UNIVERSIDAD DE SANTIAGO DE CHILE
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA INFORMÁTICA

Laboratorio N°1

Analisis Estadístico

Integrantes:	Felipe González
	Carlos Pérez
Curso:	Analisis de Datos
Sección	0-A-1
Profesor(a):	Dr. Max Chacon Pacheco

7 de Noviembre de 2020

Tabla de contenidos

Índice de cuadros

Índice de figuras

1. Introducción

El proceso normal de todas las células finaliza con su muerte, donde una nueva la reemplaza. Sin embargo, esto no sucede todo el tiempo, cuando una célula envejece y no muere o crecen células nuevas no necesarias, se empiezan a acumular y a dividir sin parar. Esto llega al punto de, en la mayoría de los casos, generar una masa sólida llamada Tumor.

Existen dos tipos de tumores. Los benignos, también llamados no cancerosos, se caracterizan por no extenderse a los tejidos cercanos, es decir, no invaden otras zonas de cuerpo. Aún por más benignos se llamen, la masa puede crecer al punto de alterar o interrumpir funciones corporales, teniendo hasta consecuencias mortales.

En su contra parte, están los tumores malignos, también llamados cancerosos o simplemente cáncer. Además de extenderse por los tejidos cercanos, células cancerosas pueden desprenderse del tumor y viajar por el sistema circulatorio o linfático y generar nuevos tumores en lugares totalmente diferentes. A esto último se le llama metástasis.

Cuando un cáncer produce metástasis en otra parte del cuerpo, el nuevo tumor posee las características de su origen. Por ejemplo, cuando un cáncer de mama se disemina a los pulmones, el nuevo tumor que se genera tiene las características de un cáncer de mama, e incluso, se le llama cáncer metastásico de mama y no cáncer de pulmón.

Existen más de 100 tipos de cáncer, y como se puede intuir del párrafo anterior, su nombre depende del lugar donde se origina. Es importante diferenciar uno de otro, porque crecen y se diseminan a velocidades distintas, además, cada uno tiene tumores de diferentes características, y por sobre todo, responden de distinta manera a cada tratamiento.

En específico, en este informe se estudia el cáncer de mama. Usualmente, el tumor puede ser visto a través de un examen de rayos x y se puede sentir. Si bien, se origina en el seno, no todo cáncer originado en la zona corresponde a cáncer de mama, pues puede corresponder a un linfoma o sarcoma.

Es crucial estudiar este tipo de cáncer, puesto que "en las mujeres, el cáncer de mama es el segundo tipo de cáncer más común"(NIC, 2020). Además, posee una mortalidad del 10 %.

2. Descripción del Problema

Ahora que se conoce mas a fondo la naturaleza del contexto del problema, resulta de gran importancia también, conocer el problema en términos de los factores que están involucrados en su descripción. Vale decir, tanto la procedencia de dichos factores, como la descripción respectiva de cada uno y qué buscan representar dentro del problema en cuestión.

En la presente sección se describe a fondo la base de datos que se utiliza en este laboratorio, junto con la descripción de cada uno de sus atributos y el dominio de sus valores.

2.1. Base de Datos

La base de datos que se utiliza proviene de uno de los tantos repositorios que ofrece el sitio web del *UCI Machine Learning Repository*, de los cuales el que se utiliza en este laboratorio es el que lleva como titulo *Wisconsin Breast Cancer Database*. Ahora bien, la procedencia real de esta base de datos se atribuye a los Hospitales de la Universidad de Wisconsin, Madison por el Doctor y Físico William H. Wolberg, donada el 15 de Julio de 1992, por lo que la composición de esta base de datos se limita a la recopilación hasta esa fecha.

Grupo	Instancias	Fecha
1	367	Enero de 1989
2	70	Octubre de 1989
3	31	Febrero de 1990
4	17	Abril de 1990
5	48	Agosto de 1990
6	49	Enero de 1991
7	31	Junio de 1991
8	86	Noviembre de 1991

Cuadro 1: Distribución de Instancias en el Tiempo

2.2. Clases y Variables

2.2.1. Descripción

Los datos corresponden a muestras de líquido tumoral de pacientes con protuberancias sólidas en su pecho. Se presentan 11 variables, todas cuantitativas discretas, y sus valores oscilan entre 1 y 10. Dos de estas variables, *code* y *class*, se escapan de esas características.

