

# PEC1. Análisis de datos Ómicos.

M<sup>a</sup> de la Vega Rodrigálvarez Chamarro

6 de noviembre, 2024

## Contents

<b>1</b>	<b>Resumen ejecutivo</b>	<b>1</b>
<b>2</b>	<b>Objetivos del estudio</b>	<b>2</b>
<b>3</b>	<b>Materiales y métodos</b>	<b>2</b>
<b>4</b>	<b>Resultados</b>	<b>3</b>
<b>5</b>	<b>Discusión, limitaciones y conclusiones del estudio</b>	<b>11</b>
<b>6</b>	<b>Notas</b>	<b>12</b>
	<b>References</b>	<b>12</b>

## 1 Resumen ejecutivo

Este estudio realiza un análisis exploratorio de datos de resonancia magnética (NMR) de muestras de orina, orientado al diagnóstico precoz del cáncer gástrico. A través del conjunto de datos 2023-CIMCBTutorial, se han analizado 149 metabolitos en cuatro categorías de pacientes: cáncer gástrico (GC), enfermedad benigna (BN), individuos sanos (HE), y un grupo de control (QC). Se evaluaron diferentes metabolitos y su relación con la clase del paciente, eliminando aquellos que no cumplían los criterios de calidad (tenían una variabilidad mayor del 20% y su completitud era menor al 90%).

Los datos fueron procesados usando herramientas como *SummarizedExperiment* de la librería *BioConductor*. En el análisis, se realizaron pruebas de varianza (ANOVA) y componentes principales (PCA), identificando 20 metabolitos significativos. Sin embargo, no se observaron agrupaciones claras ni patrones que diferencien a los pacientes con cáncer gástrico, ni en los gráficos de calor ni en los diagramas de caja. Se concluye que el análisis exploratorio inicial es insuficiente y que serían necesarios modelos adicionales de machine learning para optimizar la detección temprana del cáncer gástrico.

## 2 Objetivos del estudio

El principal objetivo del trabajo es la realización de un análisis exploratorio del conjunto de datos **2023-CIMCBTutorial** que ofrece datos de resonancias magnéticas (*NMR*) de un estudio del cáncer gástrico utilizados en un tutorial de análisis metabolómicos. El conjunto de datos contiene un total de 149 variables metabolómicas recogidas de 140 individuos humanos. Cada individuo se encuentra clasificado dentro de una categoría según su perfil metabolómico en orina  $^1\text{H-NMR}$  donde **GC** es cáncer gástrico, **BN** es benigno y **HE** es un individuo sano. Existe un cuarto grupo que es el denominado grupo de control (**QC**).

Se pretende estudiar los diferentes metabolitos extraídos y si existe alguna relación con la clase a la que pertenece el individuo con el objeto de poder realizar un diagnóstico precoz del cáncer gástrico. Se eliminarán todas aquellas variables que no sean de calidad y se decidirá si es necesario continuar analizar diferentes modelos de Machine Learning que permitan ese diagnóstico.

Como objetivo secundario, se trabajará con el objeto *SummarizedExperiment* perteneciente a la librería *BioConductor* con la finalidad de conocer mejor el uso de dicha librería.

## 3 Materiales y métodos

Los datos utilizados para el presente análisis exploratorio han sido descargados del repositorio de github *metaboData*<sup>1</sup>, publicado por Alex Sánchez Pla. El conjunto de datos seleccionado es el que proviene de un estudio de cáncer gástrico usado en un tutorial de análisis de datos metabolómicos<sup>2</sup>.

El origen de estos datos provienen de un estudio, desarrollado por la Universidad de Alberta y cuyo investigador principal es David Broadhurst (PR000699), cuyo principal objetivo del estudio era identificar si el cáncer gástrico (GC) tiene un perfil metabolómico urinario diferente en comparación con la enfermedad gástrica benigna (BN) y los pacientes sanos (HE). Para ello, se han recogido muestras de 43 pacientes con cáncer gástrico, 40 con enfermedad gástrica benigna y 40 pacientes sanos. Además, se han recogido muestras de otros 17 pacientes que actuaran como grupo de control de la calidad. En este estudio, se llegó a la conclusión que los pacientes con GC presentaban un perfil metabolómico urinario diferenciado del resto, dejando claro el potencial que el perfil metabolómico puede tener para el diagnóstico precoz de cáncer gástrico.

Para cada individuo se ha recogido la concentración de un total de 149 variables metabolómicas diferentes. Además de estas variables, se encuentra la columna *SampleType* donde se indica si el individuo forma parte del grupo de control, o de las muestras y la columna *Class* que indica el resultado clínico observado para cada individuo.

Para cada una de las medidas metabolómicas se encuentran los siguientes metadatos asociados:

- *Name*: Nombre asignado a la concentración metabolómica.
- *Label*: Proporciona un nombre único para cada metabolito.
- *Perc\_missing*: Indica el número de muestras que no tienen la medida de dicho metabolito (datos perdidos).
- *QC\_RSD*: es una puntuación de calidad que representa la variación en las mediciones de este metabolito en todas las muestras.

Estos datos se encontraban almacenados en un libro excel. En la hoja denominada *Data* se encontraban recogidas las concentraciones metabolómicas para cada uno de los individuos, así como, la información relativa al tipo de muestra y la clase de la misma. En la hoja denominada *Peak* se encuentran los metadatos de las concentraciones metabolómicas.

