

Disclaimer: Notebooks translated to Spanish are distributed as an optional aid to assist in your learning and comprehension. We make no guarantees that the translations are completely accurate nor that the translated code blocks will run properly.

¿Deberíamos desarrollar una prueba SNAP comercial para predecir la recuperación de las lesiones en la médula espinal?

Objetivo (3 min)

En este caso, usted aprenderá el propósito y las limitaciones de **visualización de datos** (**DV** por sus siglas en inglés). Es crucial que un científico de datos pueda implementar una visualización de datos adecuada para descubrir ideas, desarrollar soluciones y presentarlas a otras partes interesadas. Esperamos que al final del caso, usted pueda abordar un problema de data science y priorizar diferentes tipos de visualización de datos para ayudarlo a investigar y resolver dicho problema.

Antes de presentar el caso, comencemos con el siguiente ejercicio que muestra el poder de DV:

Ejercicio 1: (5 min)

Uno de los primeros y más impactantes ejemplos de DV es un mapeo del brote mortal de cólera de 1854 en Londres. La siguiente figura muestra las calles de Soho en Londres superpuestas con casos de cólera. La versión original de esta figura fue construida por el clínico local John Snow que trata a las víctimas del brote. Snow creó la figura para comprender mejor la naturaleza del brote e identificar su origen. El tamaño del círculo indica el número de casos en una ubicación determinada. A partir de esta imagen, identifique la fuente del brote.



Respuesta.

Este mapeo simple de la incidencia de la enfermedad ilustra un punto clave sobre DV: que DV debe ser informado por la experiencia en el dominio específico. Snow no tenía entrenamiento formal en estadística o epidemiología (¡no existía tal cosa!); todo lo que hizo fue construir un mapa para comprender mejor el brote. La elección de la visualización fue impulsada por la experiencia en el dominio específico. Como esencialmente todas las tareas de ciencia de datos, DV es menos significativo cuando no hay experiencia en el dominio para informar e interpretar los resultados.

Introducción (5 min)

Contexto de Negocio. Los modelos animales se usan comúnmente para estudiar las lesiones de la médula espinal en humanos. Por lo tanto, los avances tecnológicos realizados en medicina veterinaria para la lesión de la médula espinal (y otras áreas que comúnmente se basan en modelos animales) a menudo se hacen con el objetivo de desarrollar un producto para humanos. Uno de esos avances es el desarrollo de una prueba SNAP (una prueba que se puede ejecutar en unos minutos a partir de una sola extracción de sangre) que se puede utilizar para predecir si el paciente probablemente se recuperará de la lesión. Esta información es valiosa para médicos y profesionales clínicos. Si la información recopilada en la prueba SNAP demuestra tener suficiente poder predictivo, los investigadores buscarán una patente y llevarán el producto al mercado. La alternativa principal a una prueba SNAP es una prueba de laboratorio tradicional que requiere más tiempo y recursos; El tiempo requerido para obtener resultados de una prueba de laboratorio tradicional retrasa las decisiones de tratamiento que pueden afectar negativamente los resultados del paciente.

Problema de Negocio. Usted es un consultor de una empresa farmacéutica. Quisieran que responda la siguiente pregunta: “**¿Qué tan bien predicen las pruebas SNAP las tasas de recuperación de seis meses y deberían desarrollarse comercialmente?**”

Contexto Analítico. Interpretar gráficos y figuras correctamente y pensar críticamente sobre sus implicaciones es una habilidad crucial para un científico de datos en ejercicio. En este caso, resolveremos el problema anterior presentando una serie de tablas y gráficos, sacaremos conclusiones de ellas y tomaremos decisiones sobre qué hacer a continuación en función de eso.

El caso se estructura de la siguiente manera: primero (1) explorará estadísticas resumidas de cantidades clave; (2) verá varias formas estándar de trazar datos; (3) tomará una serie de decisiones en cada paso en función de estas gráficas; y finalmente (4) llegará a una conclusión sobre el poder predictivo de las pruebas SNAP y hacer una recomendación comercial.

Escenario para visualización de datos (5 min)

Antes de comenzar a usar la visualización de datos en este caso, hablemos sobre el marco o escenario adecuado para usar la visualización de datos.