- *code:* Corresponde al identificador de cada dato. No tiene peso a la hora de realizar un estudio.
- *clumpThickness:* Clump Thickness. Evalúa si las células son mono o multicapa. Las células benignas tienden a agruparse en monocapas, mientras que las células cancerosas a menudo se agrupan en multicapas.
- *unifCellSize:* Uniformity of Cell Size. Evalúa la consistencia del tamaño de las células de la muestra. Las células cancerosas suelen variar en tamaño, mientras que las benignas no.
- *unifCellShape:* Uniformity of Cell Shape. Evalúa la consistencia del forma de las células de la muestra. Las células cancerosas suelen variar en forma, mientras que las benignas no.
- *marginalAdhesion:* Marginal Adhesion. Cuantifica la proporción de células que se mantienen adheridas. Las células no cancerosas tienden a permanecer juntas, mientras que las malignas pierden esa capacidad.
- *epithCellSize:* Single Epithelial Cell Size. Mide el agrandamiento del tamaño de las células epiteliales. El tejido epitelial es uno de los cuatro tejidos básicos animal y está formado por una o varias capas de células epiteliales. Se encuentra en la parte externa de la piel y en la superficie interna de los órganos, vasos y todas las cavidades del cuerpo humano. Es importante considerar este dato, porque "la mayoría de los cánceres de mama son carcinomas, que son tumores que se originan de las células epiteliales que revisten los órganos y los tejidos que se encuentran en todo el cuerpo. Cuando los

carcinomas se forman en el seno, por lo general son de un tipo más específico llamado adenocarcinoma, que comienza en las células de los conductos (los conductos de la leche) o los lobulillos (glándulas productoras de leche)."(ACS, 2020). Las célúlas epiteliales cancerosas sueles estirarse significativamente.

- *bareNuclei* Bare Nuclei. Proporción de núcleos rodeados de citoplasma versus aquellos que no.
- *blandChromatin* Bland Chromatin. Califica la textura uniforme del núcleo en un rango de fino a grueso. Las células cancerosas suelen estar dentro de los gruesos
- *normalNucleoli* Normal Nucleoli. Determina si los nucleolos son pequeños y apenas visibles o más grandes, más visibles y más abundantes. Las células benignas se caracterizan por tener un nucleolo mas bien pequeño.
- *mitoses* Mitoses. Describe el nivel de actividad mitótica.
- *class:* Class. Describe si el tumor es benigno (2) o canceroso (4).

2.2.2. Dominio de Atributos

Atributo	Tipo	Dominio
Sample code number	Numérico	id
Clump Thickness	Numérico	1 - 10
Uniformity of Cell Size	Numérico	1 - 10
Uniformity of Cell Shape	Numérico	1 - 10
Marginal Adhesion	Numérico	1 - 10
Single Epithelial Cell Size	Numérico	1 - 10
Bare Nuclei	Numérico	1 - 10
Bland Chromatin	Numérico	1 - 10
Normal Nucleoli	Numérico	1 - 10
Class	Nominal	2 , 4

Cuadro 2: Dominio de Atributos

2.2.3. Escala de Atributos

Atributo	Valor	Significado
Clump Thickness	1	Las células son completamente mono-capa
	2	Las células son 90 % mono-capa
	3	Las células son 80 % mono-capa
	4	Las células son 65 % mono-capa
	5	Ligeramente más mono-capa que multi-capa
	6	Ligeramente más multi-capa que mono-capa
	7	Las células son 35 % mono-capa
	8	Las células son 20 % mono-capa
	9	Las células son 10 % mono-capa
	10	Las células son multi-capa
Uniformity of Cell Size	1	Las células son completamente uniformes
	2	Las células son 90 % uniformes
	3	Las células son 80 % uniformes
	4	Las células son 65 % uniformes
	5	Las células son mas del 50 % uniformes
	6	Las células son denos del 50 % uniformes
	7	Las células son 35 % uniformes
	8	Las células son 20 % uniformes
	9	Las células son 10 % uniformes
	10	Las células son inconsistentes con su uniformidad

Cuadro 3: Escala Clump Thickness y Uniformity of Cell Size

Atributo	Valor	Significado
Uniformity of Cell Shape	1	Completamente uniformes
	2	Las células son 90 % uniformes
	3	Las células son 80 % uniformes
	4	Las células son 65 % uniformes
	5	Las células son mas del 50 % uniformes
	6	Las células son denos del 50 % uniformes
	7	Las células son 35 % uniformes
	8	Las células son 20 % uniformes
	9	Las células son 10 % uniformes
	10	Las células son inconsistentes con su uniformidad
Marginal Adhesion	1	Todas permanecen juntas
	2	El 90 % permanecen juntas
	3	El 80 % permanecen juntas
	4	El 70 % permanecen juntas
	5	El 60 % permanecen juntas
	6	El 50 % permanecen juntas
	7	El 40 % permanecen juntas
	8	El 30 % permanecen juntas
	9	El 20 % permanecen juntas
	10	Las células no exhiben adhesion marginal

