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

Laboratorio I - Análisis estadístico

Integrantes: Maximiliano Araya Poblete
Miguel Salinas González

Curso: Análisis de Datos

Sección A-1

Profesor: Max Chacón Pacheco

23 de Abril de 2022

Tabla de contenidos

1. Introducción

La realización de esta primera experiencia se centra principalmente en los conocimientos extraíbles de la base de datos *Wisconsin Diagnostic Breast Cancer (WDBC)*, ésta se encuentra poblada por datos correspondientes a los años 1989, 1990 y 1991. Datos que fueron obtenidos de la página UCI Machine Learning Repository.
Dado el enunciado que plantea dicha base de datos como única a trabajar durante el largo del semestre. Es sumamente necesario el conocer primeramente las variables y clases que contiene, para así saber específicamente en qué se detalla cada una. Es con estos conocimientos que se podrá realizar un correcto análisis e incluso poder tomar decisiones en futuras experiencias.

El objetivo de esta experiencia es estudiar e interpretar los datos correspondientes a la base de datos *WDBC*. Adicionalmente se aplican técnicas de estadística descriptiva e inferencial según corresponda, tales como: Medidas de centralización (media, moda y mediana), distribución de probabilidades y medidas de dispersión (rango y varianza), tests de hipótesis, análisis de varianza, etc...

La estructura del documento consta de esta introducción, seguida de la descripción del problema que por un lado tiene la descripción de la base de datos y por otro lado tiene la descripción de las clases y variables. Posteriormente continúa con el análisis estadístico e inferencial. Finalmente terminando con las conclusiones respecto al problema y respecto al análisis del conjunto de datos. Además de contar con un apartado con las referencias correspondientes.

2. Marco Teórico

2.1. Correlación

Medida estadística que expresa hasta que punto dos variables están relacionadas linealmente. Es una herramienta de uso común para representar relaciones simples. En el análisis inferencial del presente trabajo se utiliza correlación de Pearson. Lalinde et al. (2018)

2.2. Tumor

Cuando células dañadas o anormales se multiplican sin control pueden formar bultos de tejido llamados tumores. Los tumores pueden ser malignos (Cancerosos) o benignos (no cancerosos). InstitutoNacionaldelCáncer (2021)

2.3. Cáncer

Enfermedad causada por células del cuerpo que se multiplican sin control y se diseminan a diferentes partes del cuerpo como tejidos cernamos, o bien pueden viajar lejos en el cuerpo y formar mas tumores en un proceso llamado metástasis. El cáncer puede comenzar en cualquier parte del cuerpo, en el presente documento se enfoca en el cáncer de mama. InstitutoNacionaldelCáncer (2021)

2.3.1. Metástasis

Presente cuando células cancerosas se desprenden del tumor original y viajan a través de la sangre o sistema linfático formando otro tumor en alguna otra parte del cuerpo. InstitutoNacionaldelCáncer (2021)

2.3.2. Ganglios linfáticos

Estructuras ovaladas distribuidas a lo largo de todo el cuerpo que permite la interacción entre antígenos y linfocitos. Ayudan a combatir infecciones al atacar gérmenes que son transportados por el liquido linfático. César E. Montalvo Arenas (2018)

2.3.3. Célula epitelial

Tipo de célula que recubre las superficies del cuerpo presentes en la piel, vasos sanguíneos, tracto urinario y órganos. César E. Montalvo Arenas (2018)

2.3.4. Anaplásico

Es una palabra que usan los patólogos para describir las células cancerosas de aspecto muy anormal. Si bien la mayoría de las células cancerosas comparten algunas características (forma o tamaño) con las células normales, las células anaplásicas no se parecen en nada a las células normales. Debido a que las células cancerosas anaplásicas no se parecen en nada a las células normales, los patólogos a menudo realizarán pruebas adicionales como inmunohistoquímica para determinar dónde comenzó el cáncer. Desafortunadamente, los cánceres con células anaplásicas tienden a ser muy agresivos y a menudo se asocian con pronóstico. MyPathologyReport (2021)

2.3.5. Apoptosis

Se refiere a la muerte celular programada, lo que sucede cuando las células cancerosas no hacen caso a las señales que indican a las células que dejen de multiplicarse o que deben destruirse. InstitutoNacionaldelCáncer (2021)