Cuando analizamos un problema complejo de data science, queremos desglosarlo en pequeñas preguntas que son más concretas o más fáciles de resolver. (Por supuesto, a veces debemos dividir las preguntas secundarias en preguntas aún más específicas, ya que los problemas de data science pueden ser bastante complejos).

A medida que exploramos los datos observando varias estadísticas resumidas y distribuciones de parámetros, también vemos qué preguntas estos resultados ayudan a responder. Las visualizaciones de datos deben elegirse para ayudarnos a responder estas preguntas secundarias. Nunca debemos usar la visualización de datos sólo por tener un componente visual en nuestro trabajo de data science.

Como un ejemplo sencillo, volvamos al ejercicio anterior sobre el brote de cólera. La pregunta general es cómo podemos detener o minimizar este brote de cólera. Una pregunta secundaria podría ser dónde está el origen del brote o cómo se está propagando. Un [mapa de calor](#) ilustra esto claramente, así que esta es nuestra elección de visualización de datos. Aunque el mapa de calor anterior ciertamente es estéticamente

agradable, tenga en cuenta que lo elegimos por razones lógicas puramente deductivas basadas en la pregunta que teníamos que responder.

Observando los datos (20 min)

Los investigadores están interesados en una prueba SNAP que registrará niveles de como máximo tres biomarcadores: GFAP, pNFH y S100B. (Para nuestros propósitos, los antecedentes sobre estos biomarcadores y los mecanismos a través de los cuales podrían afectar la recuperación del paciente no son importantes). Para investigar el valor potencial de una prueba SNAP, los investigadores trajeron datos de registros médicos de 31 pacientes caninos con lesiones de la médula espinal.

El resultado clínico de interés es si el paciente recuperó o no la función motora después de seis meses. Se extrajo sangre a cada paciente en el momento de la lesión, que se almacenó y luego se usó para extraer los niveles de los tres biomarcadores, GFAP, pNFH y S100B, utilizando una prueba de laboratorio estándar. Por lo tanto, los niveles de biomarcadores extraídos representan los niveles previos al tratamiento y no afectaron el curso de tratamiento de cada paciente.

Las primeras filas de los datos son las siguientes:

| Subject id | GFAP | S100B | pNFH | Recovered | Sex |
|------------|-------|-------|------|-----------|-----|
| 1 | 11.76 | 0.041 | 1.75 | N | F |
| 2 | 7.63 | 0.031 | 4.89 | N | F |
| 3 | 10.0 | 0.028 | 2.59 | N | F |
| 4 | 0.01 | 0.038 | 3.46 | Y | F |
| 5 | 0.0 | 0.0 | 0.60 | Y | M |
| ... | ... | ... | ... | ... | ... |

Ejercicio 2: (5 min)

¿Consideraría usted desarrollar un modelo en este momento? ¿Por qué sí o por qué no?

Respuesta.

Este es un ejemplo de un problema de clasificación — un problema en el que cada punto de datos está etiquetado con una de varias **clases** y queremos construir un modelo para predecir a qué clase debe pertenecer un nuevo punto de datos sin etiquetar. En este caso, las dos clases corresponden al estado de recuperación (Sí o No - Y o N), y queremos predecir si un paciente se recuperará o no después de seis meses debido a sus atributos GFAP, S100B, pNFH y Sexo.

Comencemos calculando algunas estadísticas de resumen sencillas para cada una de las variables observadas (como hemos visto en los casos de Python, este es un primer paso bastante lógico). Para cada variable continua, calculamos algunas estadísticas de resumen simples (mín., máx., mediana, percentil 25, percentil 75). Los resultados se muestran en la siguiente tabla.

| Variable | Min | 25th Percentile | Median | 75th Percentile | Max |
|----------|-------|-----------------|--------|-----------------|------|
| GFAP | 0.0 | 0.0 | 0.0 | 0.38 | 37.8 |
| S100B | 0.0 | 0.014 | 1.24 | 2.0 | 3.9 |
| pNFH | -0.30 | 0.23 | 0.44 | 0.94 | 65.0 |

Ejercicio 3: (3 min)

Los biomarcadores, GFAP, S100B y pNFH, miden los niveles de elementos biológicos en la sangre. ¿Notas algo inusual en las estadísticas de resumen de alguna de las variables?

