

## **General Electric Aerospace**



## **Data Preparation**

Flavio Ruvalcaba Leija - A01367631

Oscar Eduardo Nieto Espitia - A01705090

Eduardo Gonzalez Luna - A01658281

Cristian Rogelio Espinoza Diaz - A01702752

Anatanael Jesus Miranda Faustino - A01769232

## **1. Criterios de inclusión y exclusión**

Los criterios de inclusión que definimos al inicio del proyecto son los siguientes:

- Datos con Calidad: Los datos deben ser de alta calidad y precisión. Esto implica la ausencia de errores, inconsistencias o valores atípicos que puedan afectar negativamente el rendimiento del modelo.
- Datos con Limpieza: Los datos deben estar limpios, lo que implica la eliminación de valores faltantes, duplicados y ruidos que puedan afectar la calidad de los resultados del modelo.
- Datos Relevantes: Datos que puedan aportar información significativa para el modelo.

Los criterios de exclusión que definimos al inicio del proyecto son los siguientes:

- Datos Irrelevantes: Excluir datos que no aportan información significativa para el modelo.
- Datos Duplicados: Eliminar registros duplicados o redundantes que tienen exactamente los mismos valores en todas las características.
- Valores Faltantes: Evaluar la cantidad y distribución de valores faltantes. Excluir registros con un alto porcentaje de valores faltantes si no pueden ser imputados de manera razonable.

## **2. Selección de datos**

Lo primero que decidimos hacer para descartar columnas es verificar la duplicidad de las variables. Esto surge a partir de que al registrar las estadísticas descriptivas en el [diccionario de datos](#), encontramos que las columnas vienen en pares, es decir, que hay dos columnas que posiblemente estuvieran registrando los mismos valores de las distintas simulaciones.

Para confirmar un poco nuestra teoría realizamos un análisis de multicolinealidad ya que la multicolinealidad es cuando dos o más variables independientes en un modelo de regresión están altamente correlacionadas entre sí.

Para este análisis usamos el factor de inflación de la varianza (VIF), conocida en inglés como Variance Inflation Factor, es una métrica utilizada en estadísticas y análisis de regresión para evaluar la multicolinealidad entre variables predictoras en un modelo de regresión.

	Variable	VIF
1	time	NaN
2	col1_bool	NaN
3	col2_bool	NaN
4	col3_bool	NaN
5	col4_bool	NaN
6	col5_float	inf
7	col6_float	inf
8	col7_float	inf
9	col8_float	inf
10	col9_float	inf
11	col10_float	inf
12	col11_float	NaN
13	col12_float	NaN
14	col13_float	inf
15	col14_float	inf
16	col15_float	inf
17	col16_float	inf
18	col17_float	inf
19	col18_float	inf
20	col19_float	inf
21	col20_float	inf
22	col21_float	inf
23	col22_float	inf
24	col23_float	inf
25	col24_float	inf
26	col25_float	inf
27	col26_float	inf
28	col27_float	inf
29	col28_float	inf
30	stableCruise_boolean	inf
31	stableCruise_boolean.1	inf
32	col31_integer	inf
33	col32_integer	inf
34	col33_integer	inf
35	col34_integer	inf
36	col35_integer	inf
37	col36_integer	inf

Fig 2. Resultados del análisis de multicolinealidad.

Los resultados nos dieron valores como “inf” o “nan”, un VIF igual a infinito para una variable específica sugiere una multicolinealidad perfecta con otras variables independientes en el modelo. Esto significa que la variable en cuestión puede ser completamente predicha a partir de las otras variables independientes en el modelo. En otras palabras, existe una relación lineal exacta entre esta variable y al menos una de las otras variables independientes.

La aparición de "NaN" en el cálculo del VIF generalmente se debe a la presencia de colinealidad perfecta, así como a la presencia de divisiones por cero en los cálculos matemáticos del VIF. En este caso, la matriz de correlación entre las variables independientes puede no ser de rango completo, lo que genera problemas numéricos.

En la matriz de correlación podemos observar estas “relaciones perfectas” en donde, cada par de columnas hay una correlación demasiado alta.