3. Descripción del problema

Dentro de las características de las células cancerosas se puede resaltar su inmunidad a la muerte celular programada, es decir, que no responden ante la señal natural que determina su muerte tras cumplir un ciclo determinado (apoptosis). Por otro lado, las proteínas en la membrana de la célula cancerosa son diferentes a las usuales debido a la mutación en las bases nitrogenadas que las componen, provocando así que las células que podrían llevar a cabo una fagocitosis (como macrofagos, células natural killer, entre otras) no las reconozcan y por tanto, no se eliminen. La mutación de las proteínas (tanto en la membrana como internas de la célula) comenzará a causar un cambio en las características de la célula, entre las más notorias al momento de evaluarse bajo el microscopio es su morfología. En general, la célula cancerosa mostrará una distorsión en su forma que seguirá con la células generadas por su replicación. Es por ello, que la célula cancerosa queda habilitada para una replicación descontrolada, con proteínas en constante mutación e inservibles para el cuerpo. InstitutoNacionaldelCáncer (2021)

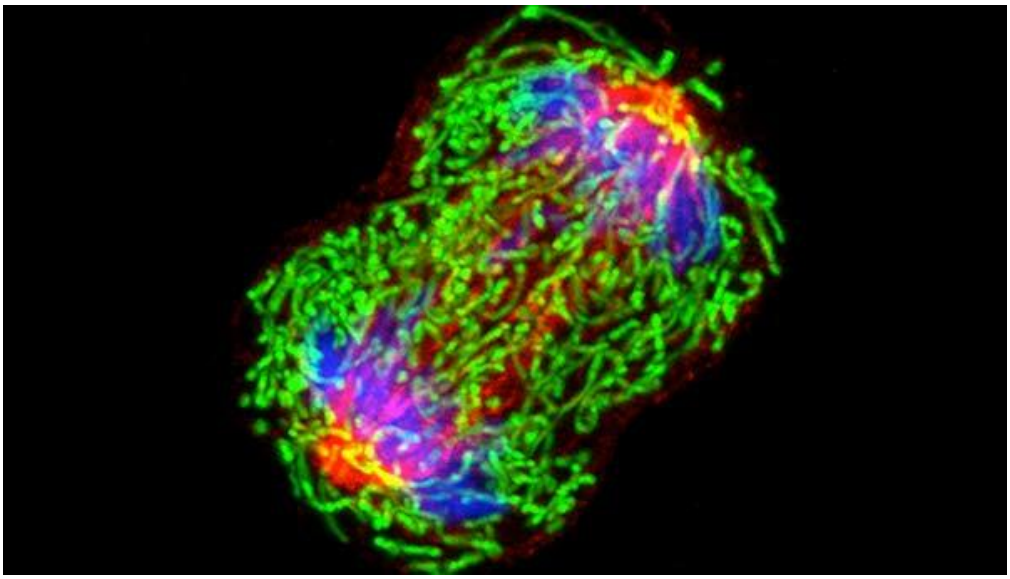

Figura 1: Una célula de cáncer de seno (mama) que se multiplica.
Fuente: Instituto Nacional del Cáncer/ Instituto del Cáncer de la Universidad de Pittsburgh

"En 2020, en todo el mundo se diagnosticó cáncer de mama a 2.3 millones de mujeres, y 685 000 fallecieron por esa enfermedad. A fines del mismo año, 7.8 millones de

mujeres a las que en los anteriores cinco años se les había diagnosticado cáncer de mama seguían con vida, lo que hace que este cáncer sea el de mayor prevalencia en el mundo." WorldHealthOrganization (2020).
Uno de los factores más influyentes en esta enfermedad es la detección temprana AmericanCancerSociety (2021). Por lo que que es sumamente importante el estudio de ésta. En el contexto de este curso de analiza una base de datos para así poder modelar herramientas de la ciencia de datos como pueden ser árboles de decisión, etc. Primeramente se debe de estudiar la base de datos en sí, ósea en que consiste cada variables y clases, junto a sus correspondientes dominios, que es lo que se hace a continuación.