Para poder utilizar los datos, se han cargados en un objeto de clase *SummarizedExperiment* que pertenece a librería *Bioconductor*. Los datos relativos a las concentraciones metabolómicas han sido cargados dentro

---

<sup>1</sup><https://github.com/nutrimetabolomics/metaboData/>

<sup>2</sup>Basic Metabolomics Data Analysis Workflow

de una matriz de ensayos (assays), para ello ha sido necesario cargar la hoja *Data* del libro Excel, tomar desde la columna 5 a la 153 y hacer su transformada para que las filas correspondan con las variables de la concentración y las columnas con las muestras para cada individuo. Dentro de *rowData* se han insertado los metadatos de las concentraciones que se encuentran en la hoja *Peak*, que previamente habían sido cargados en un dataframe. Las columnas entre la 3 y la 5 de la hoja *Data* son los metadatos de las diferentes muestras que han sido cargados en un dataframe y almacenados en *colData*. La descripción de la información que contiene este archivo se ha cargado desde el fichero *description.md* en una lista y se ha almacenado bajo *metadata*.

Antes de comenzar con el análisis exploratorio de los datos, es importante evaluar la calidad de los datos y borrar aquellas variables que no tengan la suficiente información o realizar imputaciones con valores estándar cuando la falta de información sea menor tal y como se indica en (Broadhurst et al., 2018). Para este conjunto de datos únicamente se mantendrán los metabolitos que cumplan los siguientes criterios:

- El valor de **QC-RCD** sea menor o igual del 20%, es decir, que la variación de las medidas a lo largo de todas las muestras sea menor del 20%.
- Aquellos cuyo completitud sea mayor del 90%, es decir, que más del 90% de los individuos tengan recogida esa concentración.
- Para aquellos metabolitos que tengan valores perdidos en alguna de sus muestra, pero que no supere el 10% de datos perdidos, se realizará una imputación de las variables aplicando el algoritmo de los vecinos más cercanos (knn).

Una vez descartadas todas aquellas variables que no tienen calidad suficiente, se ha procedido a realizar un análisis univariante donde se muestra el rango de valores que toma cada metabolito (valor mínimo, valor 1º cuartil, mediana, media, valor 3º cuartil, valor máximo). Además, se ha realizado un análisis de la varianza (ANOVA), para cada una de las variables, donde se puede ver si una variable por separado es significativa o no con respecto a las clases (CG, BN, HE). Por otra parte, se ha generado una matriz de correlación para estudiar aquellas variables que están estrechamente interrelacionadas unas con otras y que pueden causar un problema de multicolinealidad a la hora de crear un futuro modelo.

Una vez finalizado el análisis univariante, se realiza un análisis de componentes principales (PCA) con el fin de reducir la dimensionalidad perdiendo la menor cantidad de información posible donde los nuevos componentes principales o factores son una combinación lineal de las variables originales, y además son independientes entre sí.

Estos análisis se han realizado con todos los metabolitos que cumplen los criterios de calidad establecidos al inicio y únicamente con aquellas variables significativas en función del pValor obtenido con el análisis de la varianza.

## 4 Resultados

### *Preprocesado de la información*

Una vez realizada la limpieza de datos, la matriz de metabolitos se ha reducido considerablemente quedando un total de 53 variables tomadas de 140 individuos. A continuación, se muestra el listado de las variables que se van a someter a estudio.

Table 1: Metabolitos a estudio

	Name	Label	Perc_missing	QC_RSD
M4	M4	1-Methylnicotinamide	8.5714286	12.804201
M5	M5	2-Aminoadipate	1.4285714	9.372664
M7	M7	2-Furoylglycine	2.8571429	5.049156

	Name	Label	Perc_missing	QC_RSD
M8	M8	2-Hydroxyisobutyrate	0.0000000	5.132340
M11	M11	3-Aminoisobutyrate	5.0000000	15.476165
M14	M14	3-Hydroxyisobutyrate	2.1428571	8.905711
M15	M15	3-Hydroxyisovalerate	0.0000000	4.200837
M25	M25	6-Hydroxynicotinate	0.0000000	18.149034
M26	M26	ATP	5.7142857	16.420150
M31	M31	Adipate	0.7142857	10.475666
M32	M32	Alanine	0.0000000	4.163279
M33	M33	Anserine	0.0000000	9.499714
M36	M36	Asparagine	9.2857143	7.356297
M37	M37	Azelate	0.0000000	10.527003
M45	M45	Citrate	0.0000000	6.268332
M48	M48	Creatinine	0.0000000	3.234995
M50	M50	Ethanol	1.4285714	18.826079
M51	M51	Ethanolamine	7.1428571	14.358552
M65	M65	Glycylproline	0.0000000	10.611583
M66	M66	Hippurate	0.0000000	2.431571
M68	M68	Histidine	2.8571429	16.192165
M71	M71	Ibuprofen	0.0000000	8.715085
M73	M73	Indole-3-lactate	0.0000000	6.073295
M74	M74	Isoleucine	5.0000000	13.280452
M75	M75	Lactate	0.0000000	14.883399
M88	M88	N-Acetylglutamine	0.7142857	5.656263
M89	M89	N-AcetylglutamineDerivative	0.0000000	4.069852
M90	M90	N-Acetylornithine	0.0000000	10.542754
M91	M91	N-Acetylserotonin	0.7142857	15.704143
M93	M93	N-Methylhydantoin	0.0000000	17.293923
M101	M101	Pantothenate	1.4285714	6.648222
M104	M104	Proline	0.7142857	9.364352
M105	M105	Propylene glycol	2.8571429	8.155064
M106	M106	Pyridoxine	0.7142857	15.932042
M107	M107	Pyroglutamate	0.0000000	11.937847
M110	M110	Serotonin	6.4285714	18.625635
M115	M115	Trigonelline	2.1428571	2.975737
M116	M116	Trimethylamine	0.7142857	13.021962
M118	M118	Tropate	0.7142857	12.596865
M119	M119	Tryptophan	0.0000000	13.981804
M120	M120	Tyrosine	3.5714286	12.001142
M122	M122	Valine	2.1428571	19.146486
M126	M126	trans-Aconitate	2.8571429	3.813505
M129	M129	u11	0.0000000	4.348629
M130	M130	u1125	7.1428571	5.359377
M134	M134	u144	1.4285714	15.536401
M136	M136	u185	22.8571429	6.836917
M137	M137	u217	2.8571429	5.005221
M138	M138	u233	0.7142857	4.735334
M142	M142	u43	0.7142857	7.151166
M144	M144	u87	0.0000000	6.635486
M148	M148	-Methylhistidine	1.4285714	16.561921
M149	M149	-Methylhistidine	0.0000000	8.351801

