codificandobits

¿Cómo hacer el Análisis Exploratorio de Datos? - Guía paso a paso

Junio 11, 2021 por Miguel Sotaquirá

En este *post* les comparto una guía paso a paso sobre cómo hacer el análisis exploratorio de datos, una fase esencial en cualquier proyecto de Machine Learning o Ciencia de Datos.

Al final del artículo encontrarás las instrucciones para descargar esta guía en formato PDF.

¡Así que listo, comencemos!

Tabla de contenido

codificandobits

Video

Como siempre, en el canal de YouTube se encuentra el video de este *post*:

codificandobits

Machine Learning Engineering (o MLOps), y allí vimos todas las fases involucradas en el desarrollo de un proyecto de Machine Learning. En particular, la etapa de preparación de los datos, además de ser fundamental, requiere casi siempre entre un 60% y un 70% del tiempo de desarrollo.

Y una de esas tareas fundamentales en esta etapa es el análisis exploratorio de los datos, que es el tema de este artículo. Y aunque en Internet se encuentran muchos tutoriales y recursos sobre este tema, realmente no es fácil encontrar una guía que nos indique cómo hacer este análisis exploratorio, independientemente de las características particulares del proyecto que estemos desarrollando.

Así que en este artículo veremos precisamente una guía paso a paso sobre cómo hacer el análisis exploratorio de datos en Machine Learning o Ciencia de Datos.

Vamos a ver en qué consiste el análisis exploratorio, cuáles son los tipos de datos y las herramientas estadísticas para describirlos, hablaremos de las herramientas de visualización, del análisis bivariado y multivariado y de la sumarización. Recuerda que al final del artículo

codificandobits

¿Qué es y para qué sirve el Análisis Exploratorio de Datos?

Bien, primero tengamos en cuenta que todo lo que veremos ahora aplica únicamente para datos estructurados, es decir los que vienen en formato tabular. Para datos no estructurados o para series de tiempo el análisis exploratorio es totalmente diferente, y de esto hablaremos en artículos posteriores.

El principal propósito del análisis exploratorio es tener una idea completa de cómo son nuestros datos, antes de decidir qué técnica de Ciencia de Datos o de Machine Learning usaremos.

Y como en la práctica los datos no son ideales, debemos organizarlos, entender su contenido, entender cuáles son las variables más relevantes y cómo se relacionan unas con otras, comenzar a ver algunos patrones, determinar qué hacer con los datos faltantes y con los datos atípicos, y finalmente extraer conclusiones acerca de todo este análisis.

Y todo esto es precisamente el análisis exploratorio de datos, que es en resumen una forma de entender,

codificandobits

decidir cual sera la ruta o técnica mas adecuada para su posterior procesamiento.

Y este es siempre el paso cero en cualquier proyecto de Machine Learning o Ciencia de Datos. Siempre debemos comenzar por acá.

Bien, teniendo esto claro podemos resumir las fases del análisis exploratorio en 7 pasos:

1. Tener clara la pregunta que queremos responder;
2. Tener una idea general de nuestro dataset;
3. Definir los tipos de datos que tenemos;
4. Elegir el tipo de estadística descriptiva
5. Visualizar los datos;
6. Analizar las posibles interacciones entre las variables del dataset; y finalmente
7. Extraer algunas conclusiones de todo este análisis.

Para entender todas estas fases usaremos un dataset clásico de Kaggle: el del Titanic, un set de datos que contiene información de los pasajeros como nombres, edades, género y obviamente la categoría a la que pertenece, es decir si sobrevivió o no al hundimiento.

codificandobits

dando un vistazo a nuestro set de datos

El primer paso, la pregunta que queremos responder en este caso, es ¿que tipo de personas tenían la probabilidad más alta de sobrevivir al hundimiento del Titanic?

Y para responder esta pregunta debemos echar un primer vistazo al dataset, mirar su tamaño, determinar cuáles son las características o variables (es decir las columnas de la tabla) y dar un primer barrido a los registros u observaciones (es decir las filas del dataset):

Con esto nos haremos una idea general de los datos, viendo que por ejemplo cada pasajero estará caracterizado por variables como el nombre, la edad, el género, etc.

Paso 3: ¿qué tipos de datos tenemos?