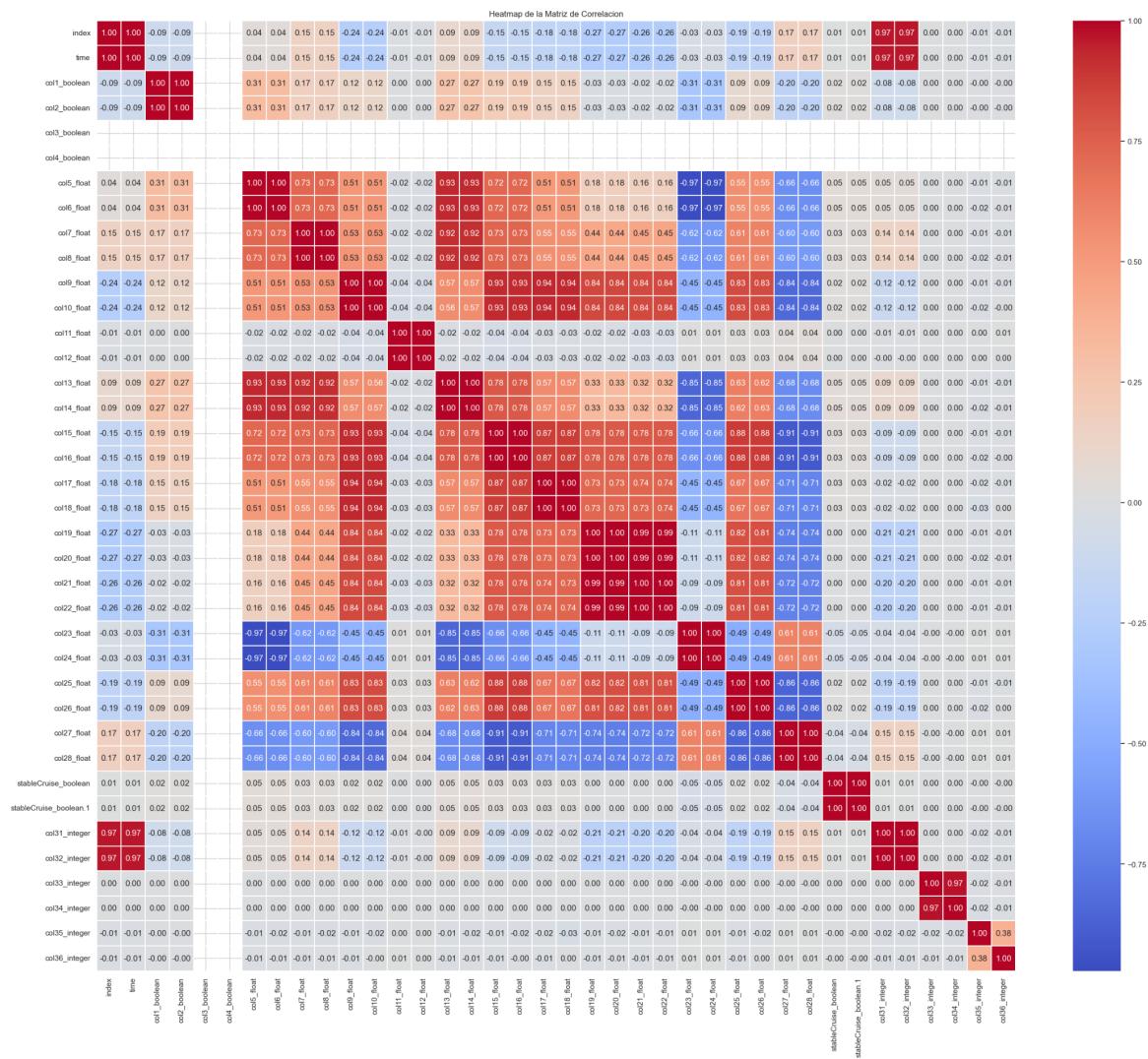


Fig 1. Matriz de Correlación

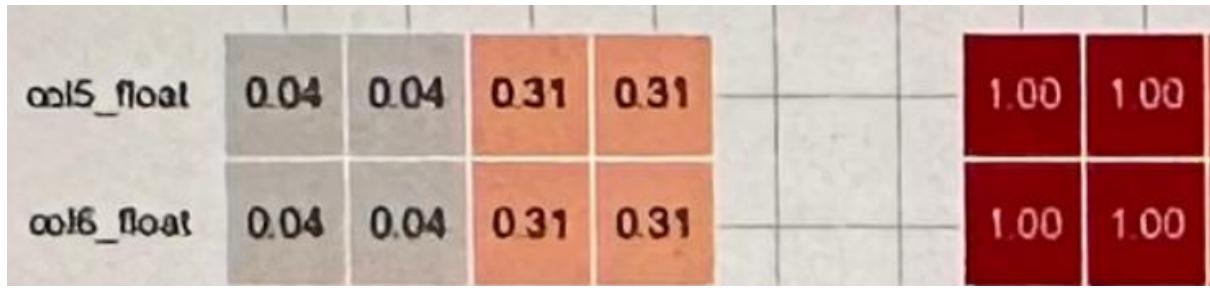


Fig 3. Ejemplo visual de la multicolinealidad en la matriz.

Posteriormente, se le cuestionó al representante de General Electric, Ignacio Mendieta, el cual nos confirmó que los datos en los csv son registros de 2 sensores distintos de las simulaciones.

Al tener una confirmación del socio formador, el equipo tuvo que tomar una decisión para cómo manejar esta situación ya que esto traería efectos negativos para la veracidad y calidad de los resultados del modelo.

Por consecuencia, el equipo decidió hacer un último análisis para corroborar que las diferencias estadísticas entre los pares de columnas es tan mínima como se observó en las estadísticas descriptivas.

Esta prueba se realizó con la técnica paired t-test. la cuál es una técnica estadística utilizada para comparar las medias de dos conjuntos de datos relacionados o emparejados. Esta prueba es útil cuando se desea determinar si existe una diferencia significativa entre las dos muestras emparejadas, lo que significa que estás interesado en comparar cómo cambian las observaciones antes y después de algún tipo de intervención o tratamiento.

```
results = paired_ttest_for_duplicated_columns(df)
for result in results:
    print(result)
#print(df.compute().head(5))
```