En primer lugar, se ha realizado un análisis univariante de las variables seleccionadas. Para cada variable,

se va extraer su el valor mínimo, el máximo, la mediana, la media, el valor al 25% (1º Cuartil) y al 75% (3º Cuartil). Además, para estudiar la significancia de cada una de las variables, se ha realizado un análisis de la varianza para cada una de las variables, donde como hipótesis nula se asume que la diferencia entre valores con respecto a la clase de la muestra, es debida al azar. El alfa tomada es del de 0.05. A continuación, se muestra una tabla resumen con todas las variables y sus valores.

### *Análisis univariante*

Table 2: Análisis univariante metabolitos

ID	Mínimo	1º Cuartil	Mediana	Media	3º Cuartil	Maximo	p-Valor
M4	0.1	18.575	34.6	43.122	51.15	242.5	0.00945
M5	1.3	67	162.2	232.954	255.25	2503	0.32255
M7	4.6	19.65	39.6	72.859	63.2	492.6	0.00205
M8	9.3	37.175	51	67.809	77.825	525	0.07107
M11	0.3	50.2	90	195.142	181.35	2676.3	0.32013
M14	0.2	30.375	50.35	68.958	83.85	437.6	0.02246
M15	7.9	28.8	44.1	57.157	71.075	212.3	0.02544
M25	0.4	12.2	18.25	24.225	28.5	171.8	0.13047
M26	2	13.8	20.9	28.098	30.8	374.6	0.13236
M31	0.2	13.6	24.15	70.886	62.2	1336.6	0.37244
M32	0.6	84.25	115.95	174.486	218.125	874.2	0.00105
M33	26.8	213.575	308.45	351.807	424.4	1127.6	0.40109
M36	0.1	13.375	21.65	44.716	59.425	305.9	0.03619
M37	4.3	31.375	49.05	85.503	105.65	1026.5	0.16544
M45	49.9	822.45	1917.1	2927.834	3769.15	16673.9	0.00001
M48	988	5791.9	7963.95	9989.246	12398.35	33766.6	0.08739
M50	0.2	98.075	150.45	354.954	270.475	17082.2	0.50306
M51	2.1	200.925	382.417	464.299	527.6	5732.7	0.13691
M65	25.6	224.775	348.6	533.42	702.35	2432.1	0.01217
M66	34.6	462.775	1112.05	1745.721	2380.1	16544.5	0.00699
M68	0.1	85.825	157.1	206.785	237.525	1639	0.28224
M71	4.4	23	31.35	47.874	56.4	490.4	0.29200
M73	10.1	46.2	90.75	98.433	133.05	366.3	0.04030
M74	0.1	10.375	16.15	27.512	32.875	171	0.15104
M75	21.6	78.4	105.7	165.529	193.225	1225.3	0.06626
M88	0.5	104.025	149.85	187.431	223.95	798.8	0.34613
M89	47.4	276.3	425.3	710.082	891.925	6317.1	0.00043
M90	2.8	32.75	56	108.087	140.725	748.7	0.02366
M91	4.6	38.45	50.25	71	91.3	337.5	0.02101
M93	5.6	39.2	65.15	80.09	97.125	505.6	0.25159
M101	0.1	18.517	29.3	43.154	42.7	428.8	0.53669
M104	0.2	186.05	249.25	323.118	384.625	1579.2	0.14342
M105	0.1	39.742	67.25	171.516	176.6	2182.2	0.20022
M106	5.3	31.6	50.4	72.849	87.65	598.7	0.09428
M107	0.4	184.6	254.15	356.076	441.575	2073.8	0.25254
M110	2.3	37.275	68.044	80.272	118.35	359.1	0.00318
M115	5.8	43.225	107.65	197.049	283.675	2134.5	0.09990
M116	0.5	13.275	27.25	36.23	39.075	318.7	0.43381
M118	4.6	73.95	179.3	248.626	269.8	1434.2	0.00074
M119	10.3	44.325	93.7	101.496	130.125	526.8	0.11256
M120	0.2	39.29	76.1	93.824	115.55	934.9	0.80299
M122	1	13.3	19.7	26.551	32.95	202.2	0.29747
M126	1.9	22.475	36.75	48.467	55.025	609.4	0.01346