3.1. Descripción de la base de datos

La base de datos *Wisconsin Diagnostic Breast Cancer (WDBC)* fue creada por el Dr.William H. Wolberg, el cuál subió las muestras a medida que informaba sus casos clínicos. La base en cuestión está compuesta por 699 observaciones en total, éstas separadas en 8 grupos, donde cada grupos indica una fecha distinta de la muestra. Cada una de las instancias contiene 11 principales variables las cuales son : ***Sample code number***, ***Clump Thinckness(C. T)***, ***Uniformity of Cell Size (UOC Size)***, ***Uniformity of Cell Shape (UOC Shape)***, ***Marginal Adhesion(M. A)***, ***Single Epithelial Cell Size (SECS)***, ***Bare Nuclei (B. N.)***, ***Bland Chromatin (B. C.)***, ***Normal Nucleoli (N. N.)***, ***Mitoses*** y ***Class*** Wolberg (1992)

3.2. Descripción de clases y variables

Para comenzar con la descripción de las variables se debe contextualizar éstas, entonces se parte desde el hecho de que las variables están derivadas de muestras histológicas, donde cada variable se refiere a las características de las células presente en la muestra. Cada variable toma valores de 1 a 10, siendo 1 más cercano a ser benigno y 10 más anaplásico.Wolberg and Mangasarian (1990)

Nombre variable	Valores posibles	Descripción
Sample Code Number	id	Básicamente es un valor serial numérico utilizado como un identificador de la muestra.
Clump Thinckness	1-10	Se refiere a la cantidad de células cancerosa en las diversas capas de la muestra. Por lo que mientras el valor de la variable (al estar estandarizada de 1 a 10) sea mayor, indica una mayor cantidad de células cancerosas en las diversas capas de la muestra,
UOC Size	1-10	Se refiere a la metástasis presente en los ganglios linfáticos. Por lo que mientras el valor de la variable (al estar estandarizada de 1 a 10) sea mayor, indica se ha propagado más el cáncer a los ganglios linfáticos.
UOC Shape	1-10	Se refiere a la variedad de tamaños de las células anaplásicas. Por lo que mientras el valor de la variable (al estar estandarizada de 1 a 10) sea mayor, indica una mayor variedad de tamaños en las células cancerosas.
Marginal Adhesion	1-10	Se refiere a la pérdida de adhesión, ósea que las células presentes pierden ciertas propiedades, significando la anaplasia de las células. Por lo que mientras el valor de la variable (al estar estandarizada de 1 a 10) sea mayor, indica que las células de la muestras tienen menos propiedades.

Sigue en la página siguiente.

Nombre variable	Valores posibles	Descripción
Single Epithelial Cell Size	1-10	Se refiere al tamaño de la célula epitelial presente en la muestra. Por lo que mientras el valor de la variable (al estar estandarizada de 1 a 10) sea mayor, indica que la célula epitelial es más grande.
Bare Nuclei	1-10	Se refiere a los núcleos desnudos en las células, ósea los núcleos sin revestimiento de citoplasma. Por lo que mientras el valor de la variable (al estar estandarizada de 1 a 10) sea mayor, indica una menor cantidad de núcleos desnudos, ya que éstos están presenten mayoritariamente en muestras benignas.
Bland Chromatin	1-10	Se refiere a la cromatina blanda, usualmente encontrada en muestras benigna. Por lo que mientras menor sea el valor de esta variable (al estar estandarizada de 1 a 10), indica más cromatina suave presente en la muestra.
Normal Nucleoli	1-10	Se refiere a los nucleolos normales presentes en la muestras. Por lo que mientras el valor de la variable (al estar estandarizada de 1 a 10) sea menor, indica que los nucleolos normales son pequeños.
Mitoses	1-10	Proliferación de cada célula, básicamente indica que cuánto se han replicando las células en cuestión mediante la mitosis.

Sigue en la página siguiente.

Nombre variable	Valores posibles	Descripción
Class	2 ó 4	Indica el tipo de muestra que es: Toma el valor 2 para muestras benignas, mientras que toma el valor 4 para muestras malignas (presencia de células anaplásicas)

Tabla 1: Diccionario de las variables y clases.

4. Análisis Estadístico e Inferencial

Para un acercamiento inicial de los datos se realiza un análisis estadístico e inferencial, se utilizan estadísticos básicos, regresiones logísticas y análisis de las componentes principales. Como ya se menciono anteriormente el atributo "Sample Code"se ignora del análisis.