Cuadro 4: Escala Uniformity of Cell Shape y Marginal Adhesion

Atributo	Valor	Significado
Single Epithelial Cell Size	1	No hay células significativamente agrandadas
	2	Las células más grandes son 20 % más grandes
	3	Las células más grandes son 30 % más grandes
	4	Las células más grandes son 40 % más grandes
	5	Las células más grandes son 50 % más grandes
	6	Las células más grandes son 60 % más grandes
	7	Las células más grandes son 70 % más grandes
	8	Las células más grandes son 80 % más grandes
	9	Las células más grandes son 90 % más grandes
	10	Las células más grandes son 100 % más grandes
Bare Nuclei	1	Nucleos completamente desprovistos de citoplasma
	2	20 % de los núcleos tienen citoplasma
	3	30 % de los núcleos tienen citoplasma
	4	40 % de los núcleos tienen citoplasma
	5	50 % de los núcleos tienen citoplasma
	6	60 % de los núcleos tienen citoplasma
	7	70 % de los núcleos tienen citoplasma
	8	80 % de los núcleos tienen citoplasma
	9	90 % de los núcleos tienen citoplasma
	10	Todos los núcleos tienen citoplasma

Cuadro 5: Escala Single Epithelial Cell Size y Bare Nuclei

Atributo	Valor	Significado
Bland Chromatin	1	Cromatina de textura completamente fina
	2	20 % de la cromatina es gruesa
	3	30 % de la cromatina es gruesa
	4	40 % de la cromatina es gruesa
	5	50 % de la cromatina es gruesa
	6	60 % de la cromatina es gruesa
	7	70 % de la cromatina es gruesa
	8	80 % de la cromatina es gruesa
	9	90 % de la cromatina es gruesa
	10	La cromatina es completamente gruesa
Normal Nucleoli	1	Los nucléolos son completamente normales (pequeños, uno por celda, apenas visibles)
	2	20 % de los nucleolos son anormales
	3	30 % de los nucleolos son anormales
	4	40 % de los nucleolos son anormales
	5	50 % de los nucleolos son anormales
	6	60 % de los nucleolos son anormales
	7	70 % de los nucleolos son anormales
	8	80 % de los nucleolos son anormales
	9	90 % de los nucleolos son anormales
	10	100 % de los nucleolos son anormales

Cuadro 6: Escala Bland Chromatin y Normal Nucleoli

Atributo	Valor	Significado
Mitoses	1	La actividad mitótica es completamente normal
	2	20 % de la actividad mitótica parece anormal
	3	30 % de la actividad mitótica parece anormal
	4	40 % de la actividad mitótica parece anormal
	5	50 % de la actividad mitótica parece anormal
	6	60 % de la actividad mitótica parece anormal
	7	70 % de la actividad mitótica parece anormal
	8	80 % de la actividad mitótica parece anormal
	9	90 % de la actividad mitótica parece anormal
	10	100 % de la actividad mitótica parece anormal
Class	2	Benigno
	4	Maligno o Canceroso

Cuadro 7: Escala Mitoses y Class

Un punto que hay que tomar en consideración antes de poder pasar a manipular los datos para realizar los análisis respectivos de la siguiente sección, es que los datos presentan 16 *Missing Values*, denotados por el símbolo "?", específicamente para 16 instancias del atributo *Bare Nuclei* entre los grupos 1 y 6. Finalmente la distribución de la clase para la totalidad de los datos se muestra en la siguiente tabla.

Distribución de la Clase	Beningno: 458 (65.5 %) Maligno: 241 (34.5 %)
Numero de *Missing Values*	16
Numero de Instancias	699

Cuadro 8: Distribución de la Clase

3. Análisis Estadístico

Ahora que se ha logrado un correcto entendimiento de cada una de las variables y clases implicadas en el problema, es posible proceder a aplicar determinadas técnicas de análisis estadístico para ahondar aun mas en este y descubrir relaciones existentes entre distintas variables, y el significado que dichas relaciones significan en el problema en cuestión.