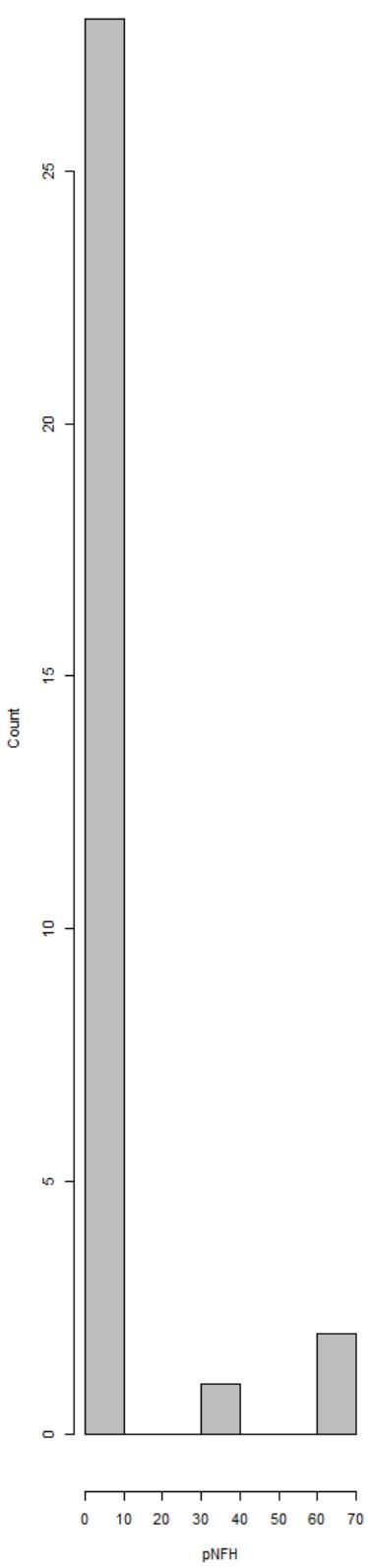
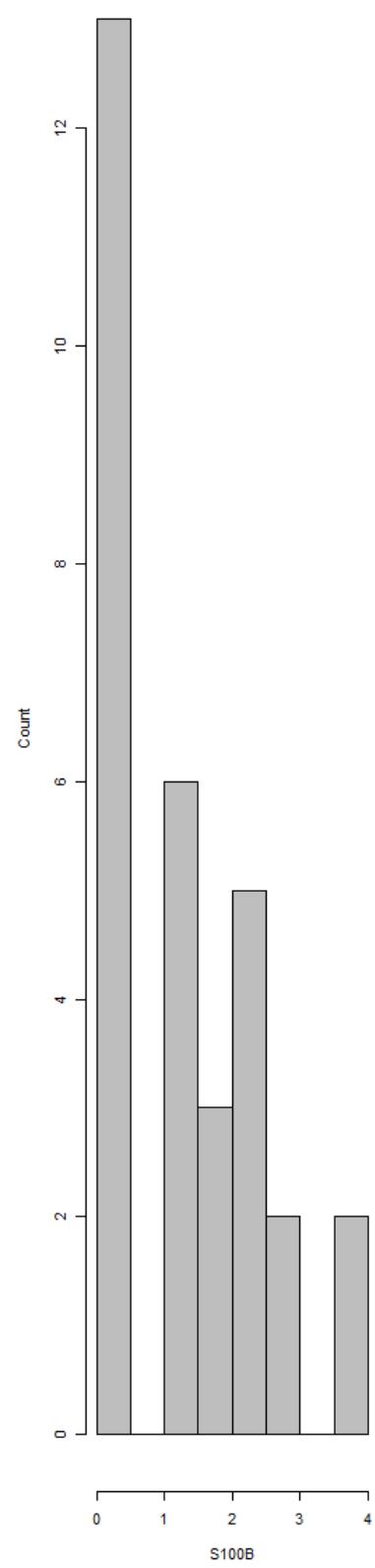
