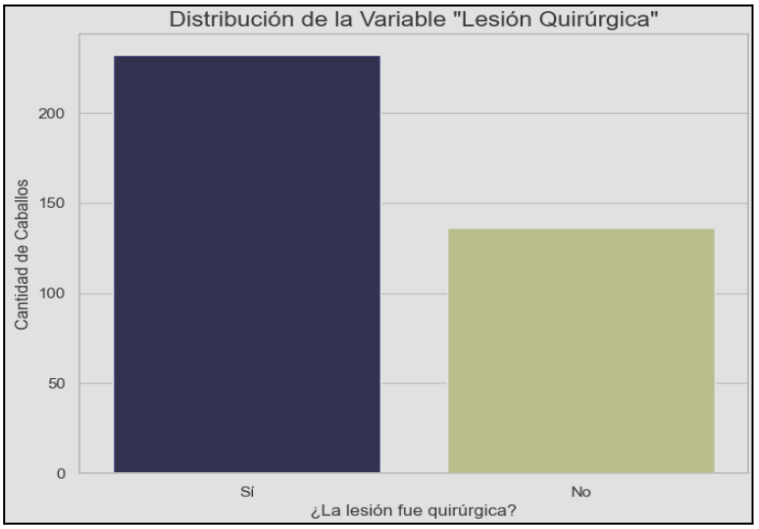


Fig. 3: Gráfico de barras de la variable lesión quirúrgica, se observa la cantidad de animales que sí poseen lesión y los que no.

En este gráfico de la variable lesión quirúrgica se observa que la categoría de que si tienen una lesión que es quirúrgica es considerablemente más alta de las lesiones que no lo son, esto significa que la mayoría de los caballos (más de 200) tenían lesiones que fueron consideradas quirúrgicas. Vemos que está fuertemente sesgado hacia casos graves que probablemente necesitaron una intervención.

Fig. 4

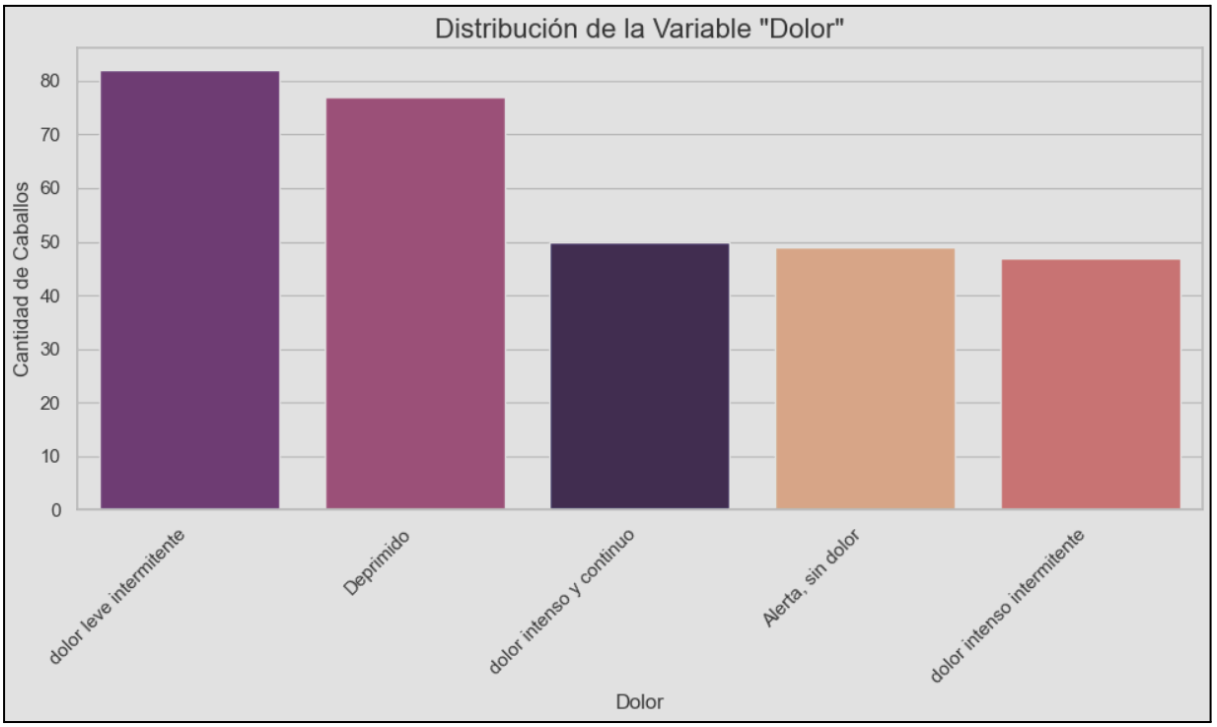


Fig. 4: Gráfico de barras de la variable “Dolor”, se observa de mayor a menor la cantidad de caballos que sufren cierto nivel de dolor a la hora de entrar a la clínica.

En el gráfico de la variable dolor, observamos que los niveles de dolor reportados están distribuidos sin un claro dominador, aunque las categorías que indican mayor severidad son muy frecuentes. El "dolor leve intermitente" y la "depresión" (que puede ser un signo de dolor intenso o toxicidad) son las dos categorías más comunes, ambas con aproximadamente 80 casos cada una. Esto refuerza la idea de que el dataset está compuesto mayormente por casos graves, ya que el dolor es uno de los principales indicadores que sugieren la necesidad de una cirugía.

Fig. 5

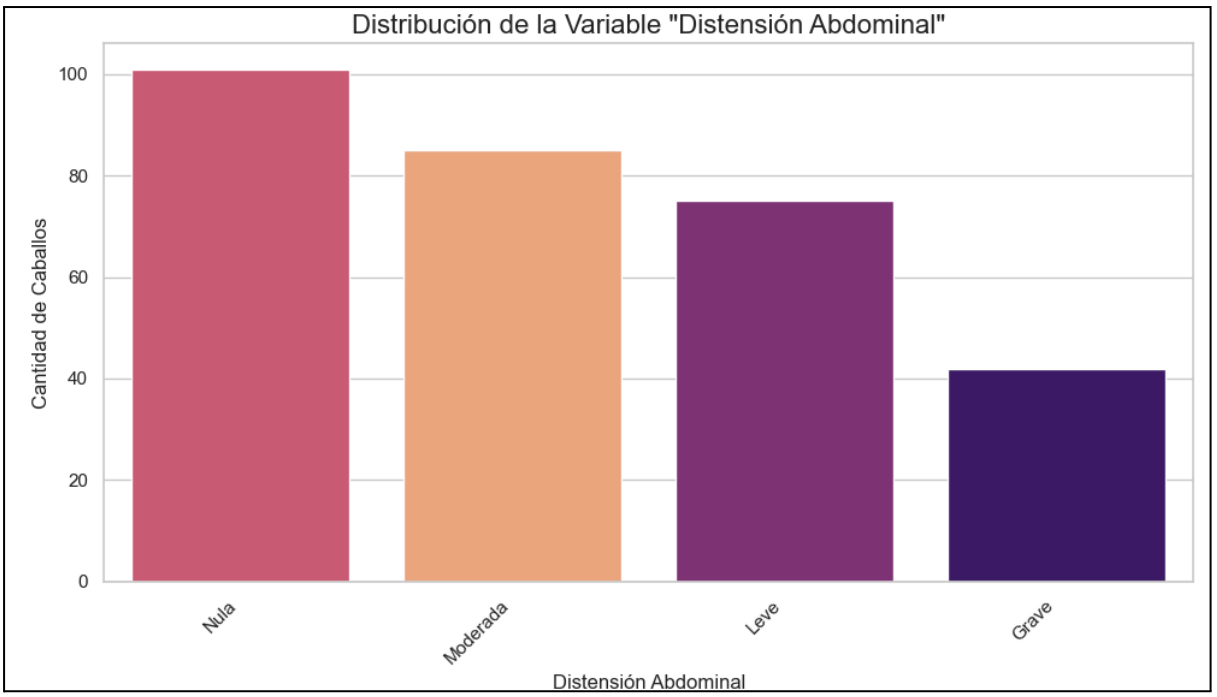


Fig. 5: Gráfico de barras de la variable “Distensión Abdominal”, se observa de mayor a menor la cantidad de caballos que sufren distintos niveles de Distensión Abdominal a la hora de entrar a la clínica.

