Fundamentos de la Ciencia de Datos

Trabajo Práctico Especial

Grupo 03

Oñatibia, Manuel
Ramundo, Alexis Nehuen
Padilla, Juan Pablo

1. Introducción

El cólico equino representa uno de los desafíos más significativos en la medicina veterinaria, siendo la principal causa de muerte en caballos. No se trata de una enfermedad única, sino de un término que agrupa un conjunto de síndromes dolorosos abdominales que pueden escalar rápidamente a emergencias médicas de alta gravedad. El pronóstico de un caballo con cólico depende de una detección temprana y de la toma de decisiones críticas por parte del personal veterinario, incluyendo la necesidad de una intervención quirúrgica.

En este contexto, el análisis de datos clínicos históricos se vuelve fundamental para entender los patrones que conducen a diferentes desenlaces. El conjunto de datos utilizado para este TPE es el "Horse Colic Dataset", con origen de repositorio de Machine Learning de UC Irvine, donde su contenido representa registros clínicos de caballos con cólico.

El conjunto de datos obtenido del repositorio contiene los datasets ya particionados para más adelante probar el modelo logrado en la hipótesis 6, (horse-colic.data) consiste en 300 instancias o tuplas (filas), además del (horse-colic.test) que consiste de 68 tuplas, cada una representando el caso de un caballo tratado por cólico, y se compone de 28 atributos. Estas variables incluyen una mezcla de mediciones clínicas numéricas (pulse y rectal_temperature,etc), observaciones categóricas ordinales (como el nivel de pain) e identificadores (hospital_number).

La utilidad de este conjunto de datos es un gran desafío. Un análisis inicial reveló problemas significativos que son comunes en los datos del mundo real: una alta cantidad de **valores nulos** (representados por '?'), la necesidad de corregir tipos de datos y la presencia de variables irrelevantes para el análisis predictivo. Como vamos a detallar en la sección de limpieza, fue necesario un pre-procesamiento exhaustivo de los datos, incluyendo la eliminación de columnas con exceso de nulos y la imputación de valores faltantes (usando la mediana y KNN para datos numéricos, y la moda para categóricos) para hacer el dataset utilizable.

Este informe está organizado mediante las siguientes secciones:

- **2: Análisis exploratorio de los datos:** Se presenta una caracterización detallada de los atributos, su distribución y el proceso de limpieza y preparación de los datos.
- **3: Hipótesis planteadas y resolución:** Se desarrollan y validan estadísticamente seis hipótesis (univariadas, bivariadas y una multivariada) para descubrir patrones y relaciones clave dentro de los datos.
- **4: Conclusiones:** Se resumen los hallazgos principales del análisis y se discute la validez de las hipótesis en el contexto del problema.

2. Análisis exploratorio de los datos

El primer paso en cualquier proyecto de ciencia de datos es realizar un Análisis Exploratorio de Datos (EDA). Necesitamos entender profundamente nuestros datos y ver que relaciones y tendencias encontramos antes de poder limpiarlos o plantear hipótesis. Este análisis se centró en entender la estructura del dataset, la naturaleza de sus atributos y, fundamentalmente, en identificar los problemas de calidad de datos que debían ser resueltos.

2.1 Carga y caracterización Inicial de los atributos

El dataset horse-colic.data se presentó en un formato de texto plano sin procesar. La carga inicial requirió una configuración específica:

- **Sin encabezado:** El archivo no contenía una fila de header, por lo que los 28 nombres de atributos debieron asignarse manualmente.
- **Separador:** El delimitador de columnas era el espacio (delim_whitespace=True).
- **Valores nulos:** Los valores faltantes (missing values) no estaban vacíos, sino representados explícitamente por el carácter '?'. Se configuró la carga para que estos se interpretaran correctamente como NaN (Not a Number).

El dataset resultante consta de **300 instancias (filas)** y **28 atributos (columnas)**.Hemos asignado los nombres de las 28 columnas basándonos en la documentación encontrada y sus descripciones. Los atributos se pueden clasificar en:

- **Identificador:** hospital_number. Esta variable es un ID único para cada caballo y no aporta valor predictivo al dataset, por lo que la marcamos para eliminarla más adelante.
- **Variable objetivo** (principal): **outcome**. Es categórica con 3 valores: 1 (vivió), 2 (murió), 3 (fue sacrificado). Esta será la variable que generalmente intentaremos predecir.
- **Variables Numéricas (Continuas/Discretas):**
  - rectal_temperature: Temperatura en grados Celsius.
  - pulse: Ritmo cardíaco (pulsaciones por minuto).
  - respiratory_rate: Tasa de respiración.
  - packed_cell_volume: Volumen de células rojas en sangre (%).
  - total_protein: Proteína total (gms/dL).
  - abdomcentesis_total_protein: Proteína en fluido abdominal.
- **Variables Categóricas (Nominales/Ordinales):**
  - surgery
  - age
  - temperature_of_extremities
  - peripheral_pulse
  - mucous_membranes
  - capillary_refill_time
  - pain
  - peristalsis
  - abdominal_distension

- nasogastric_tube
- nasogastric_reflux
- nasogastric_reflux_PH
- rectal_examination_feces
- abdomen
- abdominocentesis_appearance
- outcome
- surgical_lesion

2.2 Comportamiento de los datos, nulos y outliers

- **Valores nulos:** En la descripción de la página de donde obtenemos el dataset indica "Missing Values: Yes" para muchas variables. Al cargar los datos reemplazando '?' por NaN (Not a Number), confirmamos que hay una cantidad significativa de nulos.
  - Variables críticas por nulos: Algunas variables tienen un porcentaje altísimo de nulos (ej. nasogastric_reflux_PH, abdomcentesis_total_protein y abdominocentesis_appearance).

    Esto es algo a tener en cuenta: si tienen más del 40-50% de nulos, imputar valores puede ser perjudicial, y capaz debamos eliminarlas más adelante si no las utilizamos para plantear hipótesis.

  - Variables con pocos nulos: Variables como rectal_temperature o pulse, tienen menos nulos y podrán ser imputadas dependiendo de si las usaremos en las hipótesis.
- **Distribución y outliers (valores atípicos):**
  - pulse (Pulso): En la búsqueda del conocimiento de dominio indica que un pulso normal en reposo para un caballo es de 28-44 lpm. Nuestro dataset muestra valores de 40, 88 y hasta 184. Un valor de 184 es un outlier estadístico, pero clínicamente es posible (indica shock severo o dolor extremo).