[38] ✓ 21m 48.2s

```
... {'Columna 1': 'col1_boolean', 'Columna 2': 'col2_boolean', 't-statistic': nan, 'p-value': nan, 'Significancia': 'No hay diferencia estadísticamente significativa'} {'Columna 1': 'col3_boolean', 'Columna 2': 'col4_boolean', 't-statistic': nan, 'p-value': nan, 'Significancia': 'No hay diferencia estadísticamente significativa'} {'Columna 1': 'col5_float', 'Columna 2': 'col6_float', 't-statistic': nan, 'p-value': nan, 'Significancia': 'No hay diferencia estadísticamente significativa'} {'Columna 1': 'col7_float', 'Columna 2': 'col8_float', 't-statistic': nan, 'p-value': nan, 'Significancia': 'No hay diferencia estadísticamente significativa'} {'Columna 1': 'col9_float', 'Columna 2': 'col10_float', 't-statistic': nan, 'p-value': nan, 'Significancia': 'No hay diferencia estadísticamente significativa'} {'Columna 1': 'col11_float', 'Columna 2': 'col12_float', 't-statistic': nan, 'p-value': nan, 'Significancia': 'No hay diferencia estadísticamente significativa'} {'Columna 1': 'col13_float', 'Columna 2': 'col14_float', 't-statistic': nan, 'p-value': nan, 'Significancia': 'No hay diferencia estadísticamente significativa'} {'Columna 1': 'col15_float', 'Columna 2': 'col16_float', 't-statistic': nan, 'p-value': nan, 'Significancia': 'No hay diferencia estadísticamente significativa'} {'Columna 1': 'col17_float', 'Columna 2': 'col18_float', 't-statistic': nan, 'p-value': nan, 'Significancia': 'No hay diferencia estadísticamente significativa'} {'Columna 1': 'col19_float', 'Columna 2': 'col20_float', 't-statistic': nan, 'p-value': nan, 'Significancia': 'No hay diferencia estadísticamente significativa'} {'Columna 1': 'col21_float', 'Columna 2': 'col22_float', 't-statistic': nan, 'p-value': nan, 'Significancia': 'No hay diferencia estadísticamente significativa'} {'Columna 1': 'col23_float', 'Columna 2': 'col24_float', 't-statistic': nan, 'p-value': nan, 'Significancia': 'No hay diferencia estadísticamente significativa'} {'Columna 1': 'col25_float', 'Columna 2': 'col26_float', 't-statistic': nan, 'p-value': nan, 'Significancia': 'No hay diferencia estadísticamente significativa'} {'Columna 1': 'col27_float', 'Columna 2': 'col28_float', 't-statistic': nan, 'p-value': nan, 'Significancia': 'No hay diferencia estadísticamente significativa'} {'Columna 1': 'stablecruise_boolean', 'Columna 2': 'stablecruise_boolean', 't-statistic': nan, 'p-value': nan, 'Significancia': 'No hay diferencia estadísticamente significativa'} {'Columna 1': 'col31_integer', 'Columna 2': 'col32_integer', 't-statistic': nan, 'p-value': nan, 'Significancia': 'No hay diferencia estadísticamente significativa'} {'Columna 1': 'col33_integer', 'Columna 2': 'col34_integer', 't-statistic': nan, 'p-value': nan, 'Significancia': 'No hay diferencia estadísticamente significativa'} {'Columna 1': 'col35_integer', 'Columna 2': 'col36_integer', 't-statistic': nan, 'p-value': nan, 'Significancia': 'No hay diferencia estadísticamente significativa'}
```

Fig 4. Resultados de paired t-test

Estos resultados nos indican que la diferencia absoluta entre las observaciones emparejadas en tus columnas es muy pequeña, lo que puede potencialmente llevar a valores "nan" en el estadístico t y el valor p. Esto puede ocurrir porque 'paired t-test' implica la división por el error estándar, y si el error estándar está muy cerca de cero, puede dar como resultado valores indefinidos o "nan".

Con estos resultados, estamos garantizando que la información que posee cada par de datos duplicados no posee diferencias significativas en los datos que poseen cada par de columnas duplicadas. Esto nos indica una solución viable, la cual es quedarnos con los datos correspondientes a un sensor ya que no estaríamos perdiendo información.

Decidimos quedarnos con las columnas impares ya que estos son los datos de un sensor acorde al socio formador. Aunque esta decisión no posee mucho impacto ya que pudimos haber escogido las columnas pares y obtener el mismo resultado según nuestros análisis.

La última decisión que se tomó fue sobre escoger las variables más correlacionadas con la variable dependiente y el objetivo principal del modelo, la detección de 'Stable Cruise'.

Para esto decidimos correr otra matriz de correlación, después de eliminar las columnas pares. Con el propósito de tomar una buena decisión sobre qué variables independientes poseen las correlaciones más altas.