La distensión abdominal es un parámetro clínico crucial, y el gráfico muestra que la mayoría de los caballos presentaban algún grado de distensión. Las categorías "Nula" y "Moderada" son las más frecuentes, seguidas de cerca por "Leve". Es importante destacar que una distensión "Grave", aunque es la categoría menos frecuente, aparece en más de 40 casos, lo que es un indicador claro de una condición crítica que probablemente requiera intervención quirúrgica.

Fig. 6

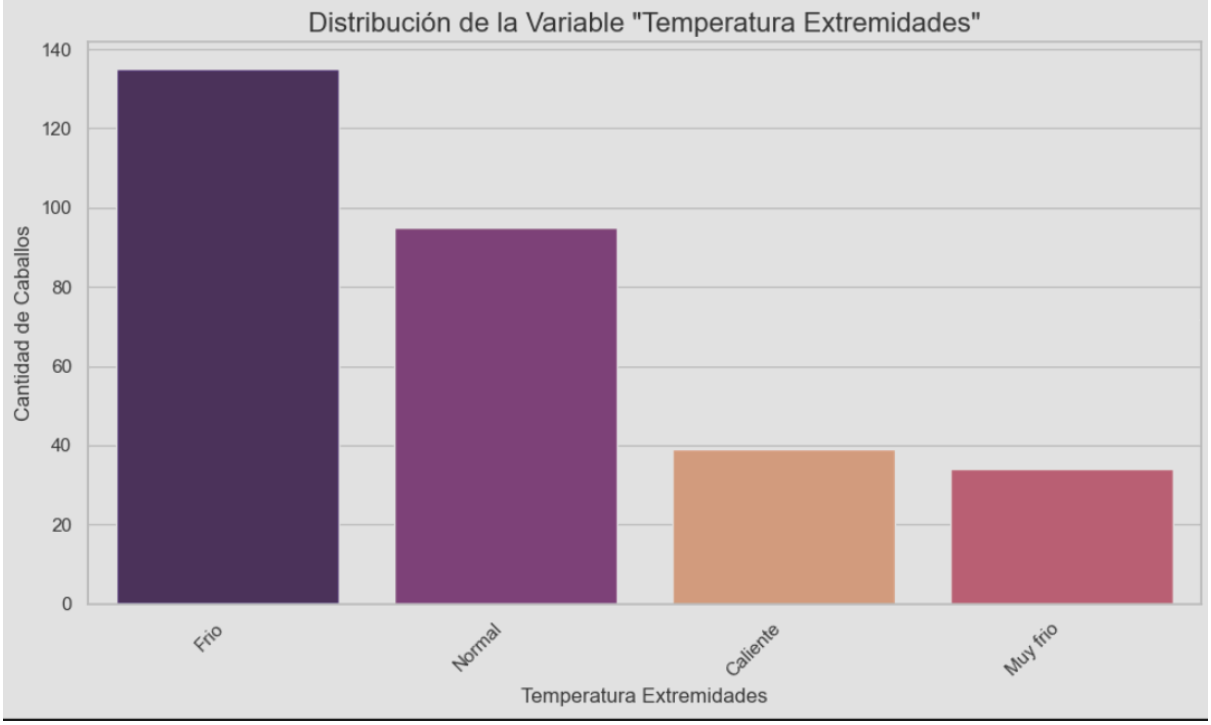


Fig. 6: Gráfico de barras de la variable “Temperatura Extremidades” muestra la distribución de la temperatura de las extremidades de los caballos

.

En este gráfico se puede notar como gran proporción de los caballos en este estudio presentaban signos de shock, esto se puede deber a que gran mayoría de los caballos que ingresaron a la clínica estaban en un estado grave.

Fig. 7:

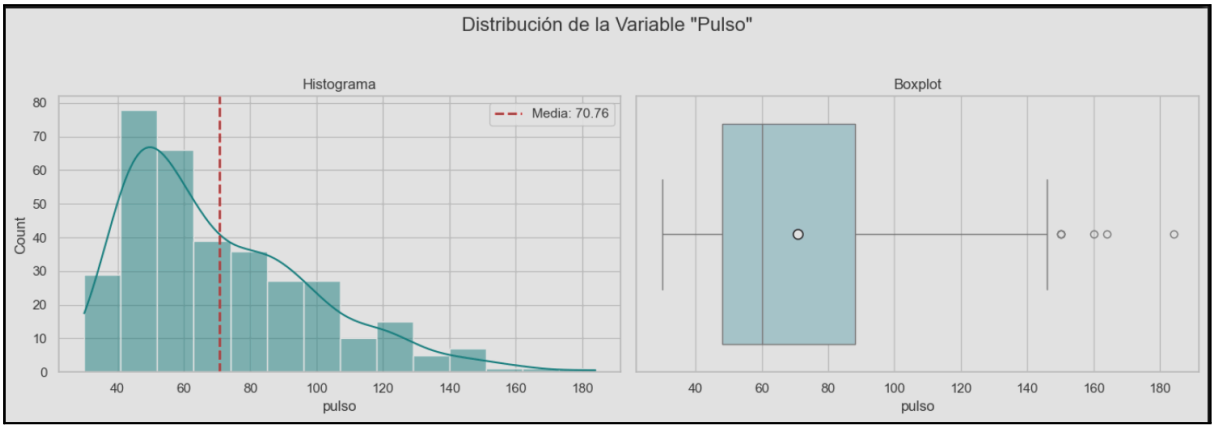


Fig. 7 (respectivamente): Esta figura presenta dos visualizaciones complementarias de la distribución del pulso.