Antes de comenzar a trabajar directamente con los datos es necesario realizar algún tipo de transformación en los datos, que permitan de alguna forma tratar los *Missing Values* mencionados al final de la sección anterior. Ante esta situación surgen dos posibles opciones.

- **Opción 1:** La primera opción contempla realizar una **Imputación**, que básicamente consiste en sustituir los valores no contemplados, en este caso los 16 valores "?" de la columna *Bare Nuclei*, por alguna otra medida como la **media** o la **mediana**.
- **Opción 2:** La segunda opción contempla simplemente excluir o eliminar las observaciones que en la columna *Bare Nuclei* tengan como valor el caracter "?".

Analizando ambas opciones, la opción numero 1 tiene factibilidad dependiendo de la naturaleza del problema y en que efecto tiene sobre el mismo. En este sentido, el atributo *Bare Nuclei* tiene directa relación con lo que es el tamaño del núcleo celular, por lo que cabe preguntarse si, el tamaño del núcleo de una célula individual tiene algún tipo de relación con la media del tamaño del núcleo de las otras células. La verdad a simple vista pareciese que no debería porque existir algún tipo de relación.

La segunda opción no parece tan mala, ya que si consideramos la totalidad de observaciones que en este caso son originalmente 669, y las 16 que presentan un *Missing Value*, estos últimos contemplan un 2,4 % de los datos, por lo que prescindir de estos no debería afectar tanto los análisis posteriores. Finalmente la opción escogida es la segunda, con lo que la distribución de la clase para la totalidad de los datos se muestra a continuación.

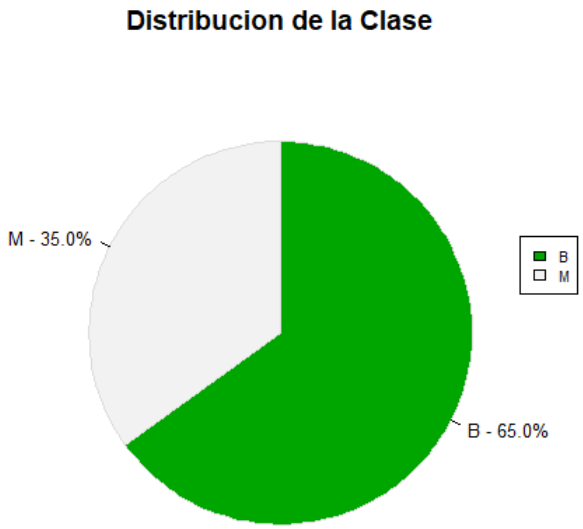

Figura 1: Pie Chart Distribución de Clase

Donde el numero de observaciones se reduce a 683, con **M = Maligno o Canceroso** indica la presencia de células cancerígenas sumando un total de 239 observaciones correspondiente al 35 % del total, y **B = Benigno** indica la ausencia de células cancerígenas sumando un total de 444 observaciones correspondiente al 65 % restante.

Ahora que los *Missing Values* han sido eliminados es posible proceder con el análisis estadístico, y de esta forma inferir aun mas sobre el conjunto de datos y que fenómenos pueden describir.

3.1. Gráficos de Caja

Para el primero de los análisis se emplean los denominados *Box Plot* o mas conocidos como "Gráficos de Caja". Estos tienen la particularidad que permiten conocer la distribución del conjunto de datos a partir de 5 medidas principales, **Mínimo**, **Primer Cuartil (Q1 o también Percentil 25)**, **Segundo Cuartil (Q2 o también Percentil 50)** que a su vez corresponde a la **mediana**, **Tercer Cuartil (Q3 o también Percentil**

75), y por último el **Máximo**. Otra medida muy importante que nos explican estos gráficos corresponden a los denominados *outliers*, que corresponden a aquellos valores que están fuera de los rangos del Mínimo y el Máximo, y explican a aquellos valores totalmente anómalos a la distribución.

A continuación, se presentan gráficos de caja de todas las variables para estudiar la dispersión de los datos y acercarnos a su relación con el tipo de tumor.