4.1. Análisis Estadístico

Se utilizan estadísticos básicos tales como mediana, media, moda, desviación estándar (**S.D**), varianza (**Var.**), máximo (**Max.**) y mínimo (**Min.**) de los atributos, los cuales pueden ser observados en la tabla 2

	C. T	UOC Size	UOC Shape	M. A	SECS	B. N.	B. C.	N. N.	Mitoses	Class
Mediana	4.0000	1.0000	1.0000	1.0000	2.0000	1.0000	3.0000	1.0000	1.0000	2.0000
Media	4.4421	3.1508	3.2152	2.8301	3.2342	3.5446	3.4450	2.8696	1.6032	2.6998
Moda	1.0000	1.0000	1.0000	1.0000	2.0000	1.0000	3.0000	1.0000	1.0000	2.0000
S.D.	2.8207	3.0651	2.9885	2.8645	2.2230	3.6438	2.4496	3.0526	1.7326	0.9545
Var.	7.9566	9.3951	8.9316	8.2057	4.9421	13.2776	6.0010	9.3187	3.0021	0.9112
Min.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	2.0000
Max.	10.000	10.000	10.000	10.000	10.000	10.000	10.000	10.000	10.000	4.0000

Tabla 2: Estadísticos básicos.

4.2. Análisis inferencial

Primero se utiliza una correlación de Pearson para medir la dependencia lineal entre las variables de todos los datos. Utilizando la función "cor()z "corrplot()"se obtiene la tabla 3 de coeficientes de correlación de Pearson y la figura 2 que es una representación visual de la tabla 3. En dicha tabla se observa que los coeficientes de correlación de Pearson

son todos mayores a cero, por lo que todas las variables poseen una correlación positiva con las otras variables.

	C. T	UOC Size	UOC Shape	M. A	SECS	B. N.	B. C.	N. N.	Mitoses	Class
C. T	1.0000									
UOC Size	0.6424	1.0000								
UOC Shape	0.6534	0.9072	1.0000							
M. A	0.4878	0.7069	0.6859	1.0000						
SECS	0.5235	0.7535	0.7224	0.5945	1.0000					
B. N.	0.5930	0.6917	0.7138	0.6706	0.5857	1.0000				
B. C.	0.5537	0.7555	0.7353	0.6685	0.6181	0.6806	1.0000			
N. N.	0.5340	0.7193	0.7179	0.6031	0.6289	0.5842	0.6656	1.0000		
Mitoses	0.3509	0.4607	0.4412	0.4188	0.4805	0.3392	0.3460	0.4337	1.0000	
Class	0.7147	0.8208	0.8218	0.7062	0.6909	0.8226	0.7582	0.7186	0.4234	1.0000

Tabla 3: Correlaciones de Pearson.

También se hace un análisis de las componentes principales con los datos con la función "PCA()"de la librería "FactoMineR", se obtiene la figura 3 que representa las variables (atributos en la base de datos) y la figura 4 que muestra los individuos, también se tienen ambas figuras en una sola como se ve en la figura 5. Estas ultimas dos figuras muestran las posiciones de cada fila de datos (las enumera), esto permite interpretar acerca de como son los valores de dichas filas respecto a las variables.

Se interpreta de las figuras 3, 4 y 5 que la "Dim 1" se relaciona a la forma de las células mientras que la "Dim 2" se relaciona con el proceso de la mitosis de las células. La tasa de variación es de $74,17\,\%$, para este problema esta un poco por debajo de lo esperado, por ejemplo en Pirchio (2022) obtienen una tasa del $90\,\%$, claro su base de datos