En este gráfico, el histograma muestra una distribución con un sesgo a la derecha, lo que indica una cola de casos con pulsos muy elevados. Aunque el pico de

frecuencia se encuentra en el rango de 40-60 latidos por minuto (lpm), la media calculada es de 70.76 lpm, como indica la línea roja. Este valor promedio de 70.76 lpm está muy por encima del rango normal para un caballo adulto, que es de 30-40 lpm. El boxplot refuerza esta idea, mostrando que la mediana (la línea dentro de la caja) se sitúa en 60 lpm, mientras que la media (el círculo) es visiblemente más alta, arrastrada hacia arriba por los numerosos valores atípicos (outliers) que llegan hasta 180 lpm

Fig. 8:

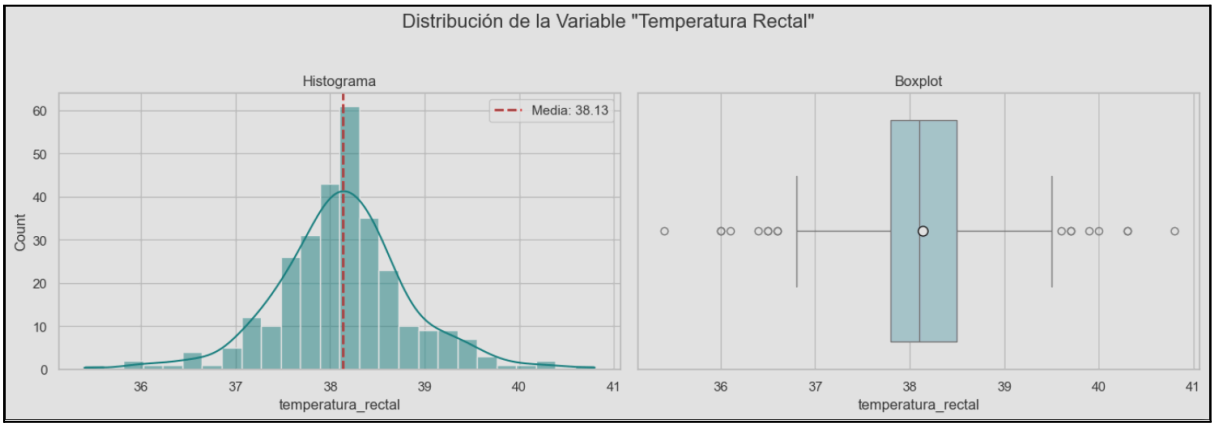


Fig. 8: Esta figura presenta dos visualizaciones complementarias de la distribución de la temperatura rectal.

La temperatura rectal presenta una distribución que se asemeja más a una campana, centrada alrededor de los 38°C, lo cual se considera una temperatura normal o ligeramente elevada. Sin embargo, el boxplot nos muestra la presencia de outliers en ambos extremos: temperaturas muy altas (fiebre, posiblemente por infección) y temperaturas bajas (que pueden ocurrir en un estado de shock tardío). Esto indica que, aunque la mayoría de los caballos mantenían una temperatura dentro del rango normal, había casos extremos que reflejaban complicaciones serias.

2.2. LIMPIEZA DE LA BASE:

Una vez completado el análisis exploratorio de datos (EDA), se identificaron dos problemas principales en el conjunto de datos que requerían corrección antes de proceder a la validación de hipótesis: la presencia de valores faltantes (NaNs) y la existencia de valores atípicos (outliers).

Tratamiento de Valores Faltantes (NaNs):

El análisis de nulos (reflejado en la Fig. 1 del informe) mostró que las variables presentaban distintos grados de ausencia de datos, lo que requirió la aplicación de múltiples estrategias.

1. Eliminación de columnas por exceso de nulos:

Las variables ph_reflujo_nasogástrico,proteína_total_abdominocentesis y apariencia_abdominocentesis presentaban una gran cantidad de nulos aproximadamente 299, 235 y 194. Observando eso decidimos eliminar las 3 columnas del dataset ya que con un porcentaje tan elevado de datos ausentes, cualquier método de imputación resulta en la invención de la gran mayoría de los datos, introduciría un severo sesgo en el análisis y nos llevaría a conclusiones poco confiables

2. imputación variables numéricas:

Variables numéricas como temperatura rectal, pulso, frecuencia respiratoria, volumen celular empaquetado y proteína total presentaban una cantidad baja de valores nulos, como acción se decidió imputar estos valores faltantes utilizando la mediana de cada columna.

La elección de reemplazar por la mediana es que es un estadístico que no se ve afectado por valores extremos(outliers) y por lo tanto para estas variables que según vimos en la fig.7 y fig.9 poseen outliers significativos podemos proveer un reemplazo mucho más representativo y confiable del valor central de los datos

3. Imputación de Variables Categóricas

Problema: Múltiples variables categóricas (tanto nominales como ordinales, ej. dolor, temperatura extremidades, peristalsis, abdomen, etc.) también tenían valores nulos.

Acción: Se imputaron estos valores utilizando la moda (el valor más frecuente) de su respectiva columna.

Justificación: Para datos de naturaleza categórica, la moda representa el valor más probable o común. Este método permite preservar la distribución de frecuencias original de la variable sin introducir categorías artificiales o sesgos indebidos.

4. Eliminación de Filas:

Problema: Tras las imputaciones anteriores, se detectaron 2 valores nulos restantes en las columnas resultado (nuestra variable objetivo) y cirugía.

Acción: Se procedió a eliminar las 2 filas completas que contenían estos nulos.

Justificación: Imputar la variable objetivo (resultado) es una mala práctica metodológica, ya que estaríamos "inventando" el dato que precisamente buscamos analizar. Dado que la cantidad de filas afectadas era mínima, la eliminación es la estrategia más segura, limpia y no representa una pérdida significativa de información.

Manejo de Valores Atípicos (Outliers):

El EDA (en particular los diagramas de Fig. 7 y Fig. 8) reveló la presencia de valores estadísticamente atípicos (extremos) en variables numéricas clave, como pulso y temperatura rectal. Como acción a esto decidimos conservar todos los valores atípicos en el dataset y no aplicar ninguna técnica de eliminación o recorte sobre ellos, ya que debido al contexto del problema y consultando la información que nos brinda el dataset concluimos que estos valores, aunque estadísticamente extremos, son clínicamente relevantes:

1. Un pulso extremadamente alto es un indicador conocido de que el animal está en shock severo.
2. Una temperatura rectal elevada es un síntoma claro de una infección grave.

Eliminar estos outliers significaría descartar la información de los casos más críticos y graves. Esto sesgaría profundamente el análisis, haciendo que nuestros resultados reflejen únicamente a los caballos con padecimientos leves o moderados. Por lo tanto, se considera que estos outliers no son "ruido" , sino información valiosa y vital para el análisis.

Conclusión:

Tras la ejecución de los pasos detallados (eliminación de columnas, imputación con mediana y moda, eliminación de filas específicas y conservación de outliers), se generó un nuevo dataset (df_limpio) que no contiene ningún valor nulo. Este conjunto de datos limpio será el utilizado para toda la validación de hipótesis que retomaremos luego.

2.3. ANÁLISIS PRELIMINAR DE HIPÓTESIS UNIVARIADAS Y BIVARIADAS:

ANÁLISIS UNIVARIADO:

Formulamos las primeras hipótesis del estudio. Estas se centran en analizar variables individuales (análisis univariado) y comparar sus métricas con valores de referencia conocidos.

Hipótesis 1: el pulso promedio de los caballos enfermos es anormalmente alto

- El análisis exploratorio no solo muestra una media muestral de 70.76 lpm, sino que el histograma presenta un claro sesgo a la derecha, confirmando la abundancia de casos con pulsos elevados

Hipótesis 2:La temperatura rectal de los caballos enfermos es superior a la normal.

ANÁLISIS BIVARIADO, relación entre variables:

Se realizó un heatmap con la matriz de correlaciones de Pearson (Figura 9), incluyendo todas las variables numéricas y categóricas codificadas. Si bien la mayoría de las correlaciones clínicas no superan el 0.7, el gráfico permite identificar 'puntos calientes' y relaciones claves.