ID	Mínimo	1º Cuartil	Mediana	Media	3º Cuartil	Maximo	p-Valor
M129	133.3	922.1	1770.3	1825.71	2384.05	8038.2	0.08761
M130	0.2	19.042	38	62.615	61.2	1188.7	0.47751
M134	64.6	452.45	734.372	1246.64	1299.9	8567.8	0.00016
M136	0.1	21.55	98.95	224.768	161.5	5014.9	0.60367
M137	0.4	261.1	604.5	1273.541	1672.85	12900.8	0.00395
M138	2.1	149.8	459.6	950.467	1335.4	6344.9	0.00001
M142	0.1	4.575	10.8	22.173	24.05	282.9	0.15833
M144	17	25.5	27.6	41.966	29.8	1251.4	0.41635
M148	1	80.2	196.55	326.404	535.65	2560.3	0.00018
M149	22.1	109.625	175.35	179.733	223.875	502.5	0.32976

Según el p-valor resultante del análisis de la varianza se puede observar:

- Aquellas variables que pueden ser significativas a la hora de la detección del cáncer gástrico son (p-valor  $< 0.05$ ): [M4, M7, M14, M15, M32, M36, M45, M65, M66, M73, M89, M90, M91, M110, M118, M126, M134, M137, M138, M148].
- Aquellas variables que pueden no ser significativas a la hora de la detección del cáncer gástrico son (p-valor  $\geq 0.05$ ): [M5, M8, M11, M25, M26, M31, M33, M37, M48, M50, M51, M68, M71, M74, M75, M88, M93, M101, M104, M105, M106, M107, M115, M116, M119, M120, M122, M129, M130, M136, M142, M144, M149].

Además, se puede observar, que los rangos de las variables son muy dispares unos de otros, por lo que será recomendable realizar una normalización cuando se hagan los análisis estadísticos.

### ***Matriz de correlaciones***

Después de normalizar las variables, se estudia su colinealidad. Se ha establecido el umbral 0.8 para decidir si unos metabolitos están estrechamente relacionados con otros.

Table 3: Matriz de alta correlación

Metabolito 1	Metabolito 2	Correlación
M88	M5	0.8115971
M88	M65	0.8139809
M90	M65	0.8215618
M88	M74	0.8152910
M104	M74	0.8068007
M122	M74	0.8256320
M5	M88	0.8115971
M65	M88	0.8139809
M74	M88	0.8152910
M90	M88	0.8281718
M104	M88	0.8656847
M107	M88	0.8679014
M138	M89	0.8139552
M65	M90	0.8215618
M88	M90	0.8281718
M74	M104	0.8068007
M88	M104	0.8656847
M88	M107	0.8679014
M74	M122	0.8256320

Metabolito 1	Metabolito 2	Correlación
M89	M138	0.8139552

Como se puede observar en la tabla, la variable M88 tiene un gran alto grado de correlación con 6 variables más, por lo que puede existir una gran dependencia entre ellas. Sería importante analizarlo con mayor profundidad para decidir si alguna de estos metabolitos debería ser eliminado en la generación de un modelo para el diagnóstico precoz de cáncer gástrico.

### *Análisis de componentes principales (PCA)*

Como se puede observar en la tabla, el primer componente sólo explica un 35.803. Para explicar el 70% se necesitarían al menos nueve componentes (71.157) y catorce componentes algo más del 80% (81.149).

Table 4: Componentes principales más representativos

PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14
0.358	0.071	0.062	0.051	0.042	0.037	0.035	0.029	0.025	0.024	0.022	0.02	0.017	0.017

Si se visualizan los dos primeros componentes se puede apreciar una pequeña diferenciación entre los grupos, quedando aquellos individuos con cáncer gástrico en la parte superior y aquellos con un tumor benigno y que están sanos más agrupados, aunque no se llega a distinguir una separación muy importante entre los grupos.