    Decisión de análisis**:** No lo eliminaremos, ya que es un indicador de gravedad y probablemente muy predictivo del outcome.

  - rectal_temperature (Temperatura Rectal): La normal es ~37.8°C. Vemos valores como 38.5, 39.2, 37.3. Estos parecen estar en un rango clínicamente razonable (fiebre o temperatura normal). Un boxplot seguramente nos sirve para ver si hay outliers extremos (ej. 30°C o 45°C), que serían errores de tipeo al cargar en la base.

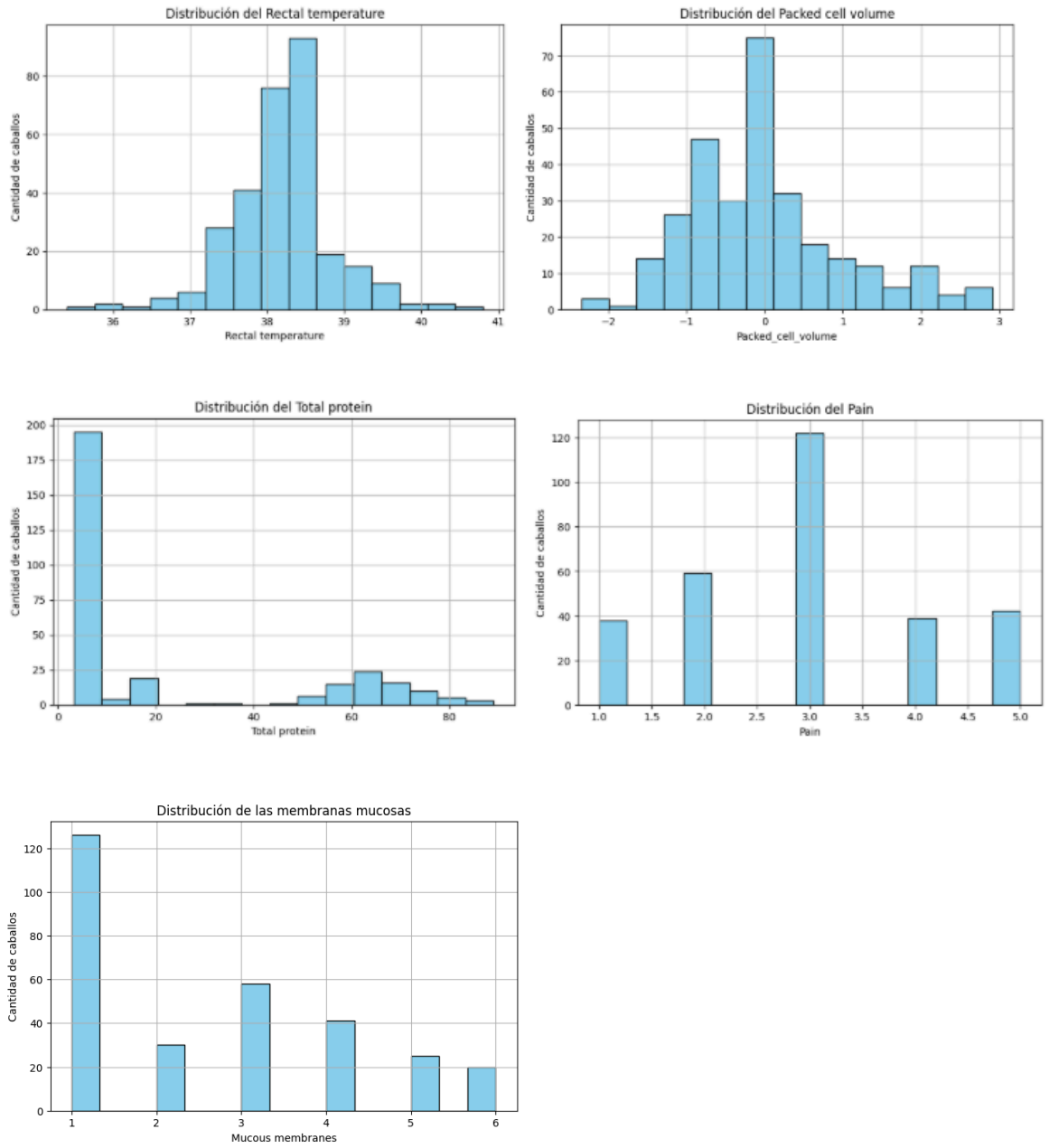


Imagen 1: Histograma de algunas variables de interés

Atributo	Cantidad de nulos	Porcentaje
Nasogas_ref_PH	247	82.34
Abdominoc_to_prot	198	66
Abdominoc_appear	165	55
Abdomen	118	39.33
Nasogas_reflux	106	35.36
Nasogas_tube	104	34.67
Feces	102	34
Periph_pulse	69	23
Rect_temp	60	20
Respir_rate	58	19.33
Temp_of_extrem	56	18.66
Abdom_dist	56	18.66
Pain	55	18.33
Mucous_membr	47	15.66
Peristal	44	14.66
Total_prot	33	11
Capill_ref_time	32	10.66
Packed_cell_volume	29	9.66
Pulse	24	8
Outcome	1	0.33
Surg?	1	0.33
Age	0	0
Surgical_lesion	0	0
Type_of_lesion	0	0

Tabla 1: Representación ordenada de los atributos de forma descendente por la cantidad de nulos identificados

2.3 Limpieza de datos

Basados en el EDA que hemos aprendido durante la cursada, debemos pre-procesar los datos para que sean utilizables por los modelos y técnicas estadísticas en las hipótesis. No podemos calcular una media si hay valores basura como ‘?’.

A continuación, detallamos las acciones correctivas implementadas y su justificación:

Carga de datos:

Acción: Cargar horse-colic.data usando pandas.read_csv, especificando header=None, delim_whitespace=True, y na_values='?'.

Justificación**:** Esto interpreta correctamente el archivo, asigna NaN a los '?' y separa las columnas por espacios.

Asignación de nombres:

Acción: Asignar los 28 nombres de las columnas manualmente según la información que obtuvimos del dataset

Justificación**:** Necesario para que el dataset sea legible y podamos referirnos a variables como pulse/pulso en lugar de "columna 5".

Eliminación de columnas irrelevantes:

Acción: Eliminar la columna hospital_number.

Justificación**:** Es un id único que no tiene valor predictivo. Incluirlo puede confundir a algunos algoritmos, no le vemos el sentido de que aporte a los datos.