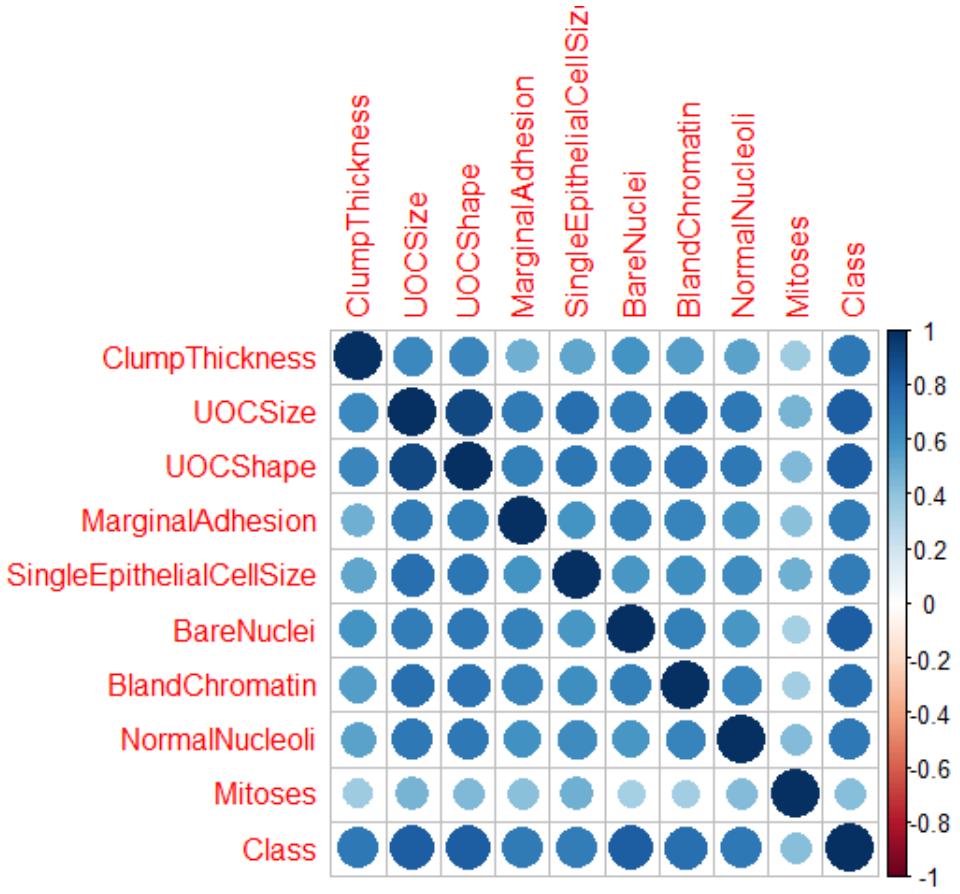

Figura 2: Correlaciones de Pearson.

es diferente pero sirve para acercase a una idea de cuanto se debe obtener, por ello el análisis de las componentes principales en este caso es utilizado principalmente para obtener una representación sencilla de entender para un mejor acercamiento a la base de datos a pesar de no tener la tasa esperada.

Para finalizar se realiza una regresión logística con “UOC size” y “UOC shape” puesto que ambas describen aspectos de la uniformidad de la célula. con la función “lm” se obtuvo la figura 6 que gráfica el modelo resultante.

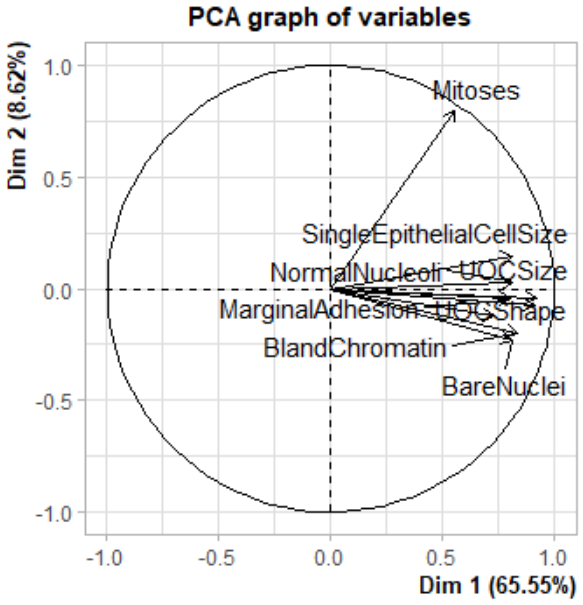

Figura 3: Análisis de las componentes principales.