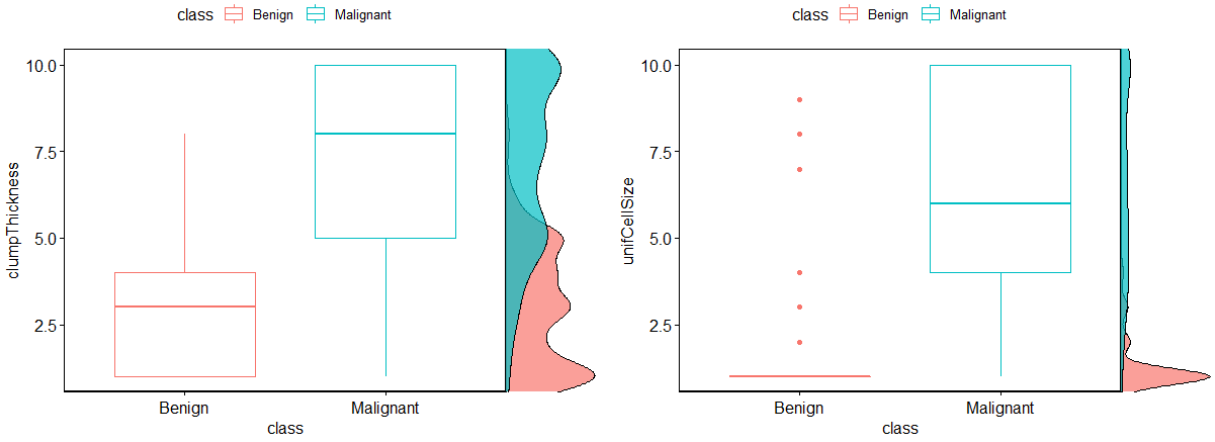

Figura 2: Gráficos de caja ClumpThickness y UnifCellSize

- *clumpThickness:* A partir de lo que visualmente ofrece el gráfico, se podría decir que la presencia de células cancerosas, esta mas relacionada con valores superiores a 5 al menos para esta variable, cuya mediana se encuentra alrededor del valor 8, lo que siguiendo la escala planteada anteriormente sugiere que alrededor del 20 % de las células son mono-capa. Por el contrario, aquellas observaciones que presentan ausencia de células cancerosas están mas relacionadas con valores menores a 5 para esta variable.

- *unifCellSize:* Lo mismo sucede con esta característica, donde se puede apreciar que para valores mayores a 5 las observaciones tienden a corresponder a la clase *Malignant*, y en caso contrario aquellas observaciones con esta característica menor a 5 corresponden a la clase *Benign*. Además es posible observar como la mayoría de los datos tienden a ubicarse en el valor 1, dejando 6 outliers que están fuera de este rango.

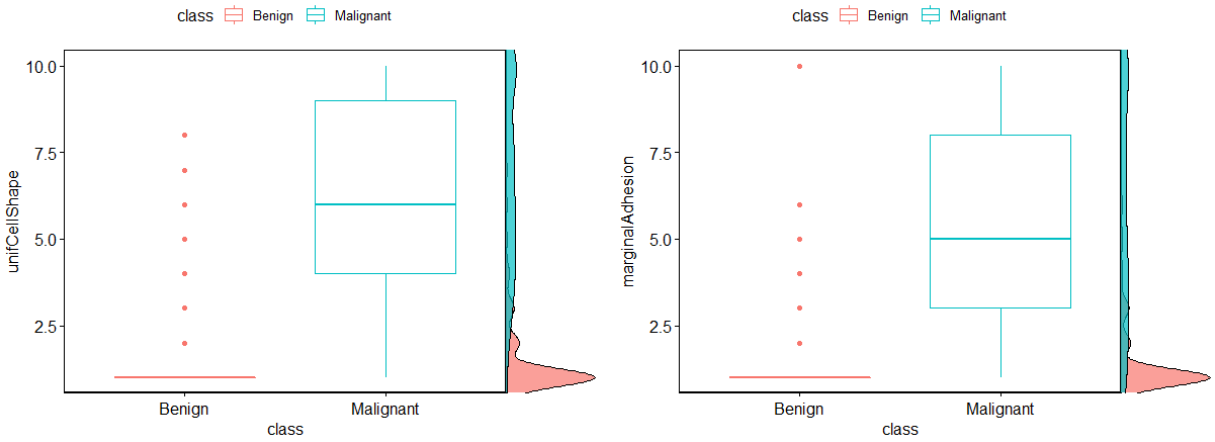

Figura 3: Gráficos de caja UnifCellShape y MarginalAdhesion

- *unifCellShape:* El gráfico de esta característica se ve bastante similar al de la característica anterior, tanto que si no se especifica la característica podrían confundirse, por lo que se podría llegar a suponer que existe una relación entre ambas variables

- *marginalAdhesion:* Recién en esta característica se puede notar con mayor claridad gráficamente, que la mediana se ubica cerca del valor 5, mientras que el cuartil 1 en el valor 2.5 y el cuartil 3 en el valor 7.5, por lo que se podría tender a creer que no existe una tendencia de asociar la clase *Malignant* con alguno de los valores de la escala. Sin embargo se puede apreciar que para la clase *Benign* la mayoría de los valores tienden a estar muy cercanos al valor 1, lo que teóricamente significa que la ausencia de células cancerosas tienden a relacionarse con una adhesión marginal fuerte, o dicho en otras palabras todas las células permanecen juntas.