Corrección de tipos:

Debido a la presencia inicial de '?', columnas como pulse se cargaron como tipo object (texto). Tras la imputación, se forzó su conversión a tipo float (numérico) para permitir cálculos matemáticos. Además Dividimos la variable Type of lesion en 4 columnas para su correcta lectura e interpretación.

Manejo de nulos:

Como tenemos 300 instancias. Eliminar filas (caballos) con nulos reduciría demasiado nuestro dataset. La estrategia preferida es la imputación (relleno) de valores, diferenciando por tipo de variable.

- **Acción en variables numéricas:**

  Para pulse, rectal_temperature, packed_cell_volume y otras numéricas. Imputamos los valores nulos usando la **mediana** de cada columna. Todas las variables con **0-10%** de nulos son las afectadas.

Justificación (de la Mediana)**:** Usamos la mediana en lugar de la media porque es más robusta a outliers. Por ejem, el pulso de 164 sesgaría la media hacia arriba, pero no afecta a la mediana.

Para (nombre de variables de ejemplos), imputamos los valores nulos usando el algoritmo de los vecinos más cercanos KNN. Todas las variables con **10-25%** de nulos son las afectadas.

Justificación: Al ser un porcentaje más elevado y con más peso de nulos los que afectan a estas variables, no nos conviene utilizar la mediana porque reduce demasiado la varianza. Usamos **KNN** porque analiza las otras variables de la fila para encontrar los vecinos (filas más similares en dataset) y utiliza los valores de estos para calcular el reemplazo. De esta manera preserva mejor la estructura de las relaciones, reduciendo significativamente el sesgo que se introduciría en el análisis posterior.

- **Acción en variables categóricas:**

  Para variables como surgery, pain, temperature_of_extremities, etc. Imputamos los valores nulos usando la **moda** (el valor más frecuente) de cada columna.Todas las variables con **0-25%** de nulos son las afectadas.

  Justificación: Porque usando la moda es la técnica estándar para imputar variables categóricas, asumiendo que el valor faltante es probablemente el más común.

- **Acción en las columnas con >40% nulos:**

  Las variables abdomcentesis_total_protein y abdominocentesis_appearance,nasogastric_reflux_PH tienen demasiados nulos.

  Justificación: Imputar casi la mitad del dataset con un solo valor (usando media/moda) nos llevaría a provocar un sesgo enorme.

  Decisión final que tomamos**:** Estas columnas serán **eliminadas** del análisis. Todas las variables con **40-100%** de nulos son las afectadas.

Una vez completadas estas operaciones, obtuvimos un conjunto de datos limpio, coherente y listo para la fase de validación de hipótesis, asegurando que los resultados estadísticos no estuvieran sesgados por datos faltantes o mal interpretados.

3. Hipótesis planteadas y resolución

Ahora presentamos las seis hipótesis formuladas a partir de las observaciones del análisis exploratorio. Cada hipótesis busca validar una intuición sobre el comportamiento del dataset, desde características univariadas básicas hasta un modelo predictivo multivariado.

Luego de llenar de preguntas sobre las variables e interrogar a la IA que utilizamos como nuestro asistente veterinario para que nos entendamos con el dominio del problema. Aca desarrollamos las 6 hipótesis planteadas y que encontramos interesantes Siguiendo la estructura del template, para cada hipótesis se detallará su definición, la estrategia estadística seleccionada para su validación y la discusión de los resultados obtenidos.

3.1 Hipótesis 1 (Univariada)

variables: categórica (cuali)

3.1.1. Definición de la hipótesis

"La mediana del nivel de severidad de las membranas mucosas es mejor (menor) que el umbral clínico de referencia igual a 3."

Justificación: La variable Mucous_membr representa un indicador ordinal del estado circulatorio del caballo, donde valores mayores implican mayor compromiso clínico. El valor clínico de 3 (“pale pink”) se utiliza como umbral que señala alteración fisiológica relevante. Contrastar la mediana contra este límite permite evaluar si la condición general de la muestra presenta un estado circulatorio dentro de parámetros aceptables.

3.1.2 Estrategia de abordaje

Técnica: Test de los Signos (Prueba de la Mediana para una Muestra).

Justificación: El Test de los Signos es un método no paramétrico apropiado para contrastar la mediana observada de una variable ordinal contra un valor fijo de referencia. No requiere supuestos de normalidad ni igualdad de varianzas, y se basa en la distribución binomial exacta para evaluar si el número de observaciones por encima y por debajo del valor hipotético (M0 = 3) difiere del esperado bajo H0.

Pasos de limpieza: dataframe utilizado (original o preprocesado)

La variable Muscous_mebr tiene 47 nulo / 15.66%

Usamos el data general

Pasos:

- Hipótesis Nula H0: La mediana de la severidad de las mucosas es igual o peor que el umbral clínico de 3.

$$H0: M => 3\}$$

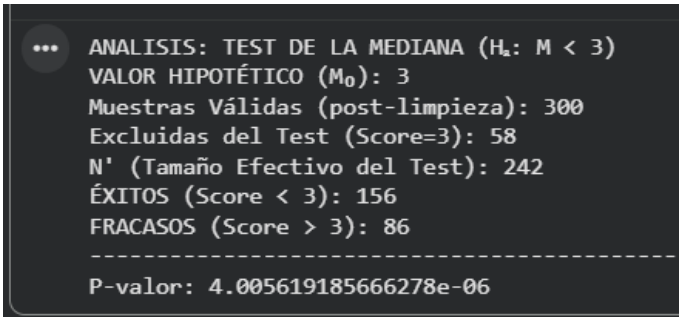
- Hipótesis Alternativa Ha: La mediana de la severidad de las mucosas es mejor (menor) que el umbral clínico de 3.

$$Ha: M < 3$$

- Se seleccionó la variable Mucous_membr y se identificaron los valores exactamente iguales al umbral clínico (3). Estos valores fueron excluidos del recuento efectivo según lo exige el Test de los Signos.
- Se computó la cantidad de valores por encima y por debajo del umbral M0 = 3, y se evaluó la probabilidad exacta mediante la distribución binomial para determinar si la mediana es significativamente menor que el valor clínico de referencia.

3.1.3 Resultados obtenidos y discusión

Resultados: El Test de los Signos arrojó un **p-valor = 0.000004**, extremadamente menor al nivel de significancia habitual (α = 0.05).



Por lo tanto, **se rechaza la Hipótesis Nula (H0)**.

Discusión: Existe evidencia estadística altamente significativa para concluir que la mediana del nivel de severidad de las membranas mucosas es mejor (menor) que el umbral clínico de referencia. Esto indica que, en términos generales, la condición circulatoria de los animales evaluados es clínicamente más favorable que el punto de corte que indica deterioro hemodinámico.