Fig. 9:

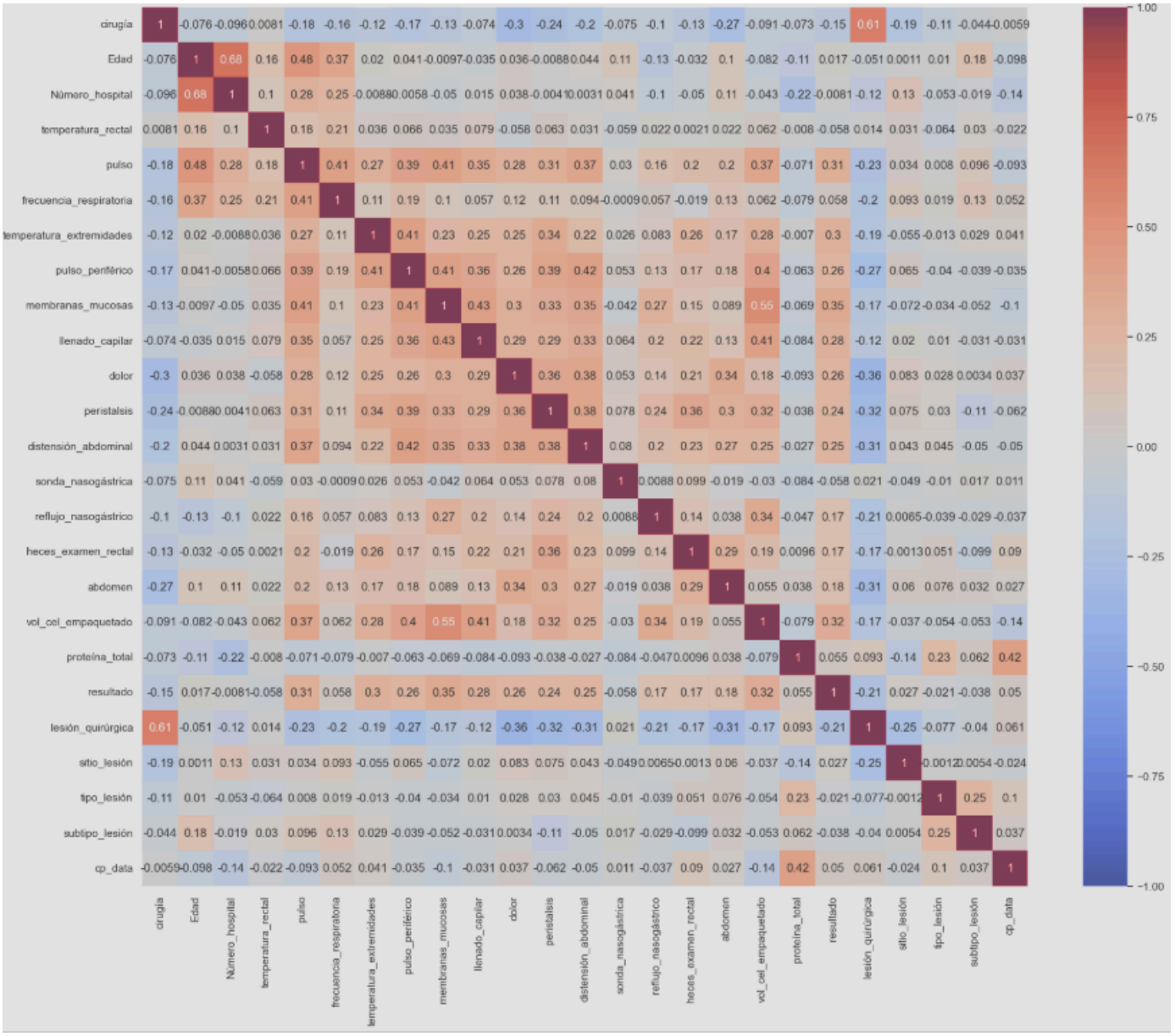


Fig. 9: mapa de calor(heatmap) que posee los coeficientes de correlación entre las variables

El mapa de calor nos sirvió para ver las correlaciones lineales. Pero para buscar patrones no lineales, donde se agrupen varias variables a la vez (ej. 'dolor' + 'temperatura' + 'membranas'), el heatmap no alcanza.

Por eso, ahora usamos T-SNE. Esta técnica nos permite "mapear" todos esos datos complejos en un gráfico 2D. La idea es ver si T-SNE puede encontrar visualmente grupos de caballos (como un grupo 'grave' y un grupo 'leve') que a simple vista no se ven.

Fig. 10:

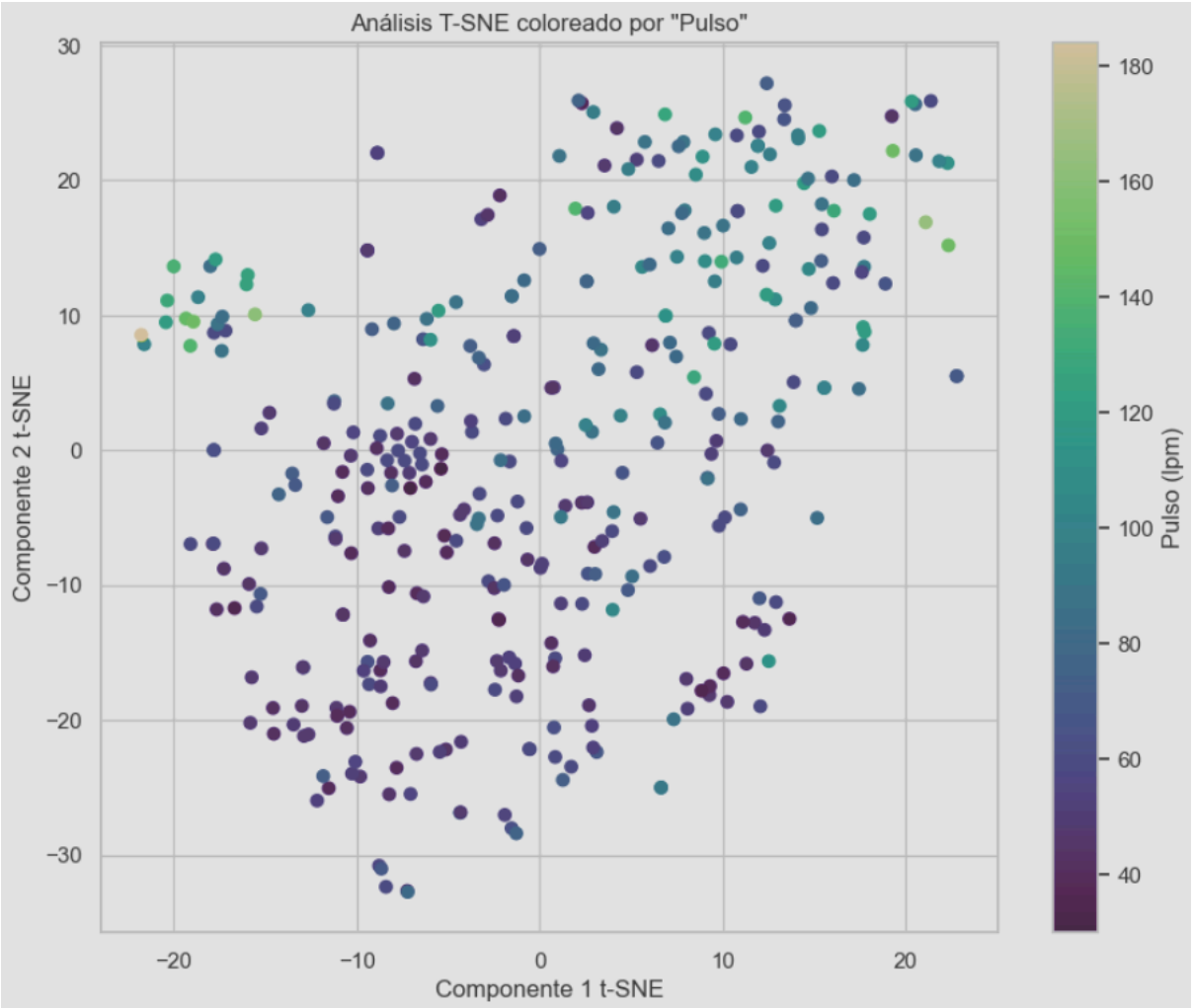


Fig. 10: gráfico t-sne coloreado por la variable pulso

Al aplicar T-SNE y colorear por la variable pulso, vemos que los caballos con pulso bajo (color oscuro) están en una zona, y los de pulso alto (color claro) en la opuesta. Esto confirma que el pulso no es una variable aislada, sino que está relacionada con el conjunto de los demás signos clínicos.