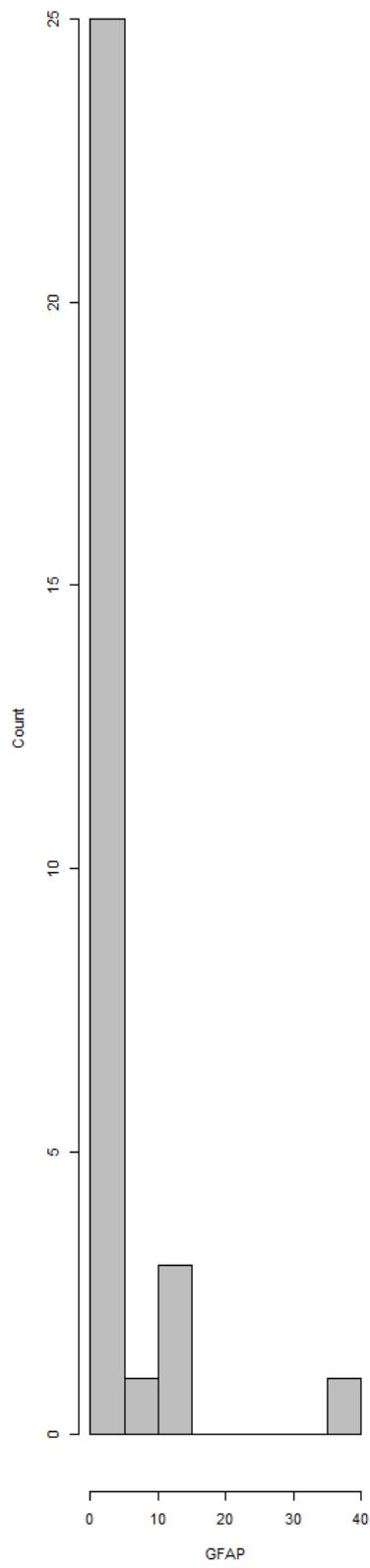
Respuesta.