Bien, después de esto podemos comenzar a analizar en detalle el dataset. El paso tres es entonces definir a qué tipo de dato pertenece cada variable. Y acá tenemos dos grandes grupos: los datos numéricos y los datos categóricos.

Los datos numéricos pueden ser discretos cuando toman sólo valores

pueden tomar cualquier valor dentro de un intervalo (como por ejemplo la tarifa del tiquete).

Los datos categóricos pueden ser nominales, binarios u ordinales. Los nominales se usan para etiquetar el dato pero no pueden ser ordenados ni medidos, como por ejemplo el género de los pasajeros: hombre o mujer.

Los datos binarios indican una de dos posibles categorías, como por ejemplo “sobreviviente” o “no sobreviviente”.

Y finalmente están los datos ordinales que, como su nombre lo indica, corresponden al orden en el que vienen representados los datos, como por ejemplo categoría del tiquete: 1, 2 o 3.

Paso 4: Descripción estadística de los datos

El cuarto paso es iniciar con la descripción estadística que depende precisamente del tipo de datos que tengamos en cada variable.

Para esto usamos dos grandes tipos de medidas: las de tendencia central y las de variabilidad.

Medidas de tendencia central

Las medidas de tendencia central nos dan una idea general del valor típico que pueden tener nuestros datos, y

La media es simplemente el promedio de los datos y por tanto se puede aplicar a datos discretos (como por ejemplo la edad de los pasajeros) o continuos (como por ejemplo el valor de los tiquetes).

La desventaja de la media es que es muy sensible a valores atípicos: si por ejemplo la mayor parte de los tiquetes está alrededor de precios bajos, pero unos pocos tienen valores muy altos, al calcular la media daría la impresión de que en promedio los pasajeros compraron tiquetes un poco más costosos.

La mediana resuelve este inconveniente, y es simplemente el valor que divide los datos en dos mitades y se puede aplicar para datos ordinales o discretos (como la categoría del tiquete o la edad).

Para calcularla debemos primero organizar los datos de manera ascendente y luego encontrar el valor tal que la mitad de los datos estarán por debajo de dicho valor y la otra mitad por encima:

$$1; 3; 3; 6; 7; 8; 9 \rightarrow mediana = 6$$
$$1; 2; 3; 4; 5; 6; 8; 9 \rightarrow mediana: (4 + 5)/2 = 4.5$$

Medidas de variabilidad

Pero resulta que no es suficiente con conocer la media o la mediana de la

tan agrupados o dispersos están los datos.

Para determinar esto usamos las medidas de variabilidad, donde las principales son la desviación estándar y el rango intercuartiles, que nos indican qué tanto se alejan los datos del valor medio o de la mediana, respectivamente.

La desviación estándar se puede calcular para cualquier tipo de dato numérico: entre más bajo sea su valor tendremos datos más agrupados, y viceversa. La desventaja de la desviación estándar es la misma de la media: es muy sensible a los valores atípicos.

Una alternativa es usar el rango intercuartiles, que es la diferencia entre el percentil 75 y el percentil 25. Si la mediana es el punto medio de los valores observados, el percentil 75 es el valor por debajo del cual se encuentra el 75% de los valores, mientras que el percentil 25 corresponderá al 25% de dichos valores. Al igual que la mediana, esta diferencia intercuartiles también es menos sensible a valores atípicos en comparación con la desviación estándar.

Así, en nuestro dataset, el percentil 75 es 38 años y el 25 es 20 años, y por tanto el rango intercuartiles será de 18

datos.

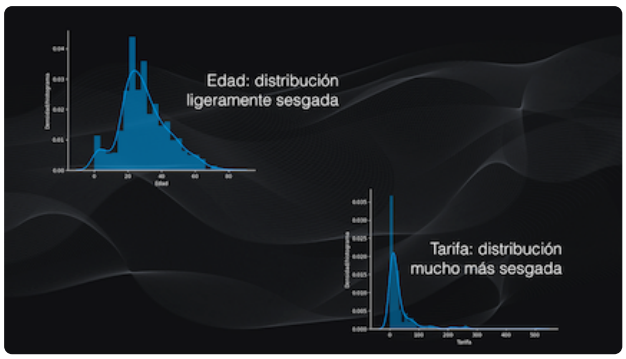
Los percentiles 25, 50 (es decir la mediana) y 75 dividen la distribución exactamente en cuatro partes llamadas cuartiles: el primer cuartil cubre del 0 al 25% de la distribución; el segundo del 25 al 50%; el tercero del 50 al 75% y el cuarto del 75 al 100%. Tengan en cuenta esta definición porque la usaremos más adelante.