Para determinar el número óptimo de clusters (K), se utilizó el Método del Codo (Figura 11). El gráfico muestra una caída pronunciada en la inercia (WCSS) de K=2 a K=3, y una segunda caída de K=3 a K=4. A partir de K=4, la curva se suaviza, indicando un rendimiento decreciente.Basado en este "codo" y en el contexto del problema (donde la variable resultado tiene 3 categorías), seleccionamos un valor de K=3 para el clustering. El objetivo será ver si estos 3 clusters se alinean con los 3 posibles resultados del paciente ('Vivió', 'Murió', 'Eutanasiado')."

Fig. 11:

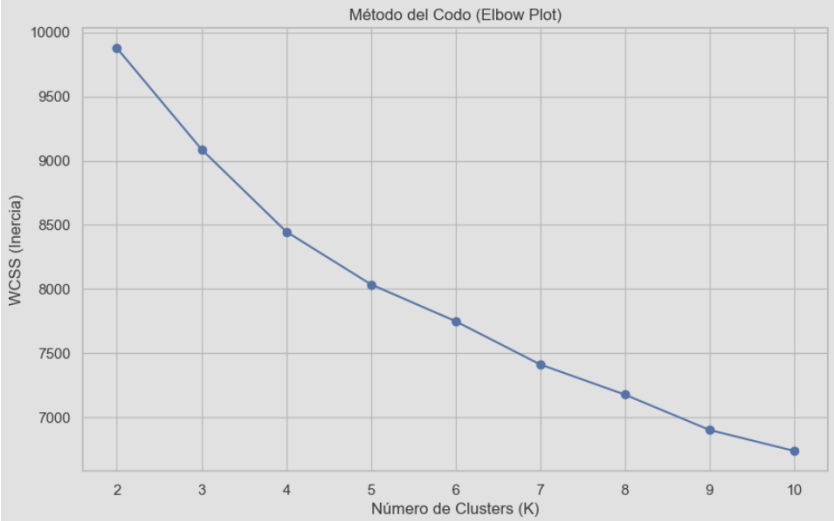


Fig. 11: gráfico del codo para determinar la cantidad de clusters conveniente

Aplicando el K-Means con K=3, el algoritmo identificó tres grupos distintos. Para visualizar estos grupos, se utilizó T-SNE (Figura 12). El gráfico muestra que los tres clusters están bien definidos y separados espacialmente, lo que confirma que el algoritmo encontró patrones reales en los datos. El Cluster 0 es el grupo más grande, mientras que los Clusters 1 y 2 representan dos perfiles de pacientes distinto

Fig.12:

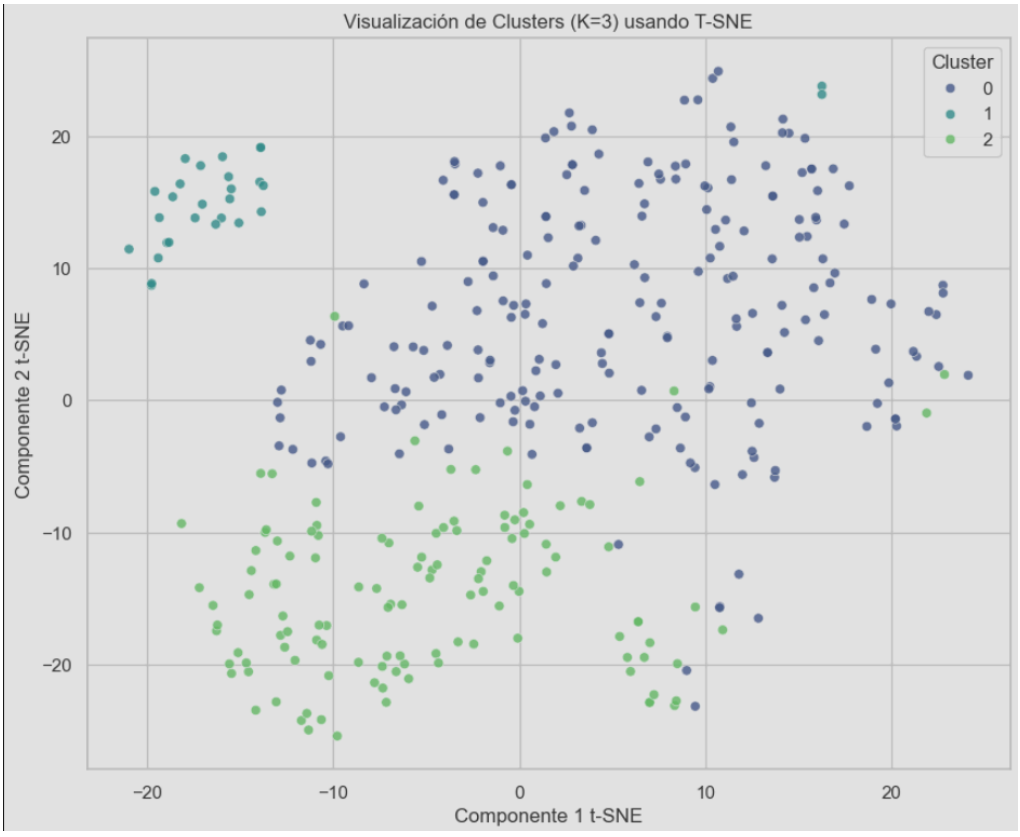


Fig. 12: el gráfico muestra claramente 3 tipos de clusters bien definidos

Luego vemos que mediante la siguiente tabla K-means logró identificar lo que representan esos 3 clusters.

Cluster 2: Los que sobrevivieron (casos leves)

Este es el grupo con el mejor pronóstico

- Resultado: El 93% de sus caballos "Vivió".
- Pulso: Tienen el pulso mediano más bajo (48 lpm), el más cercano a lo normal.
- Dolor: Casi no reportan dolor severo (solo 1-2%).

Cluster 1: los críticos (shock severo)

Este grupo representa los casos más extremos y de alto riesgo.

- Pulso: Su característica principal es un pulso mediano extremo de 120 lpm, un claro indicador de shock severo.
- Resultado: Es el grupo con la mayor tasa de mortalidad (50% "Murió").

Cluster 0: los graves (casos dolorosos)

Este es un grupo de mal pronóstico, definido por un alto nivel de dolor.