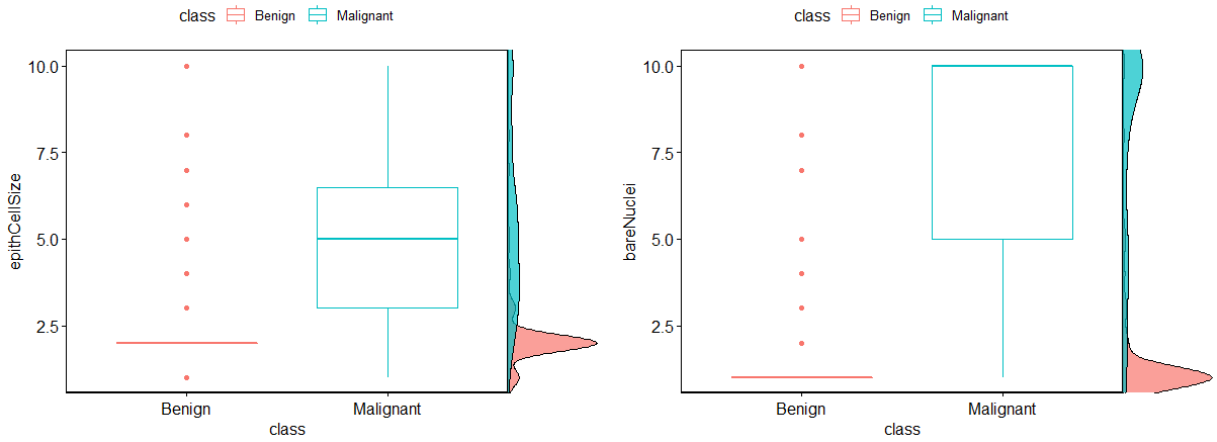

Figura 4: Gráficos de caja EpithCellSize y BareNuclei

- *epithCellSize:* Esta característica exhibe un comportamiento bastante similar al anterior, al menos visualmente viendo la distribución de la caja, y sus respectivas medidas, a excepción de que los valores para aquellas observaciones pertenecientes a la clase *Benign* están mas relacionadas al valor 2 y no el 1 como el caso anterior. Lo que denota una que aquellas células mas grandes son solo un 20 % aun mas grandes de lo normal.

- *bareNuclei:* En este caso, las observaciones pertenecientes a la clase *Malignant* tienden a presentar valores mayores a 5, sobre todo cercanas a 10 al menos para esta característica, denotando que arriba del 50 % de las células presentan citoplasma en el núcleo y aun mas cercanos al 100 %. En caso contrario aquellas observaciones denotadas como *Benign* aluden a un bajo porcentaje de células cuyo núcleo presenta citoplasma, cercano a la totalidad de estas sin presencia de este.