En base a esto, **se confirma la hipótesis inicial**. La mediana del nivel de severidad de las membranas mucosas es mejor (menor) que el umbral clínico de referencia igual a 3.

3.2 Hipótesis 2 (Univariada)

variables: numérica (cuanti)

3.2.1 Definición de la hipótesis

"La media del pulso (pulse) de los caballos en el dataset es significativamente superior al pulso normal en reposo (44 lpm)."

Justificación: El conocimiento de dominio indica que el dolor y el shock (comunes en el cólico) provocan taquicardia (aumento del ritmo cardíaco).Queremos verificar si esta condición es generalizada en la muestra.

Un caballo sano en reposo tiene 28-40 lpm. Comprobar que la media de esta población es alta (ej. 60, 70 u 80 lpm) no solo confirma que estamos tratando con pacientes enfermos, sino que cuantifica el nivel promedio de estrés fisiológico al que el hospital se enfrenta. Es la línea de base para evaluar todo lo demás.

3.2.2 Estrategia de abordaje

Técnica: Test T de una muestra

Justificación: Esta prueba estadística es la herramienta apropiada para comparar la media de una muestra (el pulse promedio de nuestros 300 caballos) contra un valor poblacional teórico o de referencia (mu = 44, el límite superior normal en reposo).

Pasos de limpieza: dataframe utilizado (original o preprocesado)

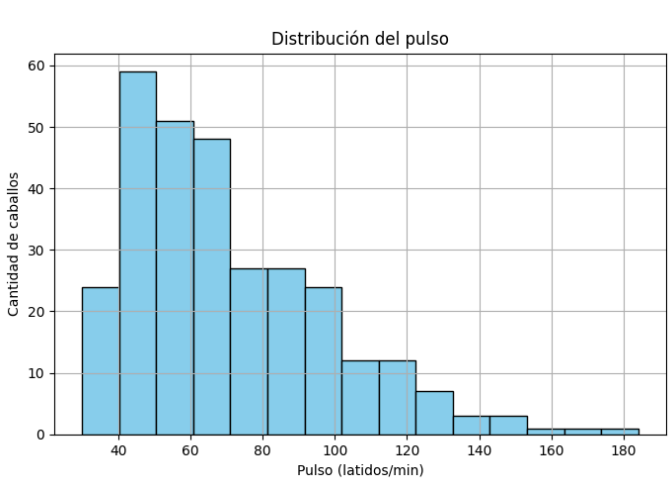
La variable pulso tiene 24 nulos / 8%

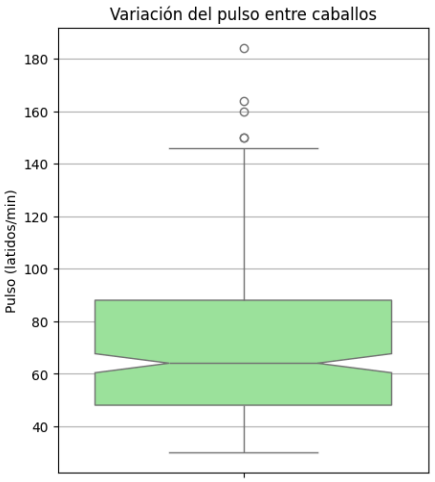
usamos el data general

Pasos:

- hipótesis nula (H0) como: "La media del pulso es igual a 44".
- hipótesis alternativa (H1) como: "La media del pulso es mayor a 44".
- Ejecutamos el T-test sobre la columna `pulse` (limpia) comparándola con el valor poblacional `mu=44`.

3.2.3 Resultados obtenidos y discusión





Resultados: Se calculó la media de la variable pulse (post-imputación con la mediana) obteniendo un valor de **71 lpm**. El Test T de una muestra, comparando esta media contra el valor 44, arrojó un p-valor extremadamente bajo: **p < 0.001**.

Discusión: Un p-valor tan pequeño (mucho menor que el nivel de significancia estándar de 0.05 o 0.01) nos permite rechazar la hipótesis nula. Indica que es estadísticamente improbable que nuestra muestra, con una media de ~71 lpm, provenga de una población con una media de 44 lpm. La diferencia observada es altamente significativa.

Por lo tanto, **se confirma la hipótesis.** La media del pulso en la muestra es significativamente superior al máximo normal en reposo.

3.3 Hipótesis 3 (Bivariada)

variables: numérica (cuanti) y categórica (cuali)

3.3.1 Definición de la hipótesis

"Los caballos que mueren o son sacrificados (outcome = 2 o 3) tienen, en promedio, un pulso (pulse) significativamente más alto que los caballos que sobreviven (outcome = 1)."

Justificación: Si un pulso elevado es un indicador de gravedad, debería estar directamente asociado con un peor desenlace (outcome). Se busca validar si pulse es un buen predictor de mortalidad.

Permite al veterinario categorizar el riesgo del paciente al momento de la admisión. Un pulso de 50 lpm tiene un pronóstico, pero uno de 100 lpm tiene otro completamente diferente. Esta relación justifica la monitorización constante y ayuda a gestionar las expectativas del propietario desde el minuto cero.

3.3.2 Estrategia de abordaje

Técnica: Visualización con Boxplots (de Pulse) y realizar un test Kruskal-Wallis si los datos no cumplen supuestos, para comparar las medias de pulse entre los 3 grupos.

Justificación: Los Boxplots son la herramienta visual ideal para comparar la distribución de una variable numérica (pulse) agrupados por las 3 categorías de outcome. El test Kruskal-Wallis es el test estadístico que nos dirá si las diferencias observadas entre las medias de los 3 grupos (vivió, murió, sacrificado) son estadísticamente significativas o si podrían deberse al azar.

Pasos de limpieza: dataframe utilizado (original o preprocesado)

La variable pulso tiene 24 nulos / 8%

La variable Outcome tiene 1 nulo / 0.33%

Usamos el data general

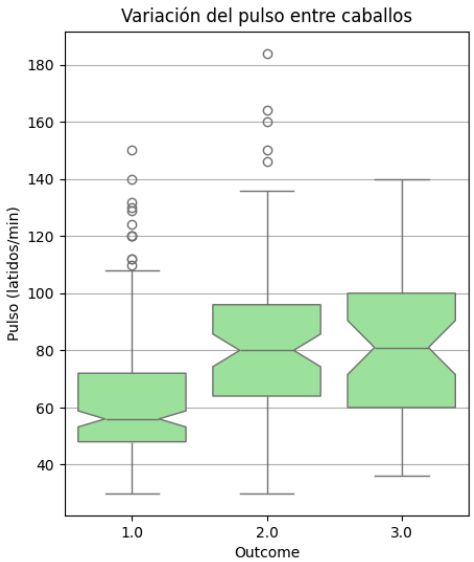
Pasos:

Visualización: Generar Boxplots de pulse agrupados por las 3 categorías de outcome.