Histogramas para biomarcadores

Los resúmenes numéricos son útiles para tener una idea general de la distribución de datos, a menudo es muy útil también obtener una representación gráfica de la distribución de datos para ayudar en la presentación y comprensión.

Un método visual sencillo y útil para ver la distribución a un nivel más granular es usar un **histograma**. En un **histograma**, el eje x se divide en diferentes **particiones** de valores para la variable de interés. Cada punto de datos se coloca en una partición según el valor de su variable de interés. El eje y corresponde al número de puntos de datos en cada partición en el eje x. Por lo tanto, los histogramas muestran qué valores de datos de una variable particular tienen una alta densidad de puntos. Las barras más grandes en un histograma indican una mayor frecuencia de observaciones en la partición correspondiente.

A continuación, hemos construido tres histogramas, uno para cada biomarcador:



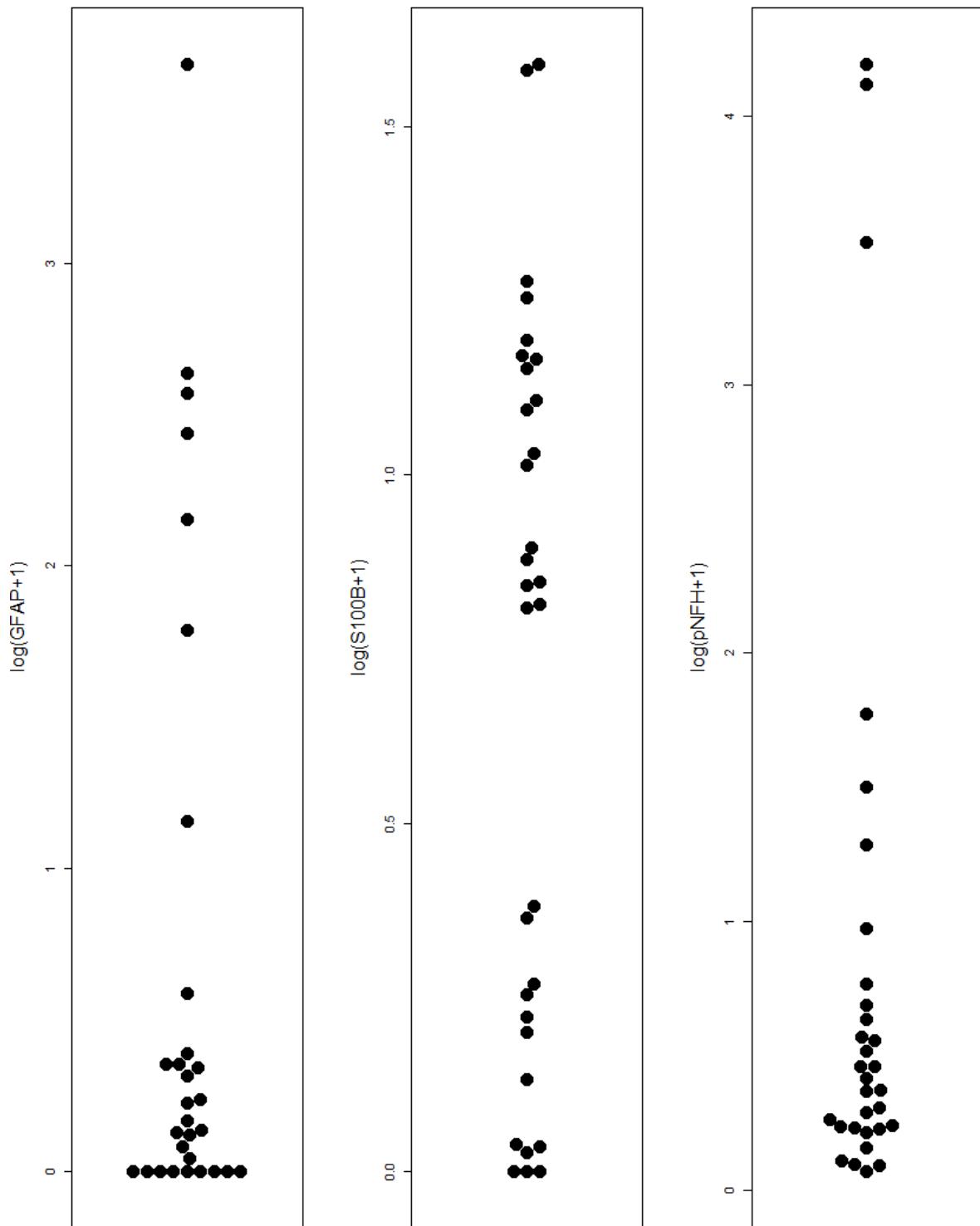
Pregunta 4: (4 min)

¿Los histogramas anteriores son informativos?

Respuesta.

Gráfico de dispersión 1D: Una alternativa al histograma (10 min)