Paso 5: visualizar los datos

La limitación de las medidas centrales y de las de variabilidad es que son sólo un número, que nos puede dar apenas una idea general del comportamiento de nuestros datos. Así que el quinto paso del análisis exploratorio es visualizar la distribución de los datos para tener una idea más detallada de su comportamiento.

Para datos continuos y discretos podemos calcular y dibujar el histograma, que se obtiene tras organizar los datos en diferentes subgrupos (o bins) y realizar el conteo del número de datos en cada uno. Con el histograma podemos verificar que la distribución es normal (es decir que tiene forma como de campana, como por ejemplo la edad) o si está

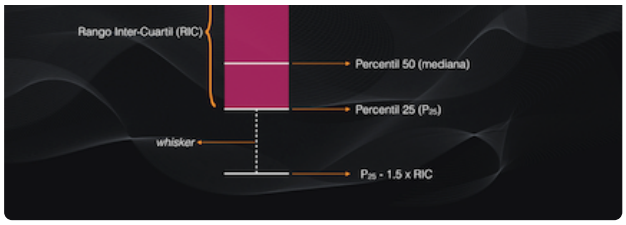
tarifa).



Histogramas para las variables Edad y Tarifa

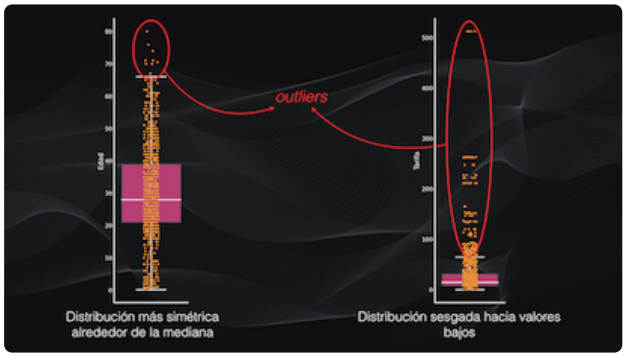
La desventaja del histograma es que no permite ver en detalle los valores atípicos, porque quedarán enmascarados al incluirlos en un bin. La alternativa en este caso, o cuando la distribución es sesgada, es usar los diagramas de caja o *boxplots*, que también se pueden usar para datos discretos y continuos.

En un *boxplot* se dibujan los percentiles: la barras superior e inferior corresponden a los percentiles 75 y 25, mientras que la línea en medio de la caja es la mediana. Por fuera de la caja hay dos líneas, conectadas por líneas punteadas (que se llaman *whiskers* o bigotes en Español). Cada una de ellas es igual al percentil 75 o 25 más o menos 1.5 veces el rango intercuartil:



Los elementos de un boxplot

Si por ejemplo dibujamos el boxplot para la edad y la tarifa y superponemos los datos originales podemos fácilmente interpretar estas variables: vemos que el rango de edades es uniforme entre 0 y 68 años, mientras que la mayor parte de los pasajeros tenía tiquetes económicos, entre 0 y 30. También podemos ver los outliers, que están más allá de las líneas de los extremos:



Boxplots para las variables Edad y Tarifa, con los datos correspondientes superpuestos (puntos de color naranja)

El tratamiento de los *outliers* lo veremos en detalle en un próximo artículo, pero por ahora tengamos en cuenta que es una parte importante del Análisis Exploratorio de los Datos.

Y bien, ¿pero qué pasa si tenemos datos categóricos? En este caso podemos usar los gráficos de barras,

categoría o el porcentaje que estas representan del total de datos.