Realizar test para validar los supuestos de normalidad y homogeneidad de varianzas

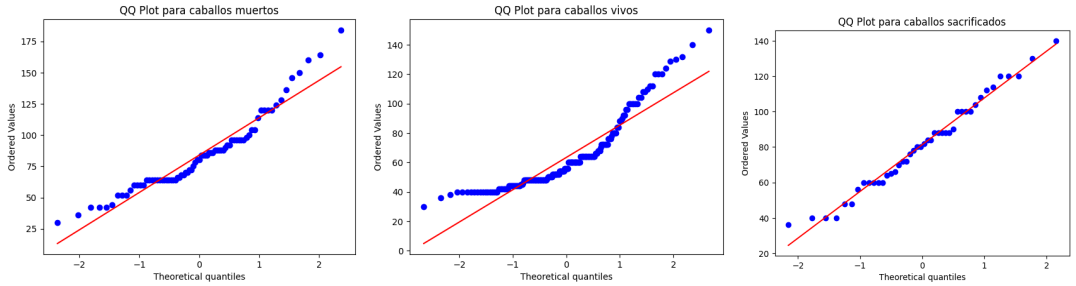
Test Estadístico: Realizar un test Kruskal-Wallis si los datos no cumplen supuestos de normalidad para comparar las medias de pulse entre los 3 grupos.

3.3.3 Resultados obtenidos y discusión



Resultados: Los boxplots mostraron una diferencia visual muy clara. La mediana del pulso para outcome=1 (Vivió) se situó alrededor de 56 lpm, mientras que las

medianas para outcome=2 (Murió) y outcome=3 (Sacrificado) fueron mucho más altas, superando ambas los 80 lpm.




```
Test de Shapiro-Wilk para caballos vivos: Estadístico=0.846, p-valor=0.000
Test de Shapiro-Wilk para caballos muertos: Estadístico=0.928, p-valor=0.000
Test de Shapiro-Wilk para caballos sacrificados: Estadístico=0.978, p-valor=0.5557470973161519
```

Luego comprobamos que no hay normalidad por test de Shapiro-wilk y gráfico QQ-plot.

Realizamos el test de levene y probamos que hay homocedasticidad, por lo tanto decidimos aplicar el test de Kruskal-Wallis.

```
Test de Levene para vivos-muertos: Estadístico=5.662, p-valor=0.018
Test de Levene para vivos-sacrificados: Estadístico=1.766, p-valor=0.185
```

Entonces después de comprobar los supuestos se aplicó Kruskal-Wallis para comparar las medias de los 3 grupos y arrojó un p-valor de **p < 0.001**, indicando que la diferencia entre el pulso de los grupos es estadísticamente significativa.

Discusión: Tanto la visualización como el test estadístico confirman la hipótesis. El pulso no solo es elevado en general, sino que es un fuerte indicador del resultado final. Los valores de pulso más altos están fuertemente asociados con la muerte o eutanasia del caballo.

Por lo tanto, **se confirma la hipótesis.** Un pulso elevado está fuertemente asociado con un pronóstico negativo.

3.4 Hipótesis 4 (Bivariada)

variables: categóricas (cuali)

3.4.1 Definición de la hipótesis

"Existe una asociación significativa entre el nivel de dolor (pain) y si el caballo presentaba una lesión quirúrgica (surgical_lesion = 1)."

Justificación: Los veterinarios suelen decidir la cirugía basándose en la gravedad del dolor. Un dolor "severo intermitente" o "severo continuo" (valores 4 o 5 en pain) sugiere una torsión u obstrucción (lesión quirúrgica) que el tratamiento médico no puede resolver. Esta hipótesis busca validar que la evaluación subjetiva del dolor por parte del veterinario es un indicador fiable para la decisión más importante: "abrir o no abrir".

3.4.2. Estrategia de abordaje

Técnica: Tabla de Contingencia (Crosstab) y Test Chi-Cuadrado (Chi-Square) de independencia.

Justificación: Ambas variables son categóricas. Una tabla de contingencia nos permite observar las frecuencias conjuntas (ejem. cuantos caballos con dolor nivel 5 tuvieron lesión quirúrgica). El test Chi-Cuadrado nos proporciona un p-valor para determinar si la asociación observada en la tabla es estadísticamente significativa o si es probable que ocurra por azar.

Pasos de limpieza: dataframe utilizado (original o preprocesado)

La variable Pain tiene 55 valores nulos / 18.33%

La variable surgical_lesion tiene 0 nulos / 0%

Usamos el data general

Pasos:

Crear una tabla de contingencia (crosstab) que cruce las categorías de pain (agrupadas si es necesario, ej. "dolor leve" vs "dolor severo") contra surgical_lesion (1: Sí, 2: No).

Aplicar el test Chi-Cuadrado a esta tabla.

3.4.3. Resultados obtenidos y discusión

/Surgical_lesion Pain/	1	2
1.0	6	32
2.0	38	21
3.0	78	44
4.0	32	7
5.0	37	5

Resultados: Se generó una tabla de contingencia entre los 5 niveles de pain y los 2 niveles de surgical_lesion. Se observó claramente que los dolores se relacionan a la

intervención quirúrgica, es decir que, mayoritariamente, a mayor dolor se eleva la probabilidad de que el paciente requiera una intervención.

El test Chi-Cuadrado aplicado a esta tabla arrojó un p-valor de **p < 0.001**.

Discusión: El p-valor confirma que las variables pain y surgical_lesion no son independientes. Existe una fuerte asociación estadística: en general a medida que aumenta el nivel de dolor reportado, aumenta significativamente la probabilidad de que el caballo presente una lesión que requiera cirugía.

Por lo tanto, **se confirma la hipótesis.** Existe una asociación estadística fuerte entre el nivel de dolor reportado y la determinación de que el caballo tiene una lesión quirúrgica.

3.5 Hipótesis 5 (Bivariada)

variables: numérica (cuanti) y categórica (cuali)

3.5.1 Definición de la hipótesis

"La distribución del Packed_cell_volume **no es igual** entre los distintos grupos de Mucous_membr; es decir, existe una relación significativa entre PCV y las membranas mucosas."