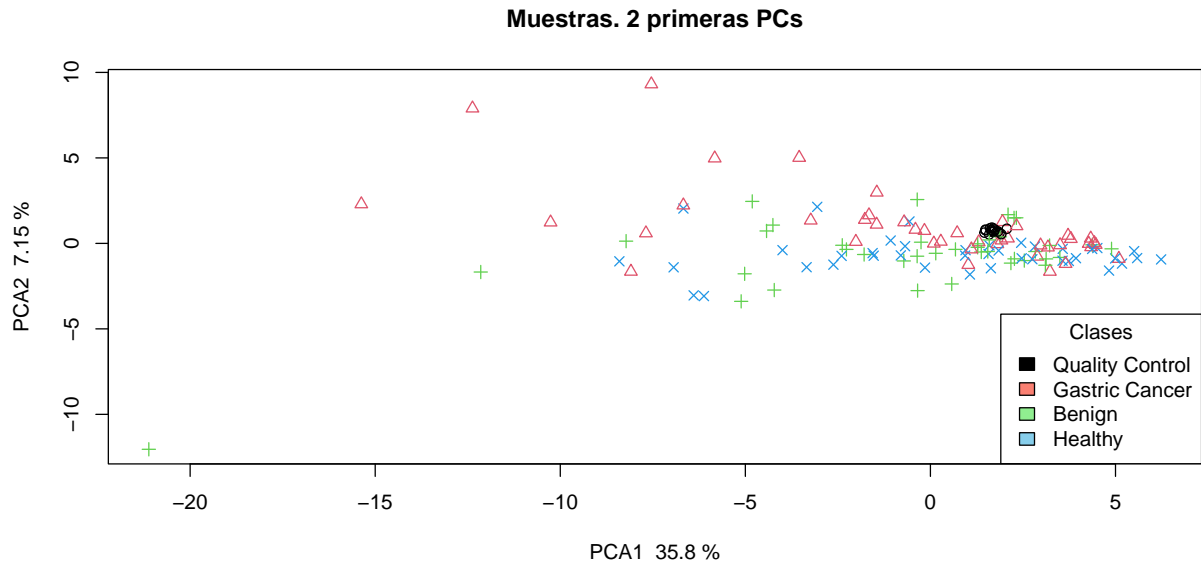


Figure 1: Muestras 2 primeras PCA's

Si se muestra el mapa de calor (Figura 2) donde las muestras se observan en el eje de las X y los metabolitos en los ejes de la Y, se puede observar que no hay una gran diferencia entre los mismos para poder agruparlos según su clase.

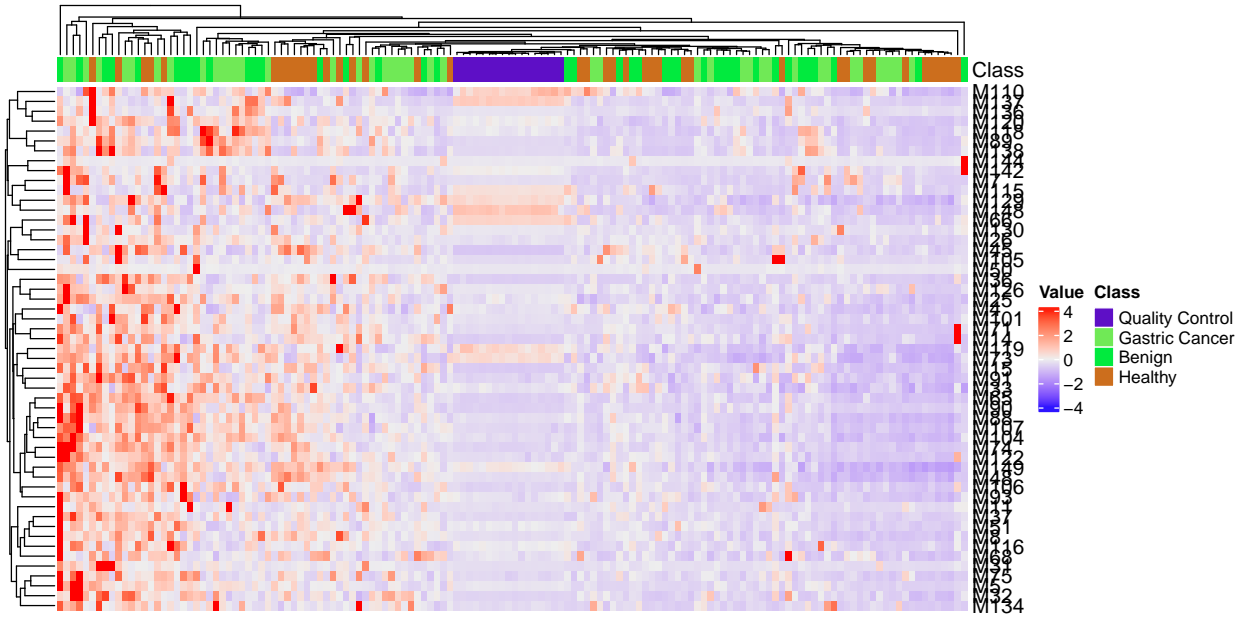


Figure 2: Mapa de calor

Si en un diagrama de cajas, se agrupan las muestras por clase, se puede ver que, excepto en el grupo de control, no hay un patrón entre las diferentes muestras, por lo tanto se podría decir que no se produce un *efecto batch*. Tampoco se puede ver un patrón claro que diferencie un grupo de otro.

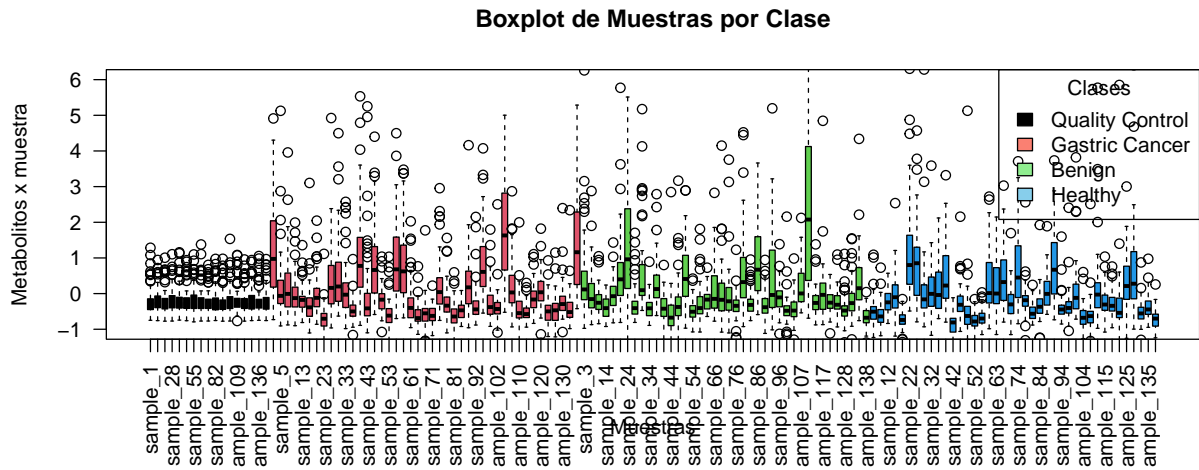


Figure 3: BoxPlot por muestra



En la figura 4 se puede ver los intervalos de cada una de las variables en función de la clase a la que pertenecen.