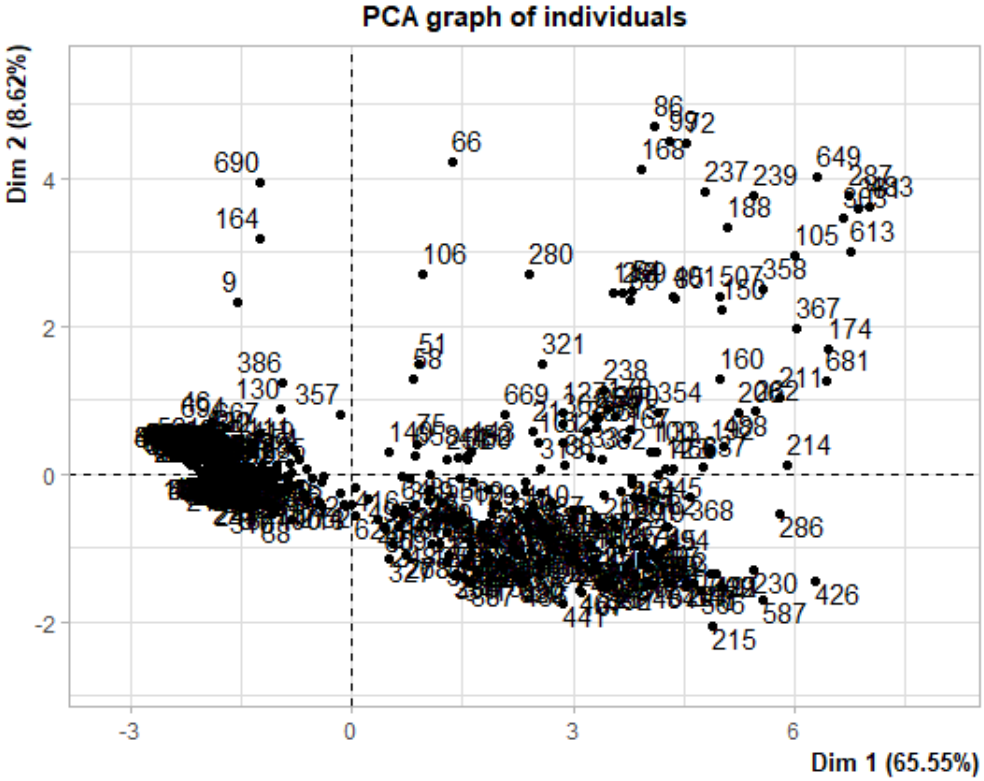

Figura 4: Análisis de las componentes principales.