- Dolor: Es el grupo que concentra el dolor más alto (el 40% de sus miembros tenían 'Dolor Intenso' o 'Continuo').
- Resultado: El 55% de sus miembros no sobrevivió (33% "Murió" + 22% "Eutanasiado").
- Pulso: Su pulso mediano (70 lpm) es anormalmente alto, aunque no tan extremo como el Cluster 1.

```
Cluster
0        70.00               38.10                    28.00
1       120.00               38.30                    46.00
2        48.00               38.10                    24.00

         vol_cel_empaquetado
Cluster
0                      45.00
1                      42.00
2                      42.00

------------------------------------------------------------

--- 2. PERFIL CATEGÓRICO (% de Resultado por Cluster) ---
resultado  1.00  2.00  3.00
Cluster
0          0.45  0.33  0.22
1          0.46  0.50  0.04
2          0.93  0.03  0.03

------------------------------------------------------------

--- 3. PERFIL CATEGÓRICO (% de Dolor por Cluster) ---
dolor    1.00  2.00  3.00  4.00  5.00
Cluster
0        0.01  0.22  0.37  0.19  0.21
1        0.07  0.18  0.50  0.11  0.14
2        0.36  0.19  0.42  0.02  0.01
```

Estos agrupamientos nos van a permitir el planteamiento de las hipótesis bivariadas que haremos a continuación

3. HIPÓTESIS PLANTEADAS,RESOLUCIÓN Y VALIDACIÓN:

3.1. HIPÓTESIS 1: (UNIVARIADA)El pulso promedio de los caballos enfermos es anormalmente alto

3.1.1: Definición de la hipótesis:

- Esta hipótesis se basa en la Figura 7 del análisis exploratorio, donde se observó una media de pulso cercana a 71 lpm. La documentación del dataset establece que un pulso normal para un caballo adulto es de 30-40 lpm. Esta prueba busca confirmar formalmente si la media observada en nuestra muestra es significativamente más alta que el límite superior normal (40 lpm).

3.1.2: Estrategia de abordaje:

- Vamos a utilizar el test de los rangos con signo de Wilcoxon, la cual como vimos en la teoría es una prueba no paramétrica, en la hipótesis buscamos comparar el valor central de nuestra muestra(pulso) contra el valor de referencia, debido a que el rango normal es entre 30-40 lpm, para probar que el pulso es anormalmente alto, vamos a testear si es mayor que 40 debido a que es el límite máximo normal.

A continuación planteamos la hipótesis: como la muestra es no paramétrica, testeamos con la mediana

- H0 : M = 40
- H1 : M > 40
- usamos un nivel de significancia = 0,05
- luego validamos la normalidad de los datos, usando Shapiro-Wilk

3.1.3: Resultados obtenidos:

Al ejecutar en la notebook las pruebas,obtuvimos los siguientes resultados:

1. test de shapiro-wilk para validar la normalidad
    - p_valor = 7.52e-15
    - conclusión: el p_valor es claramente mucho menor a 0,05, por lo que la variable pulso no sigue una distribución normal, lo cual validamos la elección de usar un test no paramétrico
2. test de wilcoxon:
    - Mediana de pulso en la muestra: 60.0 lpm
    - Valor de referencia (mediana H0): 40 lpm
    - Resultado: El P-valor es 1.71e-57

El p_valor que obtuvimos en el test de wilcoxon es mucho menor al nivel de significancia, esto nos indica que la probabilidad de observar una mediana muestral de 60,0 lpm si la mediana real de la población fuera 40 lpm es casi nula. Por lo tanto, rechazamos la hipótesis nula (H0).
confirmamos la hipótesis de que la mediana del pulso de los caballos en el dataset (60,0 lpm) es estadísticamente superior al rango normal de 30-40 lpm

3.2. HIPÓTESIS 2: (UNIVARIADA) La temperatura rectal de los caballos enfermos es superior a la normal.

3.2.1: Definición de la hipótesis:

- Esta hipótesis se basa en la Figura 9 del análisis exploratorio. En dicho gráfico, se observó que la distribución de la temperatura rectal estaba centrada "alrededor de los 38°C", lo cual se interpretó como "normal o ligeramente elevada".La documentación del dataset establece que la temperatura normal de un caballo es 37.8°C. Esta prueba busca confirmar formalmente si la mediana de la temperatura observada en nuestra muestra (que parece estar cerca de 38°C) es estadísticamente superior a ese valor normal de 37.8°C.

3.2.2: Estrategia de abordaje:

H0: M = 37.8 (La mediana de la temperatura rectal es 37.8°C).

H1: M > 37.8 (La mediana de la temperatura rectal es mayor a 37.8°C).

- En primera instancia vamos a validar mediante el test de shapiro wilk si los datos no son normales, para luego decidir qué test vamos a usar para validar la hipótesis
- En segundo lugar sabiendo que los datos no son normales,elegimos el test de wilcoxon

3.2.3: Resultados obtenidos:

- Al ejecutar las pruebas, el test de Shapiro-Wilk (p=6.43e-11) confirmó que la variable 'temperatura_rectal' no sigue una distribución normal. Por lo tanto, se procedió con el test de Wilcoxon. El resultado (p=1.06e-25) fue significativamente menor que 0.05, lo que nos lleva a rechazar la hipótesis nula (H0). Se confirma la hipótesis H2: la mediana de la temperatura rectal de la muestra (38.1°C) es estadísticamente superior a la temperatura normal de 37.8°C, validando la observación del EDA."

3.3. HIPÓTESIS 3: (BIVARIADA): Existe una asociación significativa entre el tipo de dolor reportado y si la lesión es quirúrgica.

3.3.1: Definición de la hipótesis:

- Esta hipótesis surge de la documentación del dataset, que sugiere que el dolor y la necesidad de cirugía están relacionados. Sin embargo, la misma documentación advierte explícitamente que la variable dolor no debe tratarse como una variable ordenada, ya que tiene categorías como "deprimido" (2) y el dolor puede estar enmascarado por analgésicos.
- Por lo tanto, no podemos buscar una correlación (donde si uno sube, el otro sube). En su lugar, buscaremos una asociación (Test de Independencia).

3.3.2: Estrategia de abordaje:

- Para validar esta hipótesis, vamos a probar la asociación entre dos variables categóricas no ordenadas (dolor y lesión_quirúrgica).
- Como se trata de dos variables categóricas, el test estadístico adecuado es el Test Chi-Cuadrado

  H0: Las variables dolor y lesión_quirúrgica son independientes (no hay asociación).
  H1: Las variables dolor y lesión_quirúrgica NO son independientes (sí hay asociación)

3.3.3: Resultados obtenidos:

- Al ejecutar el Test Chi-Cuadrado de Independencia, se generó primero la tabla de contingencia que cruza las 5 categorías de dolor con las 2 de lesión_quirúrgica:

```
--- Datos para H3 (Test Chi-Cuadrado) ---
Tabla de Contingencia (Observada):
lesión_quirúrgica   1   2
dolor
1.0                  9  40
2.0                 50  26
3.0                 89  56
4.0                 38   8
5.0                 44   6
----------------------------------------
--- Resultados Validación H3 ---
Estadístico Chi-Cuadrado: 63.17
P-valor: 6.244384743189642e-13
Grados de libertad (dof): 4
```

- Los resultados del test estadístico fueron concluyentes:
  Estadístico Chi-Cuadrado: 63.17
- P-valor: 6.24e-13
- Dado que el p-valor (6.24e-13) es menor que el nivel de significancia (0.05), se rechaza la hipótesis nula (H0).
- Concluimos que existe una asociación estadísticamente significativa entre el tipo de dolor reportado por el caballo y si su lesión requiere cirugía.

3.4. HIPÓTESIS 4: (BIVARIADA):

Existen diferencias significativas en el pulso de los caballos entre los distintos grupos de resultado (Vivió, Murió, Eutanasiado).