Justificación: Ambos atributos son descritos en la documentación como indicadores del estado circulatorio. Se espera que, a medida que la circulación del caballo empeora, la sangre se concentra, y ambos valores suben de forma conjunta.

3.5.2. Estrategia de abordaje

Técnica: Gráfico de Dispersión (Scatter plot) y cálculo del test no paramétrico utilizando Kruskal-Wallis

Justificación: El gráfico de dispersión permite visualizar de forma directa cómo varía el volumen de células empaquetadas (PCV) en función de cada categoría de Mucous_membr, facilitando detectar patrones, tendencias o posibles diferencias entre grupos. Dado que Mucous_membr es una variable categórica ordinal y el PCV no presenta una distribución normal en varios grupos, corresponde utilizar una prueba no paramétrica, como el test de Kruskal-Wallis, que compara si las distribuciones del PCV difieren significativamente entre las categorías sin asumir normalidad ni igualdad de varianzas.

Pasos de limpieza: dataframe utilizado (original o preprocesado)

Originalmente las variables estudiadas tenían estos valores nulos:

La variable Packed_cell_volume tiene 29 nulos / 9.66%

La variable Mucous_membr tiene 47 nulos / 15.66%

Usamos el data general que preprocesa estas variables

Pasos:

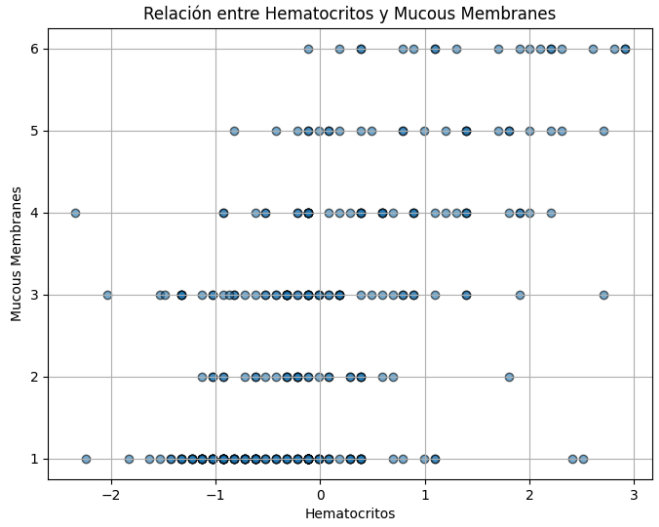
Para visualizar: generar un gráfico de dispersión (scatter plot) con packed_cell_volume en el eje X y Mucous_membranes en el eje Y.

Realizar test para validar los supuestos de normalidad y homogeneidad de varianzas.

Luego definimos las hipótesis:

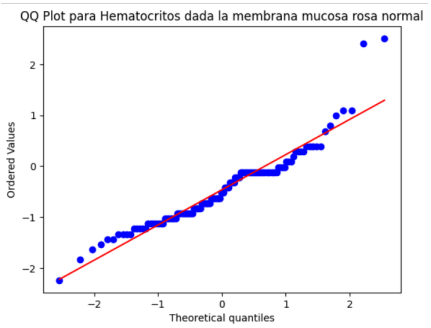
- hipótesis nula (H0) como: "packed_cell_volume es independiente de las categorías de Mucous_membr; es decir, la distribución de PCV es igual en todos los grupos de membranas mucosas".
- hipótesis alternativa (H1) como: "Packed_cell_volume varía dependiendo de la categoría de Mucous_membr".
- Ejecutamos el test Kruskal-Wallis sobre las variables comprometidas

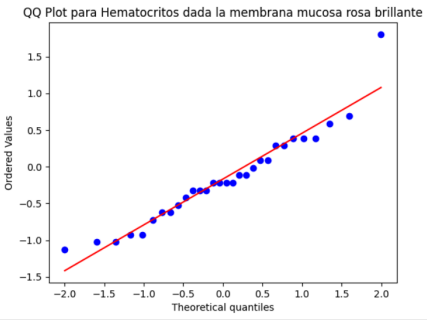
3.5.3 Resultados obtenidos y discusión

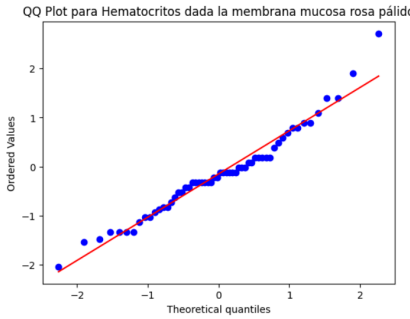


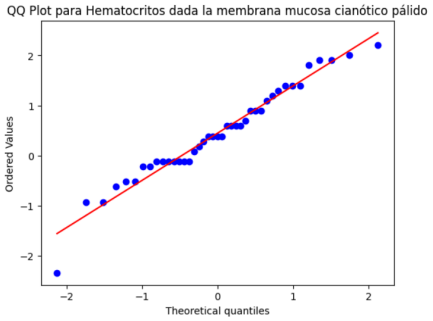
En el scatter plot se observa que, a medida que las membranas mucosas muestran colores asociados a un peor estado circulatorio (pasando de rosa normal a tonalidades pálidas o cianóticas), los valores de hematocrito (packed cell volume) tienden a aumentar. Esto indica que, a medida que se agrava la oxigenación del animal, también se eleva la concentración de glóbulos rojos en sangre, sugiriendo una relación entre ambas variables.

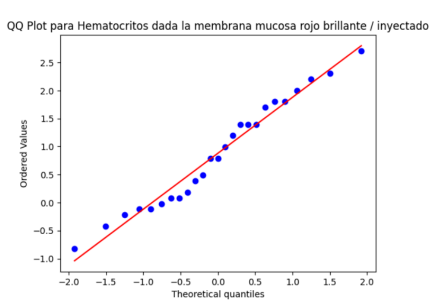
Ahora probamos los supuestos de normalidad (utilizando Shapiro-Wilk y QQ-plot) para la distribución de la variable “packed_cell_volume” que subdividimos en grupos en base a su categoria acorde a al nivel de color en las membranas mucosas.