Por ejemplo, el gráfico de barras para la variable “supervivencia” nos muestra que fueron más los “no sobrevivientes” que los “sobrevivientes”. Esto nos da una pista del posible esquema a utilizar en la predicción: el set de datos está desbalanceado y probablemente si usamos un clasificador convencional tendremos problemas para entrenarlo:

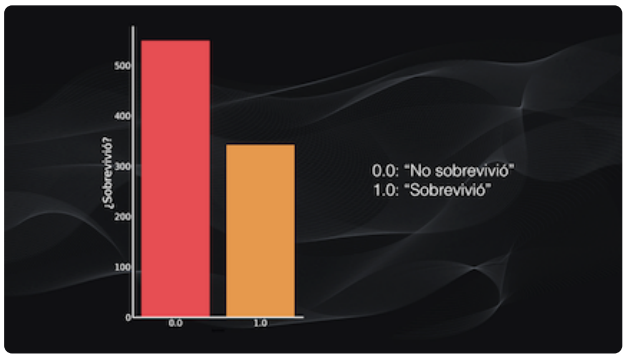


Gráfico de barras para la variable Supervivencia

Paso 6: análisis bivariado y multivariado

Hasta el momento hemos visto el análisis y visualización de una sola variable, lo que se conoce precisamente como análisis univariado. Pero también podemos ver si existe algún tipo de interacción entre dos o más variables, usando lo que se conoce como el análisis bivariado y el multivariado.

Análisis bivariado

podemos aprovechar los tipos de gráficas que vimos anteriormente para analizar estas interacciones.

Por ejemplo, si queremos comparar dos variables numéricas (como la tarifa y la edad del pasajero) podemos usar una gráfica de dispersión, donde cada punto es representado por un dato, y podemos verificar si existe alguna tendencia lineal: es decir si el aumento de una variable genera un aumento o disminución de la otra:

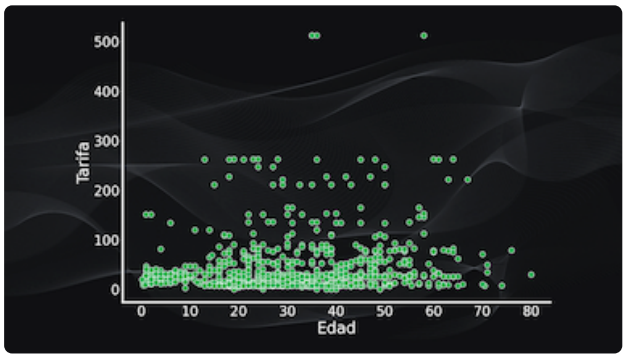
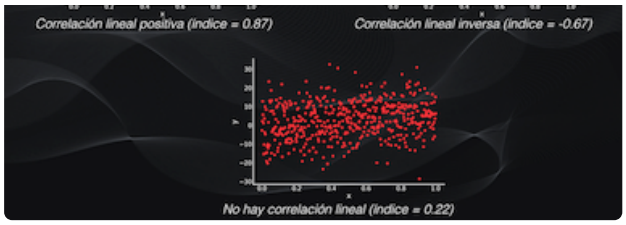


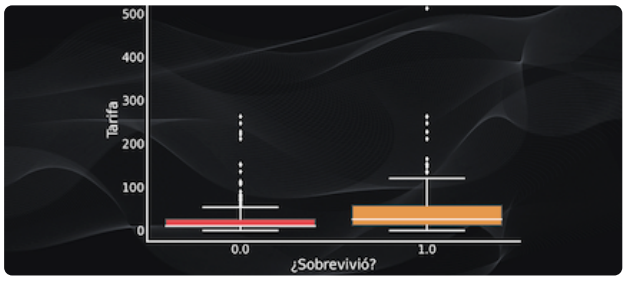
Gráfico de dispersión para las variables Tarifa y Edad

O podemos calcular el índice de correlación entre estas dos variables, donde un valor cercano a 1 nos indica una relación lineal, uno cercano a -1 una relación lineal inversa y un valor cercano a cero indica que no hay correlación lineal entre los datos (que es precisamente lo que ocurre en este ejemplo en particular):



Gráficos de dispersión para los diferentes tipos de correlación

También podemos comparar una variable numérica (como la tarifa) con una variable categórica (como la variable "supervivencia") y usar por ejemplo un *boxplot* para determinar si la tarifa está relacionada con la probabilidad de supervivencia:



Boxplots combinando una variable numérica (Tarifa) con una variable categórica (Supervivencia)

O para la misma comparación podemos usar un gráfico de violín, que es similar a un boxplot pero, además de mostrar la mediana y los límites de los cuartiles incluye una gráfica de densidad de la distribución, que es como una gráfica continua del histograma:

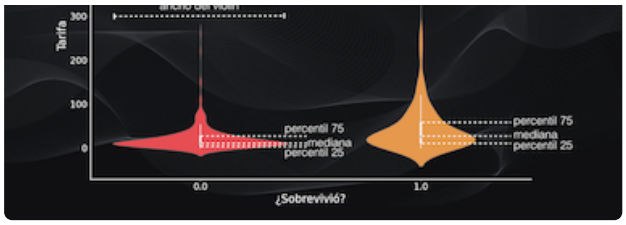


Gráfico de violín combinando una variable numérica (Tarifa) con una variable categórica (Supervivencia)

También podemos comparar dos variables categóricas (como por ejemplo el título del pasajero y la variable "supervivencia") usando gráficos de barras apiladas, lo que nos permite ver que en este caso la mayor parte de las pasajeras con categoría "señorita" sobrevivieron al hundimiento:

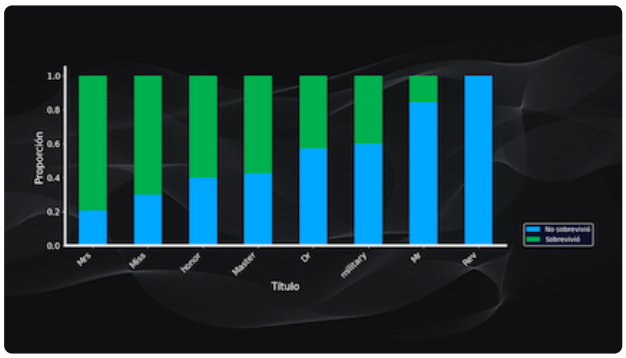


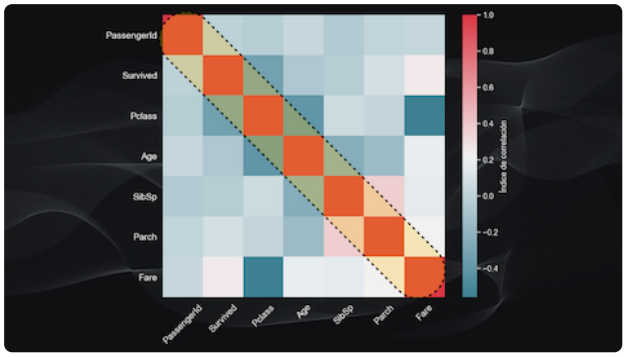
Gráfico de barras apiladas para las variables Título del pasajero y Supervivencia

Análisis multivariado

Por otra parte, en el análisis multivariado comparamos simultáneamente todos los posibles pares de variables para intentar encontrar algún tipo de relación.

Para cada comparación calculamos el índice de correlación entre diferentes pares de variables y dibujamos los

esta matriz tendremos valores iguales a 1, porque estamos comparando una variable consigo misma:



La matriz de correlación y su diagonal principal. Cada color corresponde a un nivel de correlación diferente (según la escala del lado derecho)

Pero lo que nos interesa es lo que está por fuera de esta diagonal. Por ejemplo, para el caso de nuestro dataset podemos ver que no existe relación alguna entre la clase del pasajero y la probabilidad de supervivencia, y podemos analizar en detalle diferentes pares de variables para ver si hay relaciones más relevantes que otras.

Paso 7: sumarización

Y la última fase de este análisis exploratorio consiste en sumarizar nuestras observaciones, es decir extraer las conclusiones más importantes del análisis que hemos venido realizando. En este caso sugiero escribirlas, como frases muy cortas.

Esto nos servirá para identificar por ejemplo qué variables están

codificandobits

Esto es fundamental para las etapas que vendrán más adelante en el proyecto, como el pre-procesamiento de los datos, la extracción de características o el desarrollo mismo del modelo en el caso del Machine Learning.

Enlace de descarga de la guía

Suscríbete a la *newsletter* mensual de Codificando Bits y recibe el enlace de descarga de la guía en tu e-mail:

Conclusión

Bien, con esto ya tenemos las principales fases del Análisis Exploratorio de los Datos.

También recuerda que se nos quedan por fuera dos etapas importantísimas de las que vamos a hablar en próximos artículos: el manejo de datos faltantes (es decir cuando nuestro

codificandobits

Otros artículos que te pueden interesar

- 5 pasos para aprender Machine Learning desde cero
- MLOps: el Machine Learning Engineering
- ¿Cuándo usar el Machine Learning?