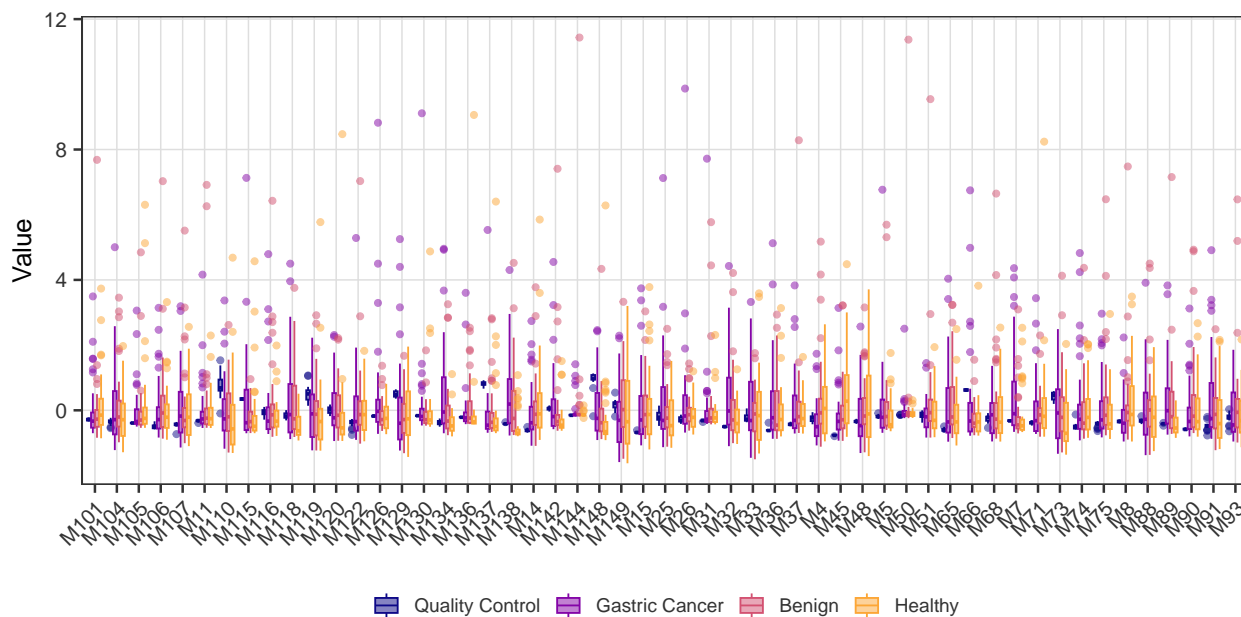


Figure 4: BoxPlot por metabolito

### *Análisis multivariante con variables significativas*

A continuación, se procede a realizar un análisis con aquellas variables que, en base al p-valor, han resultado ser significativas. Al reducir el número variable a analizar, la correlación entre las variables es bastante menor, habiendo sólo cuatro variables correlacionadas. Para evitar la multicolinealidad entre las variables, habría que estudiar si es necesario eliminar alguna de ellas para no distorsionar el modelo.

Table 5: Matriz de alta correlación

Metabolito 1	Metabolito 2	Correlación
M90	M65	0.8215618
M138	M89	0.8139552
M65	M90	0.8215618
M89	M138	0.8139552

Al realizar el análisis de componentes principales, el primer componente sólo explica un 33.824. Para explicar el 70% se necesitarían al menos seis componentes (70.48) y nueve componentes algo más del 80% (82.352). Para reducir la complejidad de modelo, se podrían seleccionar sólo las variables significativas, ya que se reduce la dimensionalidad.

Table 6: Componentes principales más representativos

PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
0.338	0.114	0.078	0.07	0.055	0.051	0.048	0.036	0.035

En esta ocasión, al analizar los dos primeros componentes, en la figura 5 no se ve una distinción tan clara entre los grupos, por lo que al reducir la dimensionalidad del modelo, es posible que no se pueda realizar un diagnóstico tan fiable.