3.4.1: Definición de la hipótesis:

- Esta hipótesis prueba una de las relaciones clínicas más importantes: si el pulso (un indicador clave de shock circulatorio) está relacionado con el resultado final del caballo.
- Se comparará la variable numérica (pulso) contra las 3 categorías de la variable resultado (1=Vivió, 2=Murió, 3=Eutanasiado).

- La expectativa clínica es que los grupos con peores desenlaces ('Murió' y 'Eutanasiado') tengan medianas de pulso significativamente más altas que el grupo que 'Vivió'. Con esta hipótesis vamos a buscar validar esa observación estadísticamente.

3.4.2: Estrategia de abordaje:

- Para esta validación, se busca comparar una variable numérica (pulso) entre más de 2 grupos independientes (resultado tiene 3 categorías: 'Vivió', 'Murió', 'Eutanasiado').
- Para seleccionar el test estadístico correcto primero vamos a validar la normalidad mediante el test de shapiro wilk.
- luego si los datos no son normales utilizaremos el test de Kruskal-Wallis o caso contrario el test de anova

3.4.3: Resultados obtenidos:

- Al calcular el supuesto de normalidad obtuvimos un p_valor para cada grupo

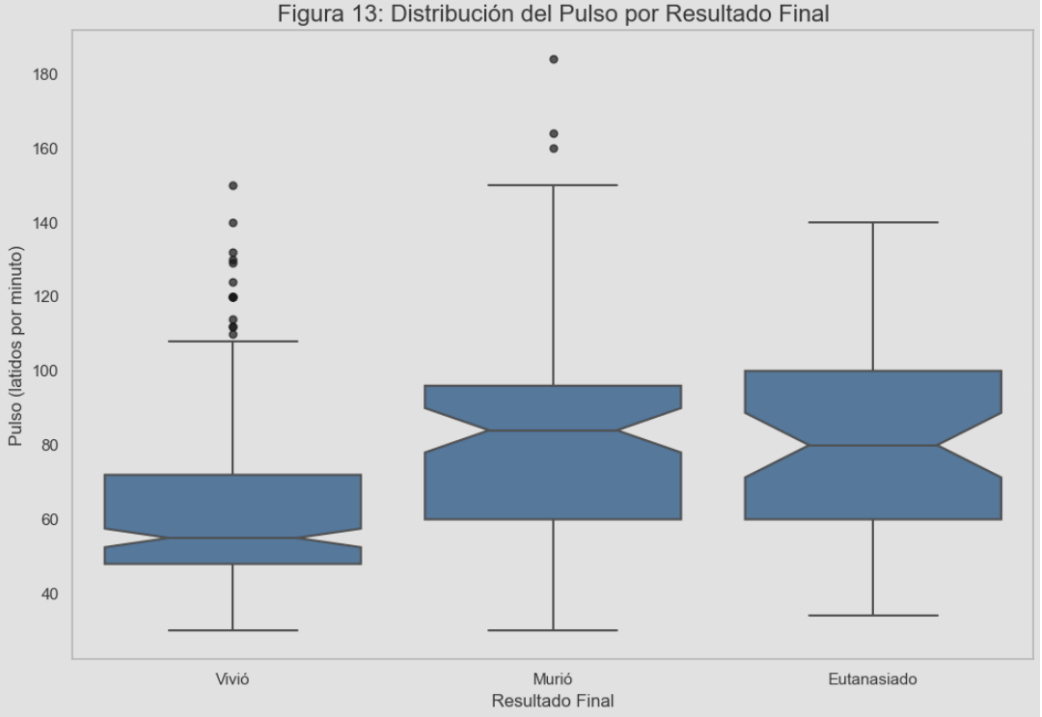
  Grupo 'Vivió' (n=225):
  P-valor Shapiro: 1.30e-14
  >> NO es Normal
  Grupo 'Murió' (n=89):
  P-valor Shapiro: 1.57e-04
  >> NO es Normal
  Grupo 'Eutanasiado' (n=52):
  P-valor Shapiro: 3.39e-01
  >> Es Normal
- Dado que al menos dos de los tres grupos no cumplen el supuesto de Normalidad, el uso de ANOVA lo descartamos y utilizamos el Test de Kruskal-Wallis, con un nivel de significancia de 0.05.

  H0: Las medianas del pulso son iguales en los 3 grupos de resultado.
  H1: Al menos una de las medianas del pulso es diferente de las demás.
- Al ejecutar el test nos dio los siguientes resultados:

```
--- Datos para H4 (Kruskal-Wallis) ---
Mediana Pulso 'Vivió': 55.00 lpm
Mediana Pulso 'Murió': 84.00 lpm
Mediana Pulso 'Eutanasiado': 80.00 lpm
----------------------------------------
--- Resultados Validación H4 ---
Estadístico H (Kruskal-Wallis): 58.17
P-valor: 2.3374616438295963e-13
```

- Dado que el p-valor es menor que el nivel de significancia (0.05), se rechaza la hipótesis nula (H0). Concluimos que existe una diferencia estadísticamente significativa en la mediana del pulso entre al menos uno de los grupos de resultado, validando así la H4.

- luego realizamos un boxplots para confirmar que lo que concluimos es correcto



- La Figura 13 muestra visualmente que la mediana del pulso para los caballos que 'Vivieron' es notablemente más baja que la de los grupos 'Murió' y 'Eutanasiado'. Las ranuras de las cajas confirman visualmente esta diferencia significativa en las medianas.

3.5. HIPÓTESIS 5:(MULTIVARIADA): El pulso de un caballo está significativamente influenciado por una combinación lineal de sus principales signos clínicos (temperatura_rectal, frecuencia_respiratoria, dolor y resultado)

3.5.1: Definición de la hipótesis:

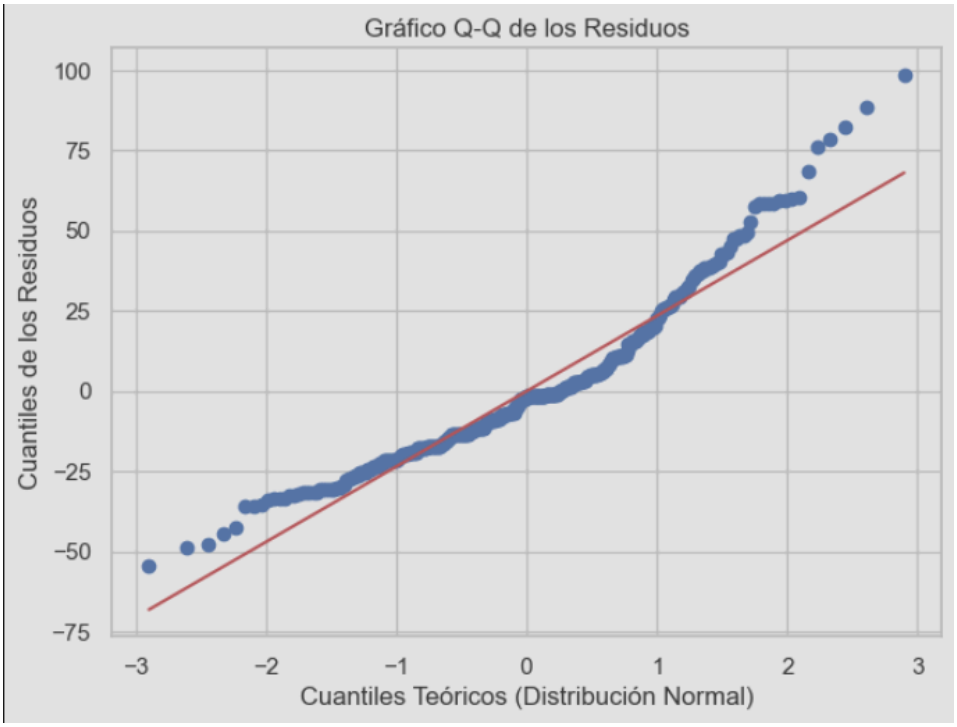
- En esta hipótesis multivariada afirmamos que el pulso (la variable dependiente) no solo se relaciona con una variable, sino que su valor puede ser explicado por un modelo lineal combinado de otras variables predictoras (las independientes).