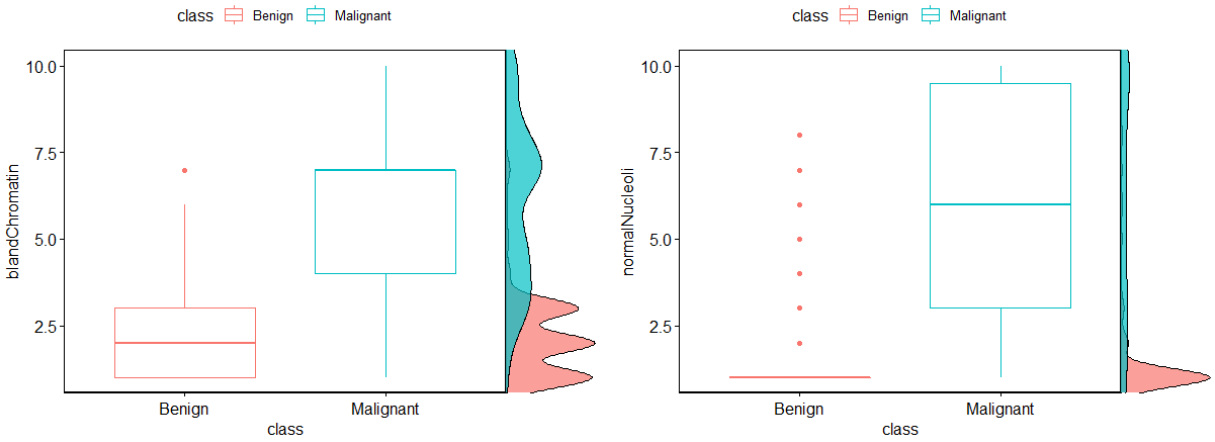

Figura 5: Gráficos de caja BlandChromatin y NormalNucleoli

- *blandChromatin:* En cuanto a las observaciones asignadas a la clase *Malignant* tienden a tener un valor superior a 5 en esta característica, sin embargo no cercana a 10 sino mas bien cercanas a 7.5, denotando que cerca del 70 % de la cromatina de las células es gruesa. Mientras que para las células de la clase *Benign*, lo es cercano a un 20 %.

- *normaNucleoli:* Para esta característica también es difícil al menos a simple vista asociarle algún valor al menos con la clase *Malignant*, sin embargo, existe una mayor tendencia aunque no muy grande a tener valores un poco mas grandes a 5. En cuanto a la clase *Benign* están prácticamente cercanas a 1.

Figura 6: Gráfico de caja Mitoses

- *mitoses:* Para esta última característica se presenta un comportamiento muy distinto a las demás, pues para las observaciones con la clase *Malignant*, una pequeña parte supera el valor 5, mientras que la mayoría se concentra entre los valores 1 y 4 aproximadamente, con su mediana cercana a 1. Por otro lado, las observaciones de la clase *Benign* tienden a estar en el valor 1 a excepción de unos pocos outliers.

En general, se observa que los datos de los tumores cancerosos son mas dispersos que el de los tumores benignos. Además, los valores de los tumores benignos están considerablemente cercanos a uno en comparación con los tumores cancerosos. Por lo que se puede asumir que para cada una de las características o atributos de las observaciones, valores cercanos a 1 o bien inferiores a 5 tienden a asociar a la observación con la clase de valor 2 o *Benign* indicando la posibilidad de ausencia de células cancerosas, de forma análoga los valores mas cercanos a 10 o bien mayores a 5 tienden a asociar a la observación con la clase de valor 4 o *Malignant* indicando la posibilidad de la presencia de células cancerosas. Esto a excepción de la característica *Mitoses* que como bien se pudo observar posee un comportamiento un tanto distinto a las demás.

Otra observación respecto a la distribución observada en la forma de los diagramas de caja observados es que en su gran mayoría no siguen una **Distribución Normal**, la cual se asimilaría a una Campana de Gauss, por lo tanto para pruebas de hipótesis es imposible emplear métodos perimétricos, razón por la cual en la siguiente sub-sección se emplea un método no perimétrico.

3.2. Correlación de Variables

Para evaluar la correlación entre las variables, primero se realizó un test Shapiro para verificar si las distribuciones son normales o no. Para complementar la información, se realizó un histograma y efectivamente no siguen una distribución normal. Por lo tanto se opta por un test no paramétrico.

Se realizó el test no paramétrico de Spearman para evaluar la correlación. En la siguiente tabla se muestran los resultados del índice de correlación de todas las variables de estudio con la variable *class*.

Variables	class
clumpThickness	0.7147899
unifCellSize	0.8208014
unifCellShape	0.8218909
marginalAdhesion	0.7062941
epithCellSize	0.6909582
bareNuclei	0.5087019
blandChromatin	0.7582276
normalNucleoli	0.7186772
mitoses	0.4234479
class	1.0000000

Cuadro 9: Correlación de cada variable con *class*

Todas las variables presentan un correlación positiva. Además, las variables que tienen un mayor índice de correlación con *class* son: *unifCellSize* y *unifCellShape*. A continuación se muestra la matriz de correlación.