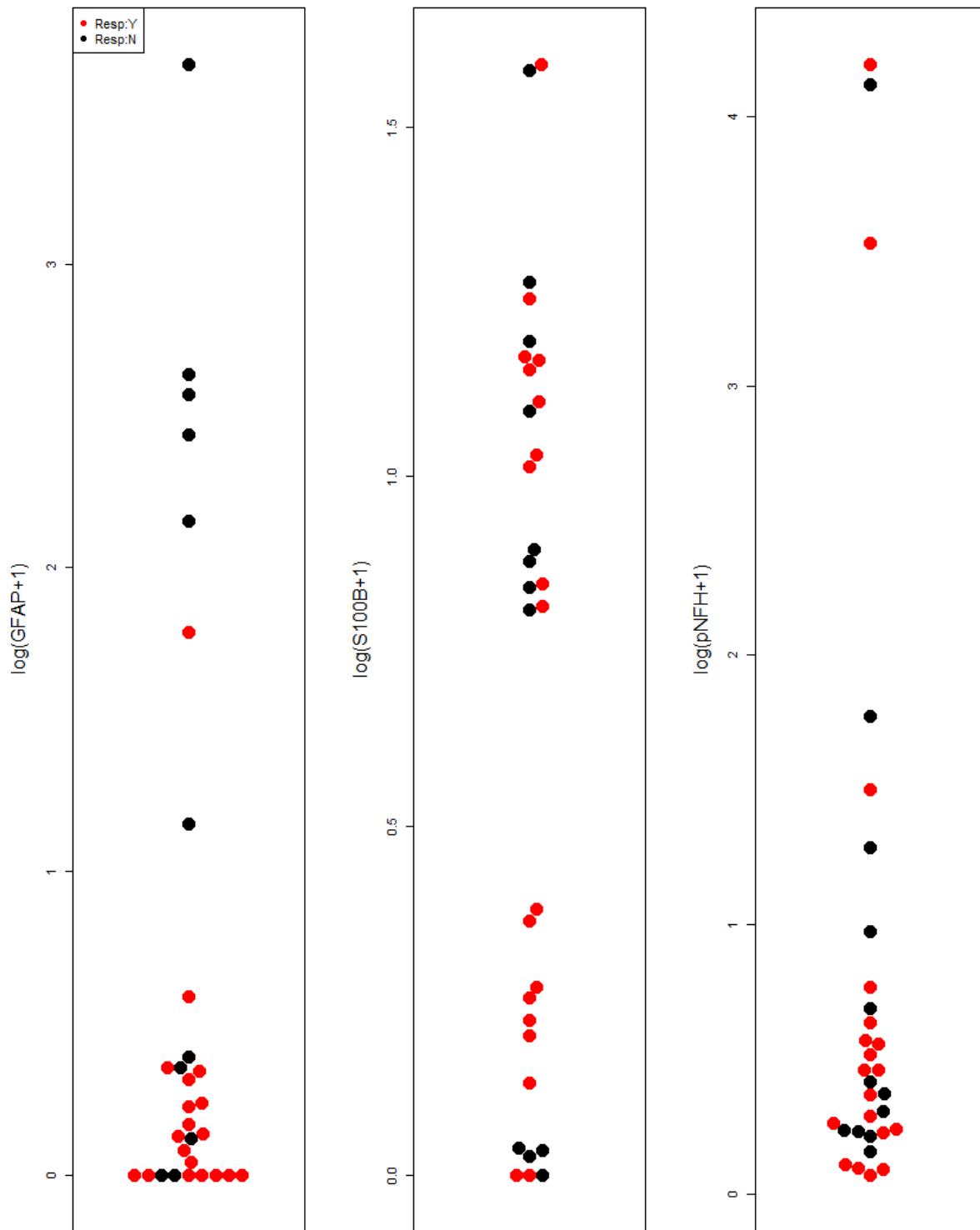
El [gráfico de dispersión 1D](#) es una alternativa al histograma, que a veces puede resultar más informativo cuando hay grupos de puntos. En un diagrama de dispersión 1D, el eje y representa la cantidad de interés (aquí, los valores de los biomarcadores), y los puntos observados se colocan de modo que su coordenada y sea igual a su valor de biomarcador y se agrupen lo más estrechamente posible a lo largo del eje x sin permitir que se superpongan. Por lo tanto, los grupos de puntos se desplazarán horizontalmente, lo que provocará que el gráfico se “hinche” en áreas donde hay muchos puntos con valores de biomarcadores similares. A continuación mostramos un diagrama de dispersión 1D para las transformaciones logarítmicas de cada uno de los biomarcadores:



(La transformación logarítmica es puramente para fines visuales; nos permite “dividir” los grupos de puntos cerca de cero y mostrarlos con mayor precisión. En casos posteriores, aprenderá sobre el poder analítico de usar la transformación logarítmica en sus datos.)

Usar el color como una herramienta para diferenciar categorías

La figura anterior da una mejor idea de la distribución de valores de cada biomarcador. Vemos que GFAP tiene un gran grupo en cero, mientras que pNFH tiene muchos puntos pequeños pero solo un cero. Sin embargo, nuestro objetivo es ver si estos biomarcadores pueden predecir el estado de recuperación. Para visualizar esto mejor, podemos colorear cada uno de los puntos según su estado de recuperación (Y si se recuperaron después de seis meses, N si no):