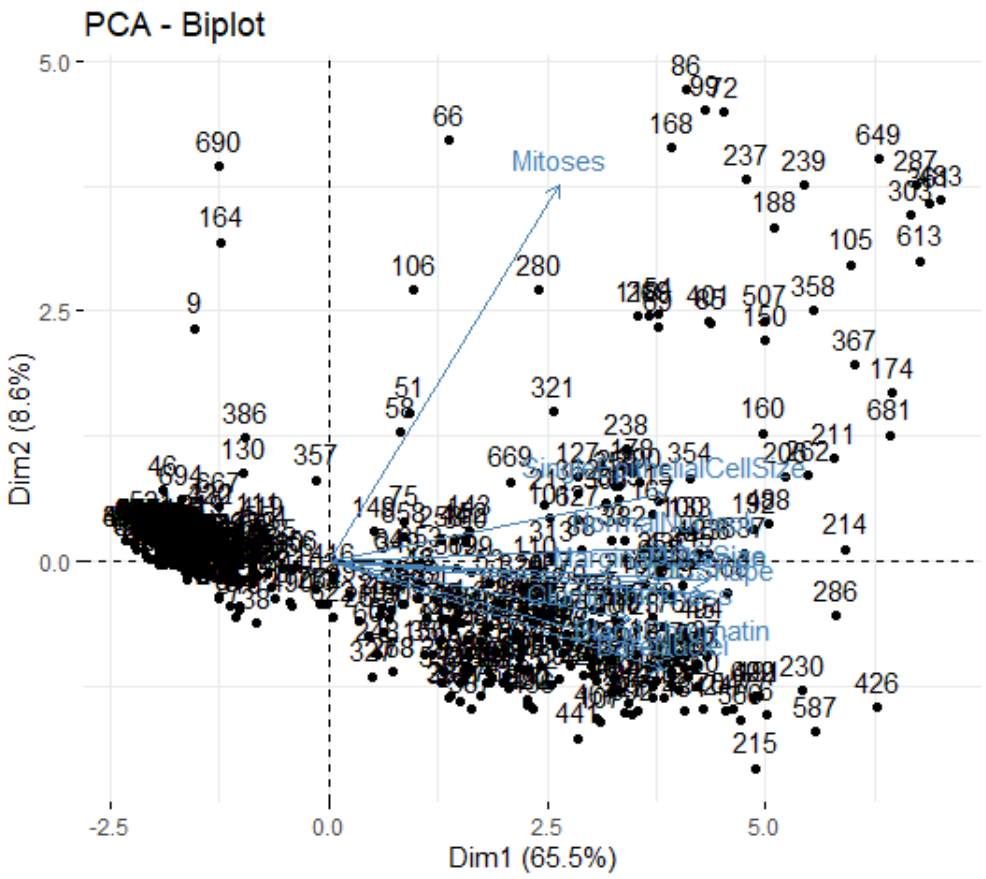

Figura 5: Análisis de las componentes principales.

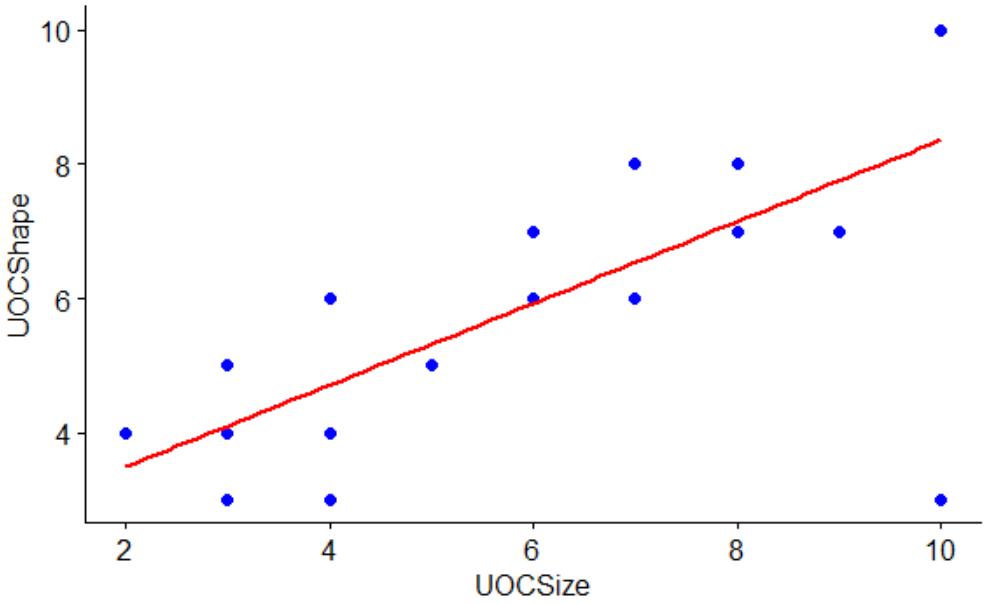

Figura 6: Regresión logística UOC size $\sim$ UOC shape.

5. Conclusión

Para concluir correctamente todo lo expuesto en esta experiencia se opta por separar la conclusión en dos partes, una conclusión estructural donde se desarticula cada apéndice que compone el informe y se analizan éstos, por otro lado una conclusión general de la experiencia, donde se analiza lo expuesto desde una vista general, aterrizando el resultado final al problema y los objetivos planteados de esta experiencia.

5.1. Conclusión estructural

Primeramente se presenta la descripción del problema, dentro de ésta se omite la contextualización, debido a que será abordada en la conclusión. Por lo que, queda la descripción de la base de datos. WDBC en el enlace original de UCI (no su versión beta) contaba con limitada información, debido a que presentaba la cantidad total de instancias de la base de datos, pero no detallaba mucho al respecto de sus variables, ni de las muestras de las que se extrajeron los datos. Es por esto que se optó por una investigación, primeramente recurriendo a la versión beta de UCI, donde detallaba un poco más acerca de las muestras. La estandarización de las variables fue extraída de la investigación del autor de la base de datos titulada, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology" Wolberg and Mangasarian (1990), mientras que la definición de cada variables se encontró en la página de MDPI, esto debido a que las variables siguen un Modelo Novel Matemático de Diagnóstico de Cáncer de Mama. Santiago-Montero et al. (2020). Juntando toda la información información es que se logra desarrollar la descripción de la base de datos y la descripción de clases y variables.