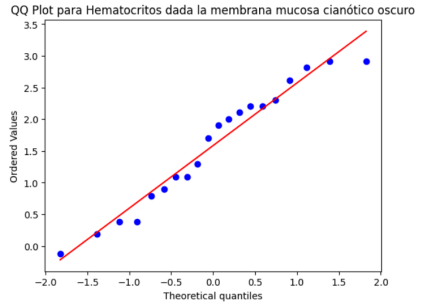












resultados test de Shapiro-Wilk

```
Categoría: rosa normal  ->  p-valor: 3.870724658908024e-06
Categoría: rosa brillante  ->  p-valor: 0.08282911040270238
Categoría: rosa pálido  ->  p-valor: 0.11716492003834428
Categoría: cianótico pálido  ->  p-valor: 0.29918082539335433
Categoría: rojo brillante / inyectado  ->  p-valor: 0.5067065469213864
Categoría: cianótico oscuro  ->  p-valor: 0.2755798738474963
```

Como se vé claramente que la categoría rosa normal no cumple con el p-valor asociado a la región de aceptación (su valor es < 0,01), se rechaza la normalidad de la variable a pesar de que el resto de variables si cumplen con el supuesto de normalidad.

Procedemos a evaluar la homocedasticidad mediante el test de levene.

```
Estadístico de Levene: 2.920169254344071
p-valor: 0.013687645216832259
```

De esta forma se confirma el supuesto de homocedasticidad teniendo en cuenta un valor $\alpha = 0,01$.

Si bien no es necesario que cumpla la homocedasticidad, se requiere que las distribuciones tengan formas similares, es decir que las varianzas no sean muy distintas.
Por lo tanto procedemos a utilizar un test no paramétrico y evaluar mediante test Kruskal-Wallis.

Resultados: Para evaluar la relación entre Packed_cell_volume y Mucous_membr se aplicó el test de Kruskal-Wallis.
Los resultados obtenidos fueron: **p < 0,0001.**

Por lo que, el p-valor resulta inferior al nivel de significancia habitual, entonces cae en la región de rechazo de la H0

Discusión: En términos clínicos, esto sugiere que el nivel de hematocritos (PCV) y la mucosidad en las membranas (Mucous_membr) **muestran una tendencia conjunta consistente** en los datos analizados. Es decir, un aumento o disminución en una de las variables predice y acompaña cambios en la otra. Se rechaza la hipotesis H0 que definimos y se acepta la hipótesis alternativa.

Por lo tanto, **se acepta la hipótesis.** Al haber encontrado evidencia estadística que respalde la no independencia entre el volumen de células empaquetadas y la mucosidad en las membranas, validando que ambos poseen una relación fuerte y positiva. Es decir, existe una relación significativa entre PCV y las membranas mucosas.

3.6 Hipótesis 6 (Multivariada)

variables: numéricas (cuanti) y categóricas (cuali)

3.6.1 Definición de la hipótesis

"El resultado final del caballo (outcome) puede ser predicho con una precisión superior al azar utilizando una combinación de variables de diagnóstico temprano: pulse, pain, rectal_temperature, peristal y packed_cell_volume."

Justificación: La realidad nunca depende de una sola variable. Esta hipótesis se basa en la idea de que un pronóstico clínico no depende de un solo factor, sino de la interacción de múltiples sistemas (signos vitales, dolor, hidratación). Si estos factores juntos pueden predecir la supervivencia, podemos crear un "índice de severidad" o algo similar. Esto es demasiado importante para la comunicación con el propietario, ayudando a tomar decisiones difíciles (como la eutanasia/mandarlo a dormir) basadas en datos objetivos y no solo en la intuición.

3.6.2. Estrategia de abordaje

Técnica: Modelo de Clasificación (Regresión Logística).

Métricas: Precisión (Accuracy) y Matriz de confusión.

Justificación: El problema consiste en predecir una variable categórica con 2 clases (outcome), ya preprocesada. La Regresión Logística es un modelo robusto e interpretable para este fin. Tambien es importante usar una matriz de confusión para entender como falla o acierta el modelo (ejem. ¿confunde "murió" con "vivió"?)

Línea base (baseline): Para probar que el modelo es "superior al azar", su exactitud debe ser mayor que la de un modelo básico que siempre predice la clase más frecuente. Como vimos en H1, la clase más frecuente es "vivió" (59.7%). Por lo tanto, nuestra **Línea base de precisión es 59.7%**.

Pasos:

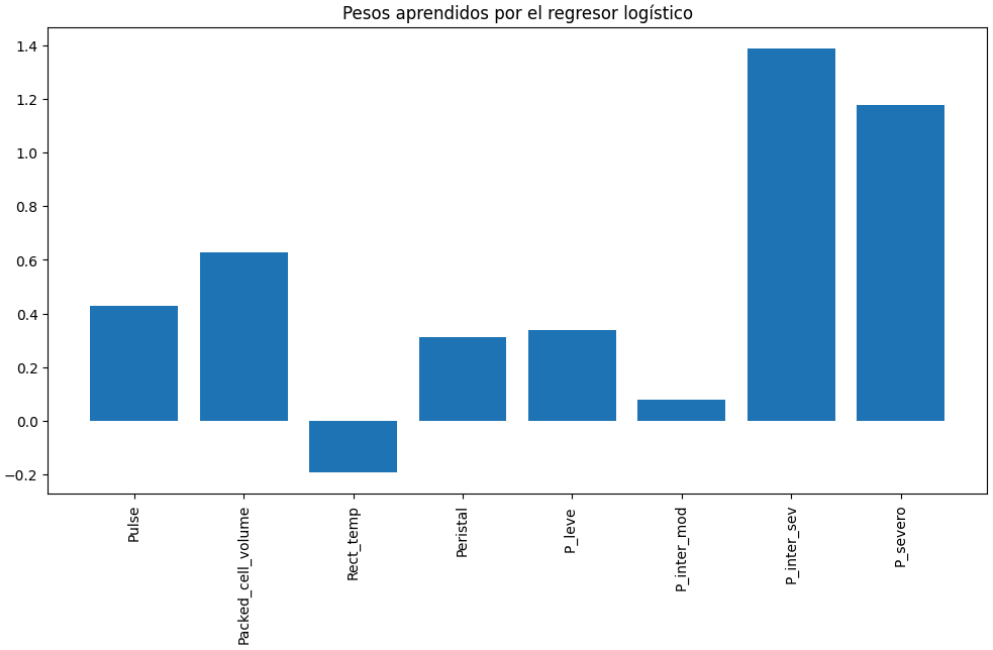
1. **Preparación:** Aplicar dummy Encoding a la variable categórica pain para que el modelo la interprete correctamente. Esto permite al modelo tratar cada nivel de dolor como una característica independiente sin asumir una relación numérica falsa entre ellas. Además agrupamos las categorías de Outcome para determinar si vivió o murió.

