Respecto a los notebooks.

Sección 2.1. Balance de clases => Equilibrado de clases

El punto 2.2 tiene un análisis de duplicados. Este punto es confuso. Normalmente, el que haya un duplicado indica que una instancia se ha metido dos veces. No obstante, esto no es así. En nuestro caso, un “duplicado” en programs son dos programas \*distintos\* que tienen las mismas características. Por ello, no deben catalogarse como duplicados, puesto que vamos a hacer ciencia da datos en lugar de machine learning. Lo mejor sería quitar esta sección.

Sección 2.3. Valores nulos. Si salen valores nulos, hay que hacer algo con ellos y decir qué es:

1. En determinados dominios, se le puede asignar un valor. Por ejemplo el 0, pero hay que discutirlo.
2. En otros dominios, se puede, simplemente, no considerar esa instancia. Por ejemplo, si calculamos la edad media, el individuo con valor nulo o se mete y se divide entre un valor una unidad menor.
3. Imputación: hay mil formas de imputar un valor. Una muy común es tomar la media (mediana si hay outliers) para numéricos o el valor más frecuente si es categórica.

necesaario => necesario

Sección 3

hará => hará

Para las variables numéricas, pones un diagrama que es un diagrama de dispersión con jitter. Conviene que expliques en algún sitio que el "jitter" en un gráfico es una técnica utilizada para agregar pequeñas cantidades de ruido aleatorio a los datos con el fin de evitar la superposición de puntos en gráficos de dispersión. Esto está bien para mostrar todos los datos (incluyes los outliers). No obstante, no es muy estándar.

Yo mantendría lo que tienes ahora, \*pero\*:

1. Indicaría que el gráfico actual incluye los outliers.
2. Pondría un histograma utilizando plt.hist (lo explico en la nueva versión de la metodología) tras quitar los outliers, indicando que esa gráfica no los incluye.

Respecto a los análisis de correlación estas utilizando Pearson. Para medir la correlación entre una variable numérica (que no necesariamente tiene una relación lineal) y una variable binaria codificada como 0 y 1, la correlación de Spearman es generalmente una opción más adecuada. La razón es que la correlación de Spearman mide la relación monotónica entre dos variables, lo cual es útil cuando no se puede suponer una relación lineal. Yo cambiaría todos los notebooks para utilizar Spearman.

A screen shot of a computer screen

Description automatically generated

![A graph of a graph

Description automatically generated]()

Si ves que un histograma es muy ancho y se ve mal, lo puedes poner en vertical, tal y como haces, por ejemplo, en “Variable category (1/9)” del notebook número 10.

Si quitando los outliers, todas las instancias tienen un mismo valor, simplemente lo muestras de forma textual. Es decir, si ves que un histograma es ridículo, no lo pongas.