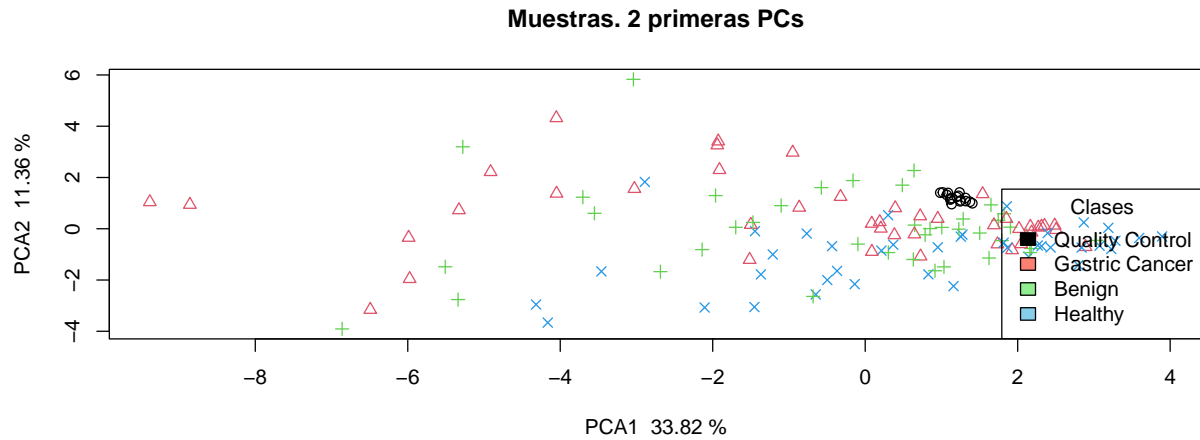


Figure 5: Muestras 2 primeras PCA's

Al igual que en el mapa de calor que se muestra en la (@ref(fig:g2)), se puede comprobar que al realizar el análisis únicamente con las variables más significativas tampoco hay una agrupación significativa en función de las clase de los pacientes.

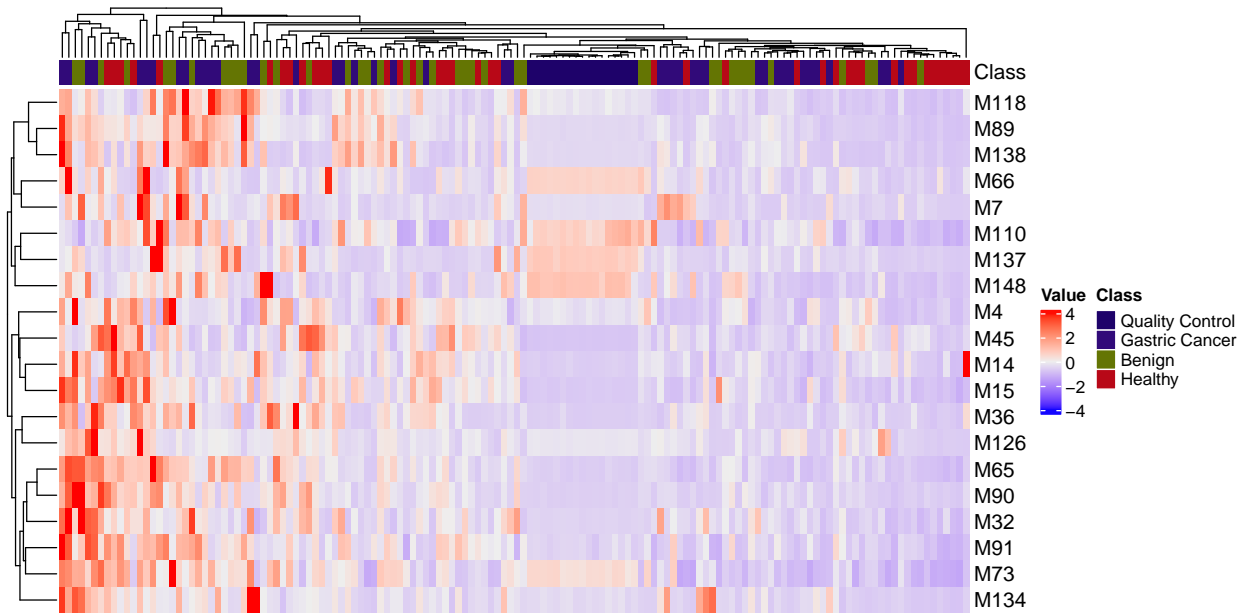


Figure 6: Mapa de calor

Si en un diagrama de cajas, se agrupan las muestras por clase, se puede ver que, excepto en el grupo de control, no hay un patrón entre las diferentes muestras, por lo tanto se podría decir que no se ve un efecto batch. Tampoco se puede ver un patrón claro que diferencie un grupo de otro.

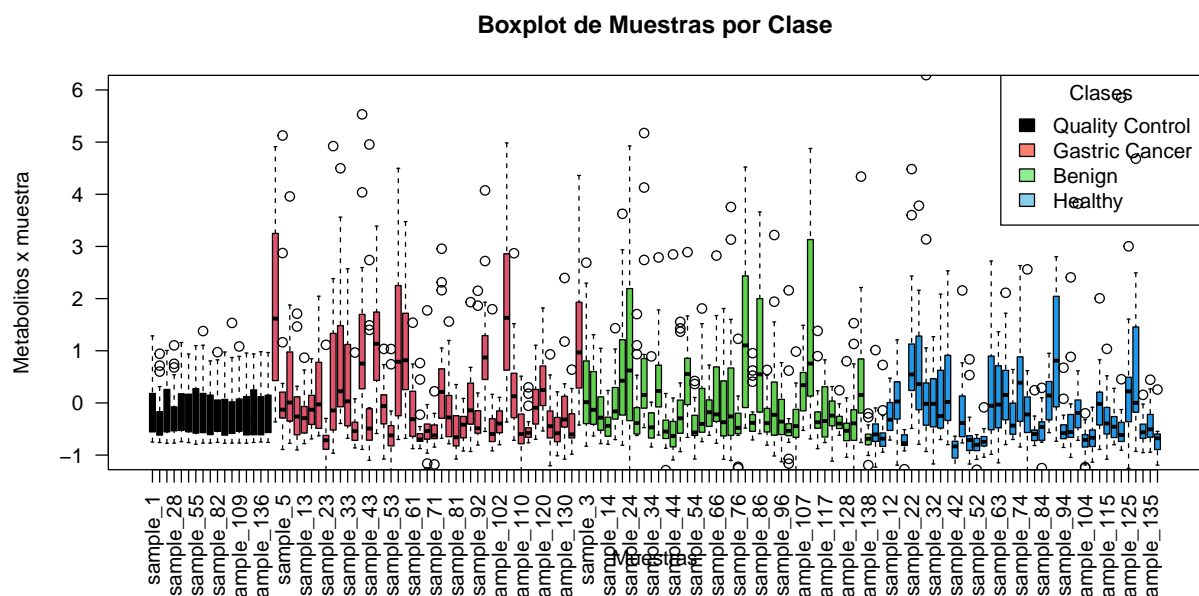


Figure 7: BoxPlot por muestra

En la siguiente figura (Figura 8), se puede ver los intervalos de cada una de las variables en función de la clase a la que pertenecen.