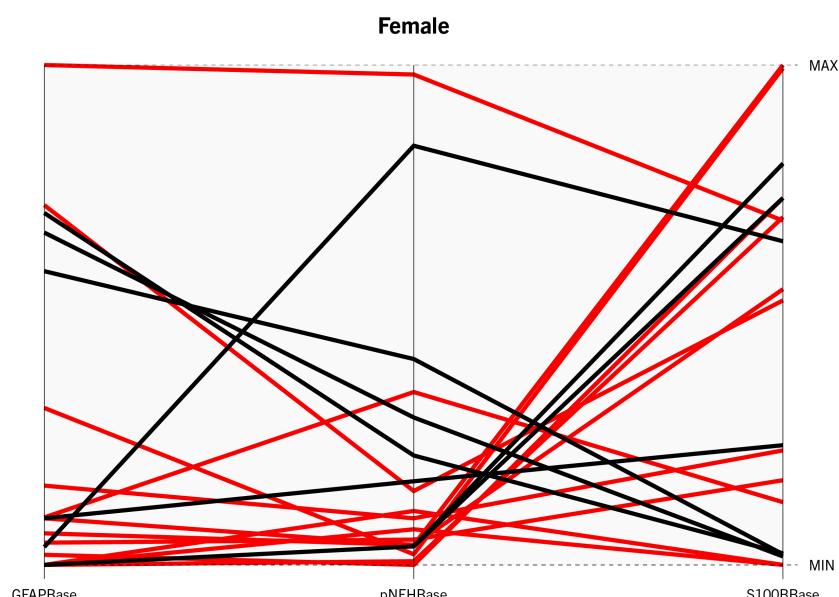
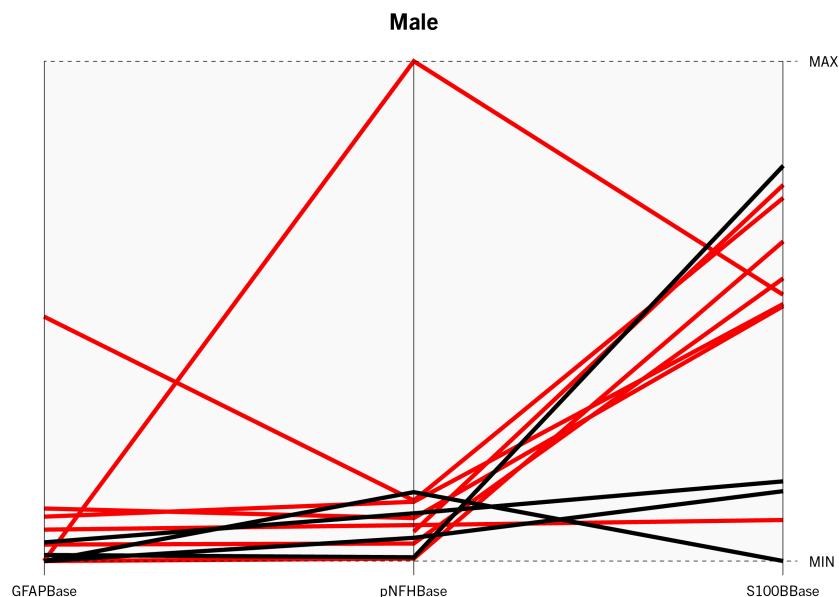
Ejercicio 5 (4 min):

De la gráfica anterior, ¿qué biomarcadores crees que afectan más la probabilidad de recuperación?

Respuesta.

Usando la última parte de la información: género (4 min)

Sin embargo, todavía no hemos utilizado una última parte de la información que tenemos disponible: el género. A continuación mostramos un gráfico de [coordenadas paralelas](#) dividido por sexo. Cada línea representa un solo sujeto y el eje y representa sus diferentes valores de biomarcadores. La gráfica sugiere que los hombres con S100B y GFAP altos tienen probabilidades de recuperarse (la recuperación se codifica en rojo), mientras que las mujeres con GFAP bajo tienen probabilidades de recuperarse. Si bien este gráfico no es un modelo y no hemos intentado calcular la tasa de error de clasificación, podemos ver que hay alguna señal. Tratemos de explorar esto más a fondo.



Gráficas de dispersión 2D (20 min)

El siguiente conjunto de gráficos que se muestran son gráficos de dispersión etiquetados por estado de recuperación. Sin embargo, en lugar de diagramas de dispersión 1D, ahora veremos [Gráficas de dispersión 2D](#). En las Gráficas de dispersión 2D se grafican datos a lo largo de dos ejes de acuerdo con los valores de los puntos de datos para los atributos representados por esos ejes; por lo tanto, dejan muy claro si hay alguna relación entre las variables en los ejes.

Para ayudar a identificar visualmente qué regiones parecen estar asociadas con la recuperación, colocamos una cuadrícula fina de puntos en el fondo y los coloreamos según si el punto más cercano correspondiente a un paciente observado se recuperó con éxito o no (es decir, para cada punto en la cuadrícula lo coloreamos de rojo si el paciente observado más cercano se recuperó y negro en caso contrario). El siguiente diagrama de dispersión de S100B frente a GFAP sugiere que S100B tiene poco valor adicional sobre GFAP en términos de precisión predictiva:

También observamos estos diagramas de dispersión separados por género:

Ejercicio 6 (3 min):

¿Qué conclusiones puedes sacar de las gráficas de dispersión 2D mostradas anteriormente?