   Definir X (variables predictoras): pulse, pain (dummies), rectal_temperature, packed_cell_volume y peristal.

   Definir Y (variable objetivo): outcome.

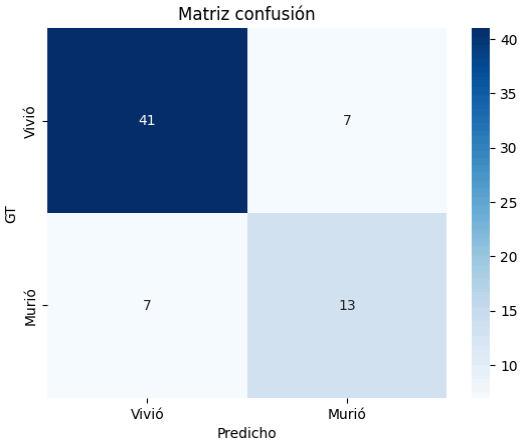
2. **División:** Separar en datos de entrenamiento (horse_dataset) y prueba (horse_datatest).
3. **Entrenamiento:** Ajustar el modelo LogisticRegression() usando los datos de entrenamiento.
4. **Evaluación:** Medir la precisión (Accuracy) del modelo sobre el dataset del test de prueba (datos que el modelo nunca ha visto).

3.6.3 Resultados obtenidos y discusión



Resultados: El modelo de Regresión Logística, entrenado con las 5 variables seleccionadas, alcanzó una **precisión (Accuracy) del 79.41%** en el conjunto de prueba. Con una **tasa de error** del **20.58%** determinando estos valores el rendimiento del modelo.

La matriz de confusión mostró que el modelo fue particularmente efectivo prediciendo las clases "vivió" (clase 1) y "murió" (clase 2).



Discusión: El resultado del modelo (79.41%) es **19.7 puntos porcentuales superior** a la línea base del 59.7%. Esto demuestra clara y cuantitativamente que la combinación de estas 5 variables tiene un poder predictivo real y útil.

La dificultad para predecir si el caballo es "sacrificado" es esperable y lógica desde el punto de vista del dominio: la eutanasia no es un evento biológico puro como "morir", sino una decisión humana que puede involucrar factores no clínicos (como el costo del tratamiento) que no están en nuestras variables. Por lo que la decisión de

agrupar Muerto con Eutanasia fue la mas adecuada para predecir si el caballo sobrevive o no.

Por lo tanto, **se confirma la hipótesis.** Las variables seleccionadas, cuando se usan en conjunto en un modelo multivariado, pueden predecir el resultado del caballo con una precisión significativa.

4. Conclusiones

Este Trabajo Práctico Especial ha permitido aplicar el proceso completo de la ciencia de datos a un problema de dominio real y complejo: el pronóstico del cólico equino. El análisis realizado sobre el **Horse Colic Dataset** permite extraer varias conclusiones relevantes tanto desde el punto de vista clínico como metodológico. A través de un proceso exhaustivo de limpieza, exploración y validación estadística, se pudieron identificar patrones que aportan valor para la predicción del desenlace clínico de caballos con cólico y para la comprensión del fenómeno.

El núcleo del trabajo, presentado en la Sección 3, fue la validación de seis hipótesis que nos permitieron construir una narrativa coherente a partir de los datos:

- La **hipótesis 1 resultó confirmada**. Se concluye que la mediana observada es significativamente menor que 3, indicando que la condición general de las mucosas en la muestra es mejor que el nivel considerado crítico.
- A pesar de la alta tasa de supervivencia, **se confirmó la Hipótesis 2**: la muestra estaba, en efecto, en un estado de estrés clínico significativo, con un pulso promedio (aprox 71 lpm) muy superior al normal.
- Las **Hipótesis 3 y 4 fueron confirmadas**, validando estadísticamente las intuiciones del dominio veterinario. Demostramos que un pulse elevado está fuertemente asociado con la muerte/eutanasia (H3) y ademas que el nivel de pain es un indicador fiable de la necesidad de una surgical_lesion (H4).
- En cuanto a la **Hipótesis 5 confirmada,** permitió rechazar la independencia entre ambas variables y concluir que el PCV varía según el estado de las membranas mucosas, evidenciando una relación consistente entre ambos indicadores clínicos**.**
- Finalmente, la **Hipótesis 6 fue confirmada** con éxito. Nuestro modelo de Regresión Logística, utilizando solo cinco variables de diagnóstico temprano, alcanzó una precisión del **79.41%**. Este resultado es notable no solo por su poder predictivo, sino porque **superó en más de 19.7 puntos porcentuales a la línea base** (59.7%), demostrando que la combinación de variables es significativamente más poderosa que adivinar la clase más común.

Conclusión General del Proyecto

En resumen, este Trabajo Práctico Especial nos permitió desarrollar de principio a fin un proceso completo de análisis estadístico y modelado predictivo aplicado al **Horse Colic Dataset**, con el objetivo de evaluar la utilidad clínica y diagnóstica de múltiples variables fisiológicas en el pronóstico del cólico equino. A lo largo del proyecto se implementaron técnicas de limpieza, exploración, análisis univariado, bivariado y multivariado, complementadas con métodos estadísticos adecuados al tipo y distribución de los datos.

Una de las conclusiones más relevantes surgió del análisis bivariado entre packed_cell_volume y Mucous_membr: se encontró una relación significativa y clínicamente interpretable entre ambas variables, lo que refuerza la idea de que, cuando la circulación del caballo empieza a fallar, también cambian los valores de hematocrito en la sangre.

Finalmente, el estudio culminó con la validación de la hipótesis multivariada mediante un **modelo de Regresión Logística** que alcanzó un desempeño altamente satisfactorio (accuracy del 79.41%), superando ampliamente la línea base establecida. Este resultado demuestra que, aunque cada variable aporta información parcial, la combinación de múltiples signos clínicos tempranos constituye un predictor robusto y útil del desenlace del caballo. La capacidad del modelo para mejorar la predicción respecto de un clasificador trivial enfatiza la importancia del enfoque multivariado en contextos clínicos reales.

En conjunto, el proyecto evidencia que el uso adecuado de métodos estadísticos y modelos predictivos aporta una comprensión más profunda del cuadro clínico, permite validar o refutar intuiciones del dominio, y abre la puerta como herramientas de apoyo a la decisión veterinaria. La integración de análisis exploratorios, contrastes estadísticos y modelado supervisado ofrece un marco sólido y reproducible para comprender y predecir la evolución del cólico equino a partir de datos objetivos.