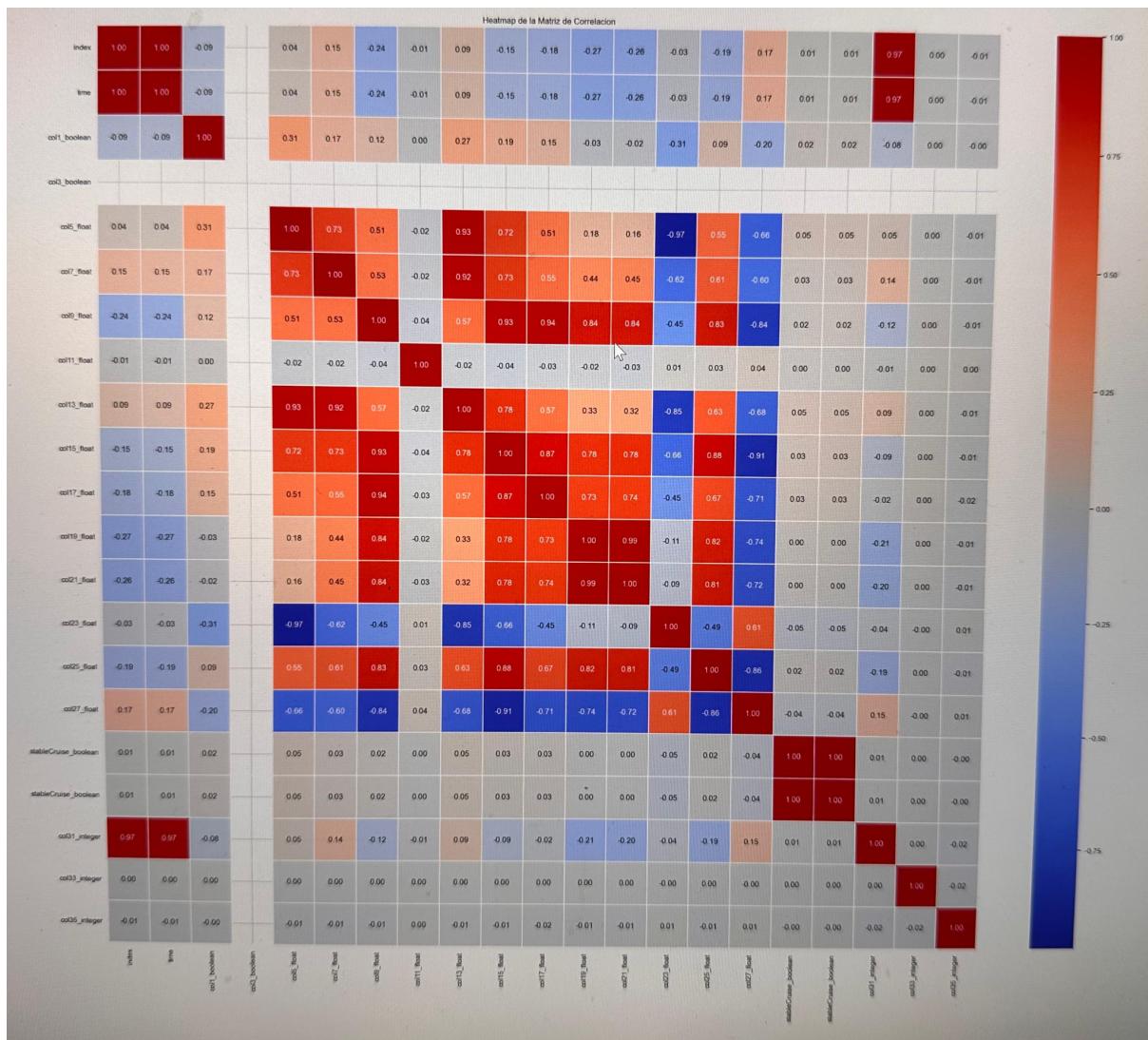


Fig 5. Matriz de correlación con solo columnas impares

Con base en la matriz de correlación, establecimos que las variables con una correlación mayor o igual a 0.03/-0.03 ya que estos nos indica una correlación moderada con los valores que tenemos a nuestra disposición.

A partir de este rango, las variables independientes que se incluirán en el data frame final son los siguientes:

- col23\_float con -0.05
- col13\_float con 0.05
- col5\_float con 0.05
- col7\_float con 0.05
- col27\_float con -0.04
- col15\_float con 0.03
- col17\_float con 0.03

Estas 7 columnas son características que cumplen con nuestros criterios de inclusión.

Un aspecto importante a tener en consideración es una posible reducción del tamaño del dataset para los modelos. Debido a las tecnologías que estamos usando consideramos que 2718 archivos pueden sobrepasar el límite de librerías como ‘Dask’, ‘Pandas’, etc.

Un acercamiento que tuvimos en la fase de Data Understanding, fue solo usar los archivos que tuvieran 1 en las columnas relacionadas a Stable Cruise. Esto nos permitió acortar el alcance de 2718 archivos csv a 737.

Otro factor importante a tener en cuenta, es que la cantidad de filas que posee la clase mayoritaria “0” es mucho mayor que las filas que posee “1”. De las

Variable	Descripción	Correlación con dependiente	t-test con su duplicado	p-value con su duplicado
time	A pesar de no ser estadísticamente significativa, esta variable nos puede ayudar a graficar las variables con respecto al tiempo. No se planea entrenar el modelo con esta variable	0.01		
col1_boolean	No es estadísticamente significativa con la variable dependiente.	0.02		
col2_boolean	Columna no aporta valor según el Pair t-test. Correlación poco significativa. Aporta poca calidad al modelo.	0.02	NaN	NaN
col3_boolean	Fue excluida debido a que todos sus registros eran cero. Lo cual no aporta nada de información ni en las estadísticas descriptivas ni en la matriz de correlación.	N/A		
col4_boolean	Fue excluida debido a que todos sus registros eran cero. Lo cual no aporta nada de información ni en las estadísticas descriptivas ni en la matriz de correlación.	N/A	NaN	NaN
col5_float	Estadísticamente significante según la correlación con la variable dependiente.	0.05		
col6_float	Columna no aporta valor según el Pair t-test. Aporta poca calidad al modelo.	0.05	NaN	NaN
col7_float	Estadísticamente significante según la correlación con la variable dependiente.	0.03		
col8_float	Columna no aporta valor según el Pair t-test. Aporta poca calidad al modelo.	0.03	NaN	NaN
col9_float	No es estadísticamente significativa con la variable dependiente.	0.02		
col10_float	Columna no aporta valor según el Pair t-test. Aporta poca calidad al modelo.	0.02	NaN	NaN
col11_float	No es estadísticamente significativa con la variable dependiente.	≈ 0.00		