Respuesta.

Aquí están las gráficas de dispersión 2D para GFAP vs pNFH y S100B vs pNFH:

Ejercicio 7: (12 min)

Utilice las visualizaciones anteriores para hacer una recomendación a los investigadores sobre si vale la pena explorar el desarrollo de una prueba SNAP para recuperación utilizando los tres biomarcadores. Trabaje con un compañero para esto. Sus respuestas también deben abordar las siguientes preguntas:

7.1

Si el costo de la prueba SNAP depende de la cantidad de biomarcadores que utiliza, ¿qué biomarcadores recomendaría incluir en la prueba?

7.2

¿Cuáles deberían ser los próximos pasos? Por ejemplo, ¿recomendaría realizar una prueba de seguimiento que confirme la precisión de la predicción de biomarcadores?

7.3

¿Cómo podría desarrollar un modelo de clasificación a este estudio? ¿Cuáles son las posibles dificultades asociadas con este estudio?

Respuesta.

Extender nuestra prueba SNAP (5 min)

Tras una consideración adicional de los protocolos de tratamiento estándar, los investigadores han decidido que las decisiones críticas de tratamiento deben tomarse dentro de las 72 horas posteriores a la lesión. En consecuencia, están interesados en si tomar o no tres mediciones, una por día, podría proporcionar mejores predicciones de recuperación de seis meses. Debido a que GFAP fue nuestro biomarcador más prometedor

al inicio del estudio, primero hagamos un gráfico de coordenadas paralelas de GFAP a intervalos regulares durante las primeras 72 horas (días 0, 1 y 2). Para facilitar el análisis visual de la trama, la dividimos por estado de recuperación:

El gráfico anterior sugiere que los niveles de GFAP tienden a ser más volátiles entre los pacientes que no se recuperaron frente a los que sí lo hicieron, y que sus picos son más altos. A continuación mostramos una gráfica de dispersión 1D del máximo (log) GFAP durante 72 horas coloreado por el estado de recuperación. Vemos que podemos separar casi perfectamente los datos en $c \approx 0.5$, excepto por tres puntos:

Conclusiones (5 min)

Evaluamos el valor potencial de mercado de una prueba SNAP para lesiones de la médula espinal en perros. Descubrimos que el biomarcador GFAP parece ser el más discriminatorio entre los tres biomarcadores considerados en términos de identificación de la probabilidad de recuperación. Resulta que si las mediciones se toman durante las primeras 72 horas después de una lesión, es posible construir predicciones significativamente mejores al tomar el máximo de mediciones GFAP durante ese período. Debido a que el tamaño de la muestra era pequeño, recomendamos que los investigadores consideren un estudio de seguimiento con un grupo mucho mayor de sujetos antes de comprometer los recursos de la compañía para desarrollar una prueba SNAP comercial.

Ideas que nos debemos llevar (5 min)

En este caso, usted ha aprendido sobre el poder de las visualizaciones de datos (DV) para obtener información sobre las problemáticas incluso antes de comenzar el modelado. También aprendió a interpretar correctamente, sacar conclusiones y descubrir los próximos pasos de las visualizaciones de datos. Algunos consejos útiles de acuerdo a lo visto incluyen:

1. Los histogramas, las gráficas de dispersión y los gráficos de coordenadas paralelas proporcionan información que no está contenida en las estadísticas de resumen tabular. En particular, estos permiten ampliar de manera más efectiva puntos de datos atípicos, así como partes específicas de la distribución general de datos.
2. El uso adecuado del color es una herramienta poderosa cuando se combina con tramas; en este caso, usamos color para separar sujetos de sexo masculino y femenino en las gráficas de dispersión. El color es una forma fácil y común de agregar otra dimensión a un gráfico 2-D sin cambiar la estructura fundamental del gráfico en sí.

La visualización de datos es una parte indispensable de la toma de decisiones basada en datos. En muchos casos, la visualización de datos es la parte más informativa y más lenta del proceso de análisis y data science. Una serie de visualizaciones de alta calidad puede proporcionar información crítica para construir modelos más adelante. Además, las visualizaciones de datos pueden facilitar las discusiones entre el científico de datos y sus colaboradores menos técnicos.