UNIVERSIDAD DEL QUINDÍO

Supervivencia de pacientes con cáncer

Gabriela Diaz

Sebastian Marin

Física computacional

Programa de Física, Facultad de Ciencias Básicas y Tecnologías

Universidad del Quindío