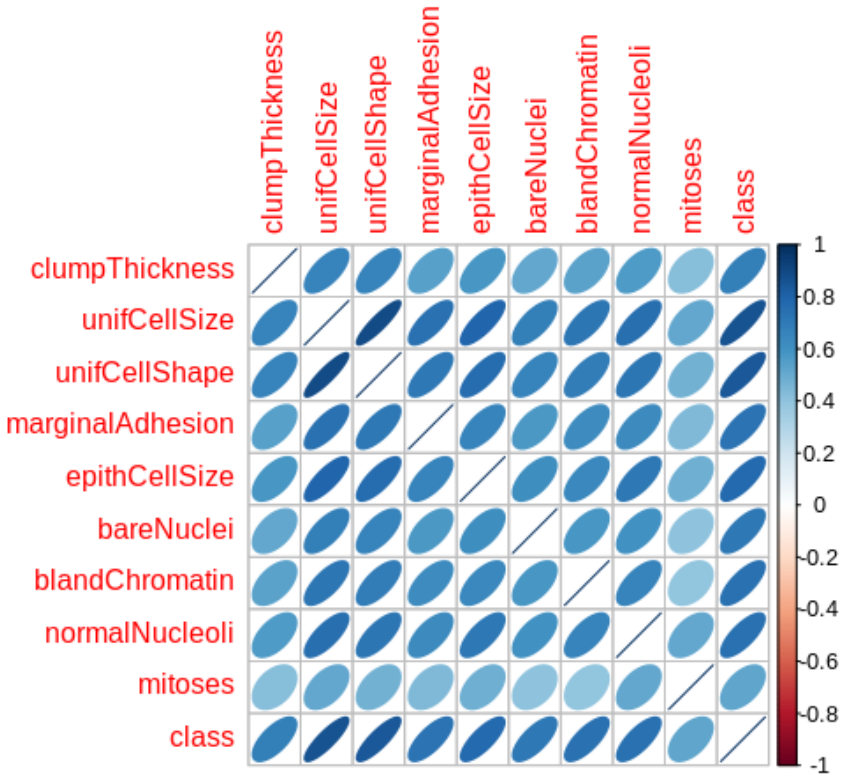

Figura 7: Gráfico correlaciones

Como se puede observar, todas son positivas. Esto se explica por la manera que se han planteado las variables. Finalmente, se asocia los valores cercanos a 10 de todas las variables a mayor riesgo de que el tumor del paciente sea canceroso.

4. Conclusiones

La base de datos utilizada presenta datos de fluidos extraídos de tumores encontrado en el pecho de pacientes. Además, contiene una variable llamada *class* que indica si el tumor es canceroso o no. Por esto, el estudio realizado se basó en identificar si las demás variables ayudan a pronosticar si el paciente padece de cáncer.

Las variables fueron planteadas de tal forma que, a mayor valor, presenta mayor probabilidad de que el tumor se canceroso. Esto se puede comprobar en la tabla de correlaciones y en los gráficos de caja. Por esto mismo, los índices de correlación resultaron ser todos positivos.

Además las conclusiones obtenidas a partir de esta primera experiencia significan una base de la que sera la siguiente experiencia en el *Agrupamiento de K-Medias*, para tal vez antes ver la posibilidad de excluir algunas variables que no son significativas en la representación de los datos y dejar el modelo aun mas simple.

A futuro, se espera poder realizar un método confiable que pueda pronosticar si un paciente efectivamente padece de cáncer de mama, a partir de una muestra del fluido de su tumor.

Bibliografía

Borges, L. (2015). Analysis of the wisconsin breast cancer dataset and machine learning for breast cancer detection.

Contreras, J. A. (2020). Laboratorio 1 - análisis estadístico. [Online] https://www.overleaf.com/read/khyrqwqtbbcq.

Institute, N. C. (2020a). Breast cancer—patient version. [Online] https://www.cancer.gov/espanol/tipos/seno.

Institute, N. C. (2020b). Cancer stat facts: Female breast cancer. [Online] https://seer.cancer.gov/statfacts/html/breast.html.

Institute, N. C. (2020c). ¿qué es el cáncer? [Online] https://www.cancer.gov/espanol/cancer/naturaleza/que-es.

MERZOUKI, R. (2017). User manualbreast cancer diagnosis web user interface. [Online] https://www.rai-light.com/docs/BCD_User_Manual_v01.pdf.

Salom, E. V. (2017). ¿qué son las células epiteliales? [Online] https://cienciatoday.com/que-son-celulas-epiteliales/.

Society, A. C. (2020a). Tipos de cáncer de seno. [Online] https://www.cancer.org/es/cancer/cancer-de-seno/comprension-de-un-diagnostico-de-cancer-de-seno/tipos-de-cancer-de-seno.html.

Society, A. C. (2020b). What is cancer? [Online] https://www.cancer.org/cancer/cancer-basics/what-is-cancer.html.

Society, A. C. (2020c). What is cancer? [Online] https://medlineplus.gov/spanish/cancer.html.

UCI (2020). Breast cancer wisconsin (original) data set. [Online] http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original).