col12_float	Columna no aporta valor según el Pair t-test. Aporta poca calidad al modelo.	$\approx 0.00$	NaN	NaN
col13_float	Estadísticamente significante según la correlación con la variable dependiente.	0.05		
col14_float	Columna no aporta valor según el Pair t-test. Aporta poca calidad al modelo.	0.05	NaN	NaN
col15_float	Estadísticamente significante según la correlación con la variable dependiente.	0.03		
col16_float	Columna no aporta valor según el Pair t-test. Aporta poca calidad al modelo.	0.03	NaN	NaN
col17_float	Estadísticamente significante según la correlación con la variable dependiente.	0.03		
col18_float	Columna no aporta valor según el Pair t-test. Aporta poca calidad al modelo.	0.03	NaN	NaN
col19_float	No es estadísticamente significativa con la variable dependiente.	$\approx 0.00$		
col20_float	Columna no aporta valor según el Pair t-test. Aporta poca calidad al modelo.	$\approx 0.00$	NaN	NaN
col21_float	No es estadísticamente significativa con la variable dependiente.	$\approx 0.00$		
col22_float	Columna no aporta valor según el Pair t-test. Aporta poca calidad al modelo.	$\approx 0.00$	NaN	NaN
col23_float	Estadísticamente significante según la correlación con la variable dependiente.	-0.05		
col24_float	Columna no aporta valor según el Pair t-test. Aporta poca calidad al modelo.	-0.05	NaN	NaN
col25_float	No es estadísticamente significativa con la variable dependiente.	0.02		
col26_float	Columna no aporta valor según el Pair t-test. Aporta poca calidad al modelo.	0.02	NaN	NaN
col27_float	Estadísticamente significante según la correlación con la variable dependiente.	-0.04		
col28_float	Columna no aporta valor según el Pair t-test. Aporta poca calidad al modelo.	-0.04	NaN	NaN
stableCruise_boolean	Variable dependiente	1		
stableCruise_boolean.1	Columna no aporta valor según el Pair t-test. Aporta poca calidad al modelo.	1	NaN	NaN
col31_integer	No es estadísticamente significativa con la variable dependiente.	0.01	NaN	NaN
col32_integer	Columna no aporta valor según el Pair t-test. Aporta poca calidad al modelo.	0.01		
col33_integer	No es estadísticamente significativa con la variable	$\approx 0.00$	NaN	NaN

	dependiente.			
col34_integer	Columna no aporta valor según el Pair t-test. Aporta poca calidad al modelo.	≈ 0.00	NaN	NaN
col35_integer	No es estadísticamente significativa con la variable dependiente.	≈ -0.00		
col36_integer	Columna no aporta valor según el Pair t-test. Aporta poca calidad al modelo.	≈ -0.00	NaN	NaN

### 3. Limpieza de los datos

#### 3.1 Análisis de multicolinealidad

Antes de empezar a trabajar sobre estas variables decidimos hacer un análisis de multicolinealidad para verificar que las variables no estuvieran altamente relacionadas entre sí. La multicolinealidad es un término utilizado en estadísticas y regresión, esta se refiere a una situación en la que dos o más variables independientes en un modelo de regresión están altamente correlacionadas entre sí, lo que dificulta la capacidad del modelo para distinguir el efecto individual de cada variable en la variable dependiente. En otras palabras, puede hacer que sea difícil determinar qué variable está contribuyendo de manera única a los resultados, lo que puede afectar la interpretación y precisión del modelo de regresión. Es importante identificar y abordar la multicolinealidad para obtener resultados confiables en el análisis estadístico. A pesar de que la multicolinealidad es un problema afecta mayormente a modelos de regresión y no a clasificación, consideramos que es importante hacer este análisis antes de empezar a limpiar los datos.

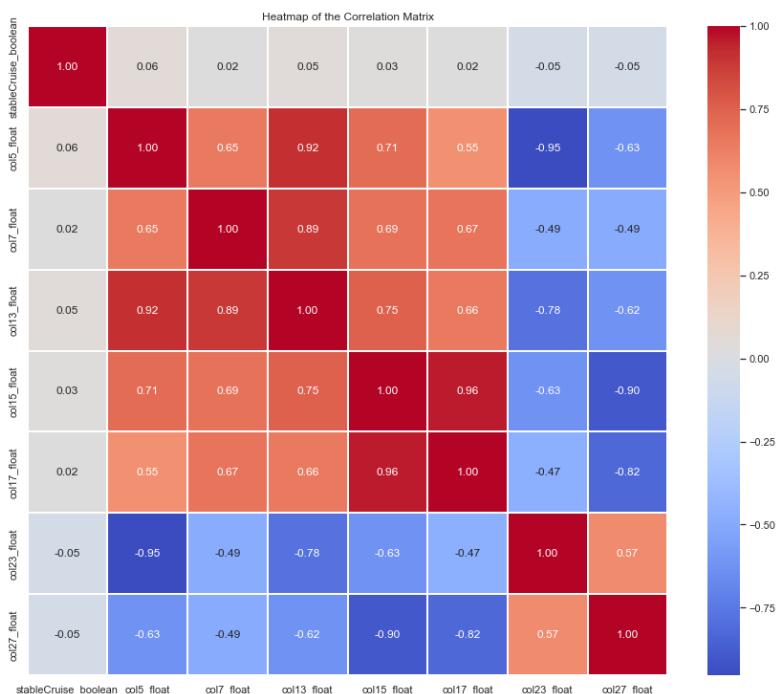
Tras ejecutar el análisis de multicolinealidad obtuvimos el siguiente resultado:

Variable	VIF
col5_float	718.07
col7_float	392.53
col13_float	1360.49
col15_float	177.66
col17_float	187.65
col23_float	43.57
col27_float	15.70

Como se puede observar las variables “col5\_float” y “col13\_float” son aquellas variables que tienen mayor multicolinealidad, usualmente un valor debajo de 5 es indicativo de que no existe multicolinealidad. Estas son las maneras de lidiar con la multicolinealidad:

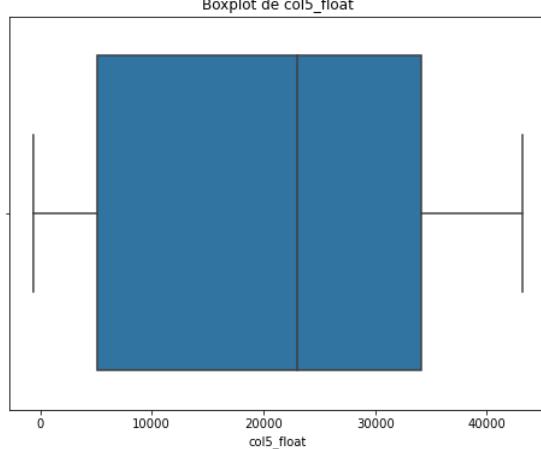
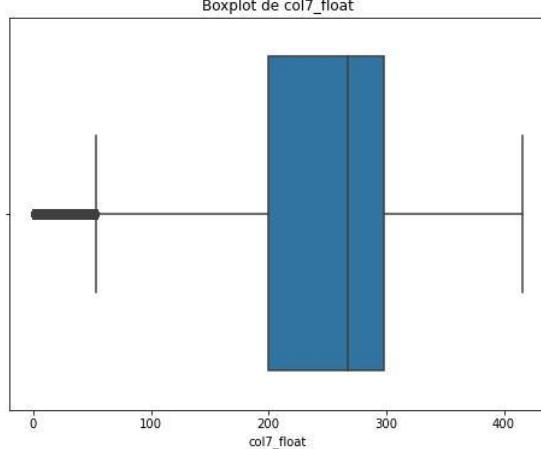
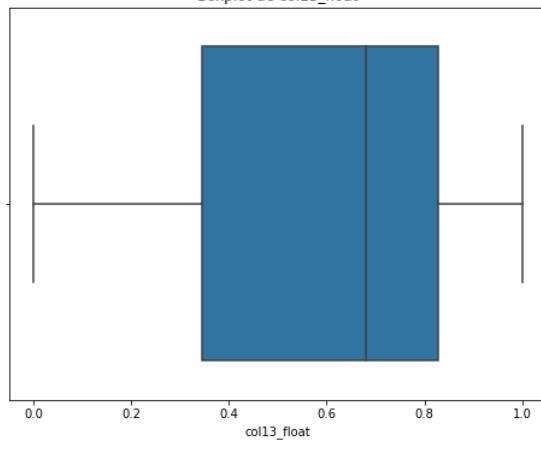
- **Eliminación de variables:** Eliminar las variables altamente correlacionadas con otras.
- **Transformación de variables:** Se puede aplicar transformaciones a las variables para reducir la multicolinealidad. Por ejemplo, realizar una reducción de dimensionalidad como Análisis de Componentes Principales o una combinación de variables puede ayudar.
- **Regularización:** Utilizar técnicas de regularización como Lasso o Ridge para penalizar coeficientes altos y reducir la multicolinealidad.

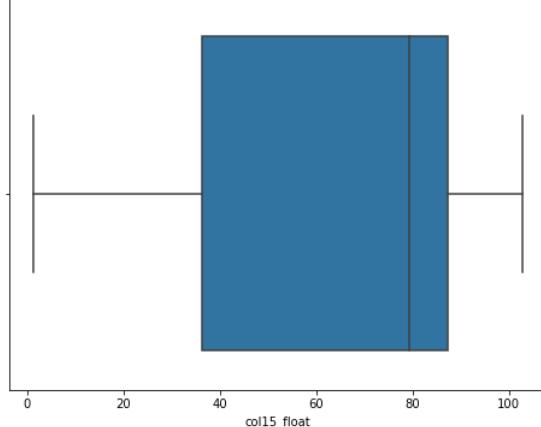
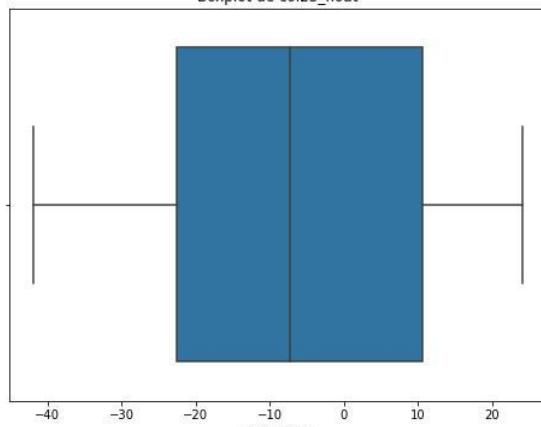
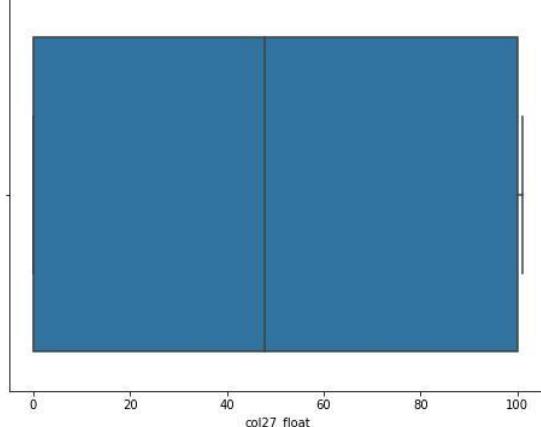
Se decidió eliminar la variable “col17\_float” debido a que tiene un índice de correlación muy alto con “col15\_float” que tuvo una correlación de 0.96. Esta correlación es la más alta que se obtuvo tras filtrar las variables la cual indica que prácticamente es la misma variable al tener un índice de correlación tan alto.

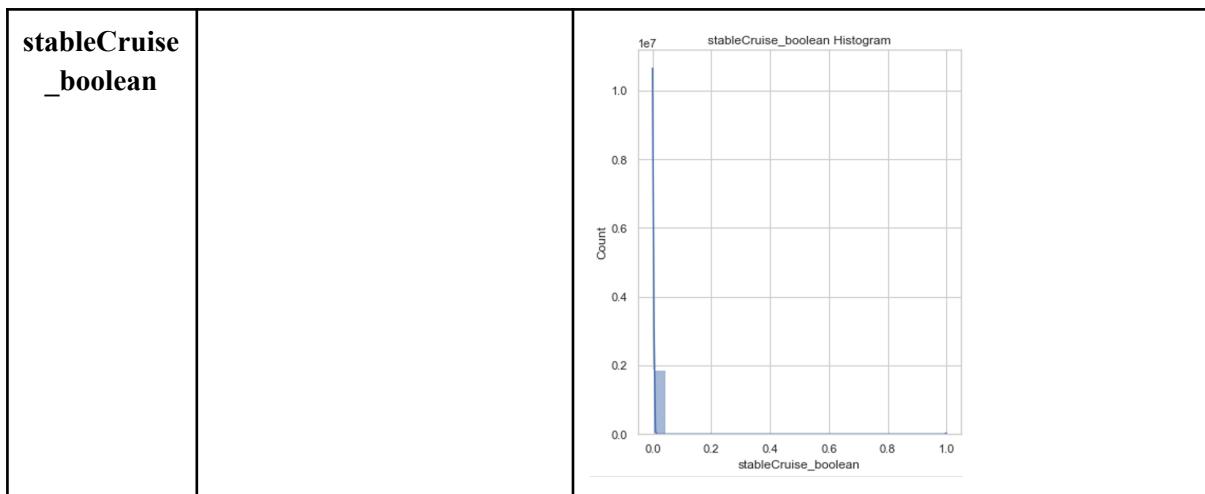


Para el resto de variables por ahora se optó por utilizar técnicas de regularización para atender los problemas de multicolinealidad. Estos métodos de regularización penalizan a las variables con mayor multicolinealidad para así penalizar los pesos a la hora de entrenar el modelo.

### 3.2 Manejo de valores atípicos

col5_float	<p><b>Puntos atípicos (Outliers):</b> No tiene puntos individuales fuera de los whiskers lo cual significa que no posee valores atípicos.</p>	 <p>Boxplot de col5_float</p> <p>Este boxplot muestra que la variable col5_float no contiene valores atípicos. Los whiskers extienden desde alrededor de 1000 hasta 40000, y el cuadro interquartil (IQR) cubre el rango de aproximadamente 15000 a 35000. La media (línea horizontal interior) se sitúa entre el primer cuartil (Q1) y el tercer cuartil (Q3).</p>
col7_float	<p><b>Puntos atípicos (Outliers):</b> Cómo se puede observar, hay puntos fuera de los whiskers del gráfico lo que significa que esta variable tiene valores atípicos.</p>	 <p>Boxplot de col7_float</p> <p>Este boxplot muestra que la variable col7_float contiene valores atípicos. Los whiskers extienden desde alrededor de 20 hasta 400, pero existen valores individuales (outliers) que se extienden más allá de estos límites, situados por debajo de 100 y por encima de 400.</p>
<p>La técnica de winsorization es una forma efectiva de tratar valores atípicos. La winsorization implica reemplazar los valores extremos por límites predefinidos, en lugar de eliminarlos por completo. Esto ayuda a reducir el impacto de los valores atípicos en el análisis de datos sin eliminar información importante.</p> <p>Para tratar los valores atípicos vamos a usar esta técnica.</p>		
col13_float	<p><b>Puntos atípicos (Outliers):</b> No tiene puntos individuales fuera de los whiskers lo cual significa que no posee valores atípicos.</p>	 <p>Boxplot de col13_float</p> <p>Este boxplot muestra que la variable col13_float no contiene valores atípicos. Los whiskers extienden desde 0.0 hasta 1.0, y el cuadro interquartil (IQR) cubre el rango de aproximadamente 0.4 a 0.8. La media (línea horizontal interior) se sitúa entre el primer cuartil (Q1) y el tercer cuartil (Q3).</p>

col15_float	<p><b>Puntos atípicos (Outliers):</b> No tiene puntos individuales fuera de los whiskers lo cual significa que no posee valores atípicos.</p>	<p>Boxplot de col15_float</p>  <p>col15_float</p>
col23_float	<p><b>Puntos atípicos (Outliers):</b> No tiene puntos individuales fuera de los whiskers lo cual significa que no posee valores atípicos.</p>	<p>Boxplot de col23_float</p>  <p>col23_float</p>
col27_float	<p><b>Puntos atípicos (Outliers):</b> No tiene puntos individuales fuera de los whiskers lo cual significa que no posee valores atípicos.</p>	<p>Boxplot de col27_float</p>  <p>col27_float</p>



### 3.3 Calidad de los datos

## 4. Construir los datos

### 3.3 Tipos de escalamiento de datos

En el proceso de escalamiento de datos, investigamos los diferentes tipos de herramientas que se acoplan a nuestro diseño del modelo. Posteriormente, hicimos una tabla para identificar el más beneficioso para el proyecto:

	<b>Normalización</b>	<b>Estandarización</b>
<b>Descripción</b>	La normalización, también conocida como escalamiento min-max, transforma las características a un rango específico, generalmente entre 0 y 1. Esto se logra restando el valor mínimo y dividiendo por la diferencia entre el valor máximo y mínimo.	La estandarización, también conocida como normalización Z-score, transforma las características restando la media y dividiendo por la desviación estándar. Esto resulta en una distribución con media cero y desviación estándar igual a uno.
<b>Ventajas</b>	La normalización mantiene la forma de distribución original de los datos y garantiza que todas las características estén en el mismo rango. Esto puede ser útil para algoritmos que requieren características en un rango específico, como redes neuronales con funciones de activación limitadas.	La estandarización es menos sensible a los valores atípicos, lo que significa que los valores extremos no afectarán significativamente el resultado final. También puede mejorar la convergencia y el rendimiento de los algoritmos de aprendizaje automático.

<b>Desventajas</b>	<p>La normalización es sensible a los valores atípicos, lo que significa que los valores extremos pueden afectar significativamente los resultados. Además, la normalización puede comprimir los datos si hay una gran variabilidad en los valores, lo que puede resultar en una pérdida de información.</p>	<p>La estandarización no garantiza una distribución específica de los datos y puede resultar en valores negativos y positivos. Esto puede ser problemático si se requiere una distribución específica, como una distribución entre 0 y 1.</p>
--------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

En nuestra investigación, encontramos otros dos métodos: Binarización y Discretización. Desafortunadamente, no los utilizamos por las siguientes cuestiones:

**Binarización:** La binarización convierte los datos en valores binarios (0 o 1), lo que implica una pérdida de información. Esta pérdida puede ser significativa, especialmente si los datos originales tenían valores continuos o categorías múltiples. Es mayormente utilizado en One Hot Encoding.

**Discretización:** La discretización puede llevar a una pérdida de información valiosa al agrupar los datos en intervalos o categorías. Esto puede resultar en una representación inexacta de los datos originales y, a su vez, en una interpretación errónea de los resultados. Además, la discretización puede introducir sesgos o distorsiones en los datos. Dependiendo de cómo se realice la discretización y de los valores de corte seleccionados, es posible que se creen categorías que no reflejen adecuadamente la distribución real de los datos.

Finalmente, con la investigación efectuada, conocemos las ventajas y desventajas de cada una de las herramientas que nos pudieron ayudar a proceder el escalamiento de datos. Con nuestra decisión final de utilizar el método de estandarización para todo el dataset.

## 5. Integrar los datos

### 5.1 Generación de atributos derivados

Generar nuevos registros o atributos derivados puede ser beneficioso en muchos casos, pero debe hacerse con precaución, especialmente si los datos están anonimizados y no se conocen las variables originales. Estos son los puntos que consideramos para tomar esta decisión:

- **Dominio del problema:** La generación de atributos derivados generalmente se basa en un profundo conocimiento del dominio del problema. Si no se conoce la naturaleza de las variables originales debido a la anonimización, es difícil aplicar este enfoque de manera efectiva.
- **Riesgo información sensible:** Si los datos están anonimizados para proteger información sensible, la generación de nuevos atributos podría inadvertidamente revelar información confidencial. Esto podría tener implicaciones legales y éticas.

Dada la falta de información sobre las variables originales y el anonimato de los datos, se decidió no utilizar ninguna variable derivada.

## 6. Formatear los datos

### **Descripción de Dataset:**

Una vez realizado el dataset final, nos quedamos con las siguientes seis columnas:

col5\_float: Esta columna float, mediante el análisis de multicolinealidad, definimos que es estadísticamente significante según la correlación con la variable dependiente.

col7\_float: Esta columna float, mediante el análisis de multicolinealidad, definimos que es estadísticamente significante según la correlación con la variable dependiente.

col13\_float: Esta columna float, mediante el análisis de multicolinealidad, definimos que es estadísticamente significante según la correlación con la variable dependiente.

col15\_float: Esta columna float, mediante el análisis de multicolinealidad, definimos que es estadísticamente significante según la correlación con la variable dependiente.

col23\_float: Esta columna float, mediante el análisis de multicolinealidad, definimos que es estadísticamente significante según la correlación con la variable dependiente.

col27\_float: Esta columna float, mediante el análisis de multicolinealidad, definimos que es estadísticamente significante según la correlación con la variable dependiente.

stableCruise\_boolean: Esta columna booleana, es la variable dependiente que utilizaremos para realizar el modelo.