Respecto al análisis estadísticos se utilizaron estadísticos básicos como mediana, moda, media, mínimo, máximo, desviación estándar y varianza. Del análisis de estos estadísticos podemos comprobar que en WDBC se cumple una hipótesis planteada por AmericanCancerSociety (2021) diciendo "Es importante que sepa que la mayoría de los bultos en los senos son benignos y no cancerosos" lo que se puede comprobar en la base de datos viendo la moda obtenida en la clase que es 2, valor que indica que la muestras es

benigna.

Respecto al análisis inferencial es importante centrarse en el ACP, esto debido a que en esta experiencia fue utilizado el entendimiento neto de WDBC, ya que el ACP generó un mapa de coordenada de 2 dimensiones, así reduciones de 9 variables a 2, y a su vez conservando un 74.17 % de la información. Este eje coordenado de dos dimensiones nos describe claramente en el Eje X (haciendo símil a un mapa cartesiano) la anaplasia de las células presentes en la muestras apuntando positivamente a la derecha, mientras que en el eje Y (haciendo símil a un mapa cartesiano) la proliferación de las células presentes en las muestras apuntando positivamente hacia arriba.

5.2. Conclusión general

Dados los objetivos planteados se puede afirmar que se cumplió con ellos, ya que se investigó WDBC, logrando entenderla, comprenderla y analizándola. También se aplicaron las técnicas de estadística descriptiva.

Respectos a las problemáticas presentadas en la experiencia, se pueden separar claramente 2. Por un lado se presentó la problemática de las casi nulas bases en biología por parte de los autores de este documento, lo que se solventó a través de mucho estudio e investigación, ésta siendo plasmada principalmente en el marco teórico de este documento. Por otro lado, se presentó la problemática de los datos, debido a que al ser muestras histológicas, éstas presentan un corte de células a su vez, lo que las células cortadas son interpretadas como células anaplásicas, lo que deriva en altos niveles de anaplasia, lo que también se puede dar al contexto de WDBC, ya que son muestras de un doctor, entonces se asume que hubo una sintomatología previa antes de acudir a supervisión médica. Sin se puede atribuir más a la primera posibilidad, ya que en el documento de investigación del autor Wolberg and Mangasarian (1990) presentan casos de falsos positivos, lo que le da sentido a la problemática del corte histológico.

Bibliografía

AmericanCancerSociety (2021). ¿qué es el cáncer de seno?

César E. Montalvo Arenas, M. M. B. (2018). Tejido linfÁtico y Órganos linfÁticos. *BIOLOGÍA CELULAR E HISTOLOGÍA MÉDICA*.

InstitutoNacionaldelCáncer (2021). ¿qué es el cáncer?

Lalinde, J. D. H., Castro, F. E., Rodríguez, J. E., Rangel, J. G. C., Sierra, C. A. T., Torrado, M. K. A., Sierra, S. M. C., and Pirela, V. J. B. (2018). Sobre el uso adecuado del coeficiente de correlación de pearson: definición, propiedades y suposiciones. *Archivos venezolanos de Farmacología y Terapéutica*, 37(5):587–595.

MyPathologyReport (2021). Anaplástico.

Pirchio, R. (2022). Clasificación de cáncer de mama con técnicas de análisis de la componente principal-kernel pca, algoritmos de máquina de vectores de soporte y regresión logística. *Medisur*, 20(2).

Santiago-Montero, R., Sossa, H., Gutiérrez-Hernández, D. A., Zamudio, V., Hernández-Bautista, I., and Valadez-Godínez, S. (2020). Novel mathematical model of breast cancer diagnostics using an associative pattern classification. *Diagnostics*, 10(3).

Wolberg, W. (1992). Breast Cancer Wisconsin (Original). UCI Machine Learning Repository.

Wolberg, W. H. and Mangasarian, O. L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the National Academy of Sciences*, 87(23):9193–9196.

WorldHealthOrganization (2020). Cáncer de mama.