3.5.2: Estrategia de abordaje:

- Para validar la hipótesis usamos el modelo de regresión lineal en donde la variable dependiente(Y) es el pulso y variables independientes(X) la temperatura, dolor y resultado
- También medimos el R^2 para ver cuánto del pulso explica el modelo.
- Primero validamos el supuesto de normalidad usando el Test de Shapiro-Wilk.
    1. Si los residuos son normales ($p > 0.05$): El modelo es válido. Aceptamos H5.
    2. Si los residuos no son normales ($p < 0.05$): El modelo no es válido. La relación no es lineal. Rechazamos H5.

3.5.3: Resultados obtenidos:

- El modelo arrojó un coeficiente R^2 de 0.2122. Esto indica que este modelo, logra explicar un 21.22% de la varianza de la variable pulso.
- Luego se evaluó si el modelo era estadísticamente válido testeando la normalidad de sus residuos.



- El Gráfico Q-Q plot muestra que los residuos (puntos azules) no se distribuyen de forma normal. No siguen la línea roja teórica, sino que se desvían claramente en los extremos.
- Con el Test de Shapiro-Wilk sobre los residuos confirmamos lo que se ve en el gráfico, el p-valor es de 1.03e-11
- Dado que el p-valor (1.03e-11) es menor que el nivel de significancia (0.05), se rechaza la hipótesis de normalidad de los residuos.
- Aunque el R^2 indica que existe una relación entre las variables, el hecho de que el supuesto de normalidad falle invalida el modelo lineal.
- No se puede afirmar que el pulso esté influenciado por una combinación lineal de dolor y resultado; la relación real es más compleja.

3.6. HIPÓTESIS 6: (MULTIVARIADA):

Las variables pulso, dolor y vol_cel_empaquetado tienen un poder predictivo estadísticamente significativo para clasificar el desenlace de un caballo ('Vive' vs. 'No Vive').

3.6.1: Definición de la hipótesis:

- Con esta hipótesis buscamos construir un modelo predictivo utilizando únicamente las variables que demostraron ser más relevantes en el análisis exploratorio y bivariado

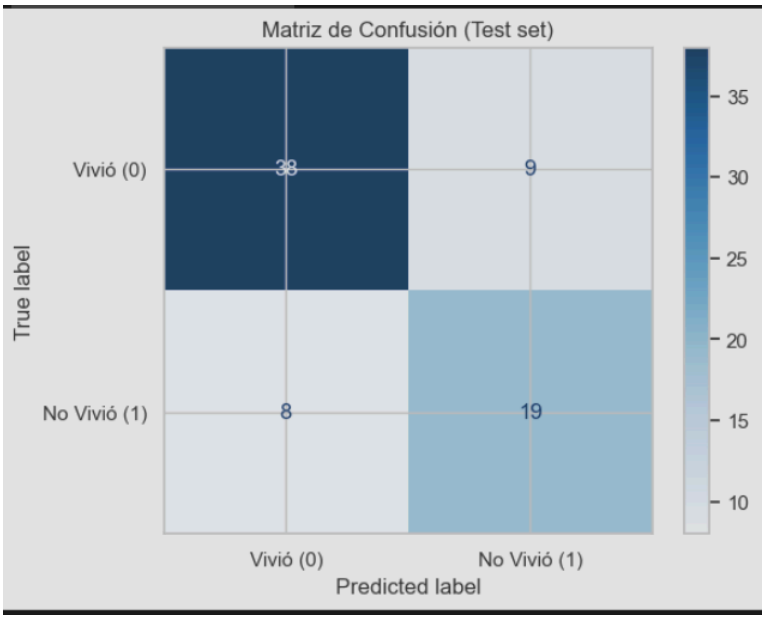
- vol_cel_empaquetado: Un signo vital clave que apareció en el heatmap como un fuerte indicador de shock/deshidratación, ambos ligados al resultado.

3.6.2: Estrategia de abordaje:

- Para validar esta hipótesis predictiva, utilizamos el modelo de regresión logística
- la variable resultado(Y) pasará a ser binaria, donde 0 = 'vivió' y 1 = ‘no vivió, se identifica que no vivió si fue eutanasiado o se ubica en la categoría ‘murio’
- las variables pulso, vol_cel_empaquetado y dolor componen la (X)
- Vamos a medir el Accuracy (exactitud) y analizar la matriz de confusión para ver en qué se acierta y en qué se equivoca.
- Vamos a aceptar H6 si el modelo obtiene una exactitud alta en el conjunto de prueba.

3.6.3: Resultados obtenidos:

- Se entrenó el modelo de Regresión Logística sobre el 80% de los datos (292 muestras) y se lo evaluó en el 20% de datos de prueba (74 muestras) que el modelo nunca había visto.
- El resultado principal que obtuvimos es el Accuracy (exactitud), que mide el porcentaje total de aciertos del modelo el cual nos dio un 0.7703 (77.03%), es decir que el modelo muestra un gran poder predictivo
- Por último realizamos la matriz de confusión que muestra la cantidad de aciertos y errores del modelo sobre el test



- Aciertos: el modelo predijo correctamente 38 casos de 'Vivió' y 19 de 'No Vivió'.
- Errores: el modelo se equivocó en 17 casos (9 Falsos Negativos y 8 Falsos Positivos).

- Como conclusión el modelo obtiene un 77.03% de precisión en datos nuevos, validando que el conjunto de variables (pulso, vol_cel_empaquetado y dolor) sí tiene un poder predictivo estadísticamente significativo y por lo tanto se acepta la H6.

4. CONCLUSIÓN FINAL:

Para finalizar, este trabajo nos resultó de gran interés y utilidad. Nos permitió plasmar los conocimientos adquiridos tanto en las clases teóricas como en las prácticas, aplicándolos a un dataset que encontramos particularmente relevante.
En cuanto a la resolución, iniciamos con un análisis general de los datos, seguido de una etapa de limpieza. En esta fase, aplicamos las diversas técnicas aprendidas en la cursada para asegurar la calidad y fiabilidad de los datos..
Posteriormente, llevamos a cabo un análisis más profundo que escaló en complejidad. Arrancamos con un análisis univariado para entender cada variable individualmente. Luego, en el análisis bivariado, comenzamos a identificar las relaciones claves, dándonos cuenta de que variables como dolor y pulso eran las que más predominaban en el conjunto de datos.
Finalmente, todo concluyó en el análisis multivariado. En esta etapa integramos todos los hallazgos anteriores para plantear un conjunto de hipótesis formales, las cuales validamos (o rechazamos) utilizando las metodologías de regresión y clasificación correspondientes.

BIBLIOGRAFÍA:

https://archive.ics.uci.edu/dataset/47/horse+colic
El archivo horse-colic.names
Material de la cátedra(collabs y filminas)