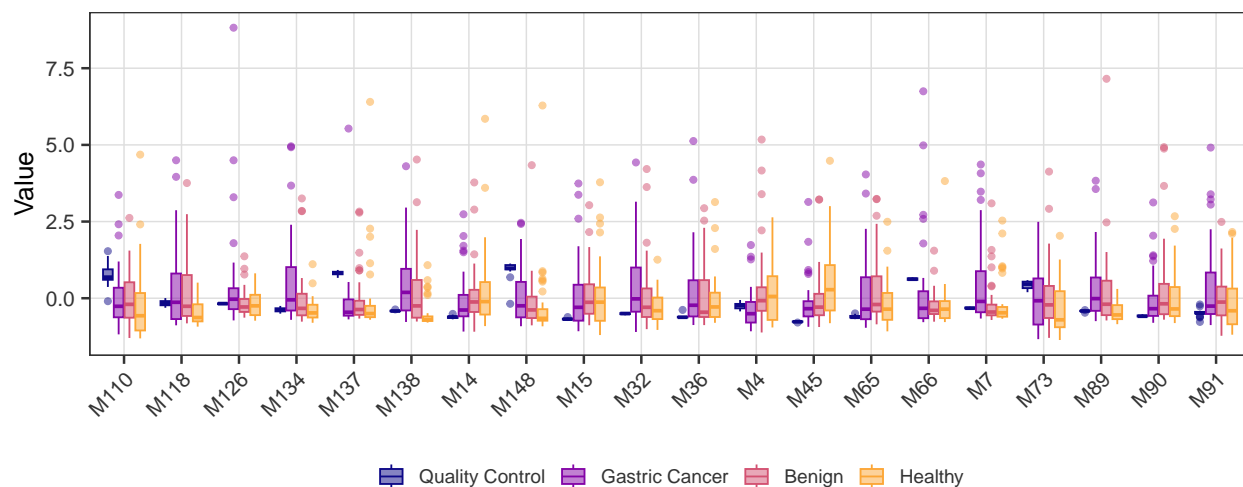


Figure 8: BoxPlot por metabolito

## 5 Discusión, limitaciones y conclusiones del estudio

Los datos del presente estudio se basan en el artículo “<sup>1</sup>H-NMR urinary metabolomic profiling for diagnosis of gastric cancer” de (Chan et al., 2015) donde se refleja que los pacientes con cáncer gástrico tienen un perfil urinario metabólico diferente y que se podría crear un perfil metabólico para el diagnóstico precoz de cáncer gástrico.

En el presente estudio, se ha realizado un análisis exploratorio de los diferentes metabolitos para identificar si los valores que toman son diferentes cuando el paciente padece cáncer gástrico y si es posible su diagnóstico

en base a un perfil urinario metabolómico. En primer lugar, se han eliminado aquellas variables que no tenían la calidad adecuada, habiéndose reducido el conjunto de metabolitos a 52 de los 149 originales. Dentro del análisis univariante, se ha realizado un análisis de la varianza de cada una de las variables y se han obtenido 20 metabolitos significativos.

Posteriormente, se ha realizado un análisis multivariante basado en el análisis de componentes principales para ambos grupos y no se han visto resultados concluyentes. Con ambos grupos de variables se ha observado que se necesita un número mayor de 6 componentes para poder explicar al menos el 70% del comportamiento, por lo que la dimensionalidad sigue siendo bastante compleja.

Tanto en los mapas de calor como en los boxplot, tampoco se aprecian agrupaciones claras dentro de las diferentes muestras que permitan identificar el cáncer gástrico. Ni siquiera se pueden apreciar que algunas variables tomen valores diferentes si la clase es cáncer gástrico o no.

Se puede llegar a la conclusión que el análisis exploratorio realizado no es suficiente y que sería necesario estudiar más modelos que ayuden a realizar el diagnóstico precoz de cáncer gástrico.

## 6 Notas

Todo el código y los datos utilizados en la presente práctica se encuentran en el siguiente repositorio de Github: <https://github.com/VegaUOC/Rodrigalvarez-Chamarro-MariadelaVega-PEC1>

## References

- Broadhurst, D., Goodacre, R., Reinke, S. N., Kuligowski, J., Wilson, I. D., Lewis, M. R., & Dunn, W. B. (2018). Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. *Metabolomics*, 14(6), 72. <https://doi.org/10.1007/s11306-018-1367-3>
- Chan, A. W., Mercier, P., Schiller, D., Bailey, R., Robbins, S., Eurich, D. T., Sawyer, M. B., & Broadhurst, D. (2015). (1)H-NMR urinary metabolomic profiling for diagnosis of gastric cancer. *Br J Cancer*, 114(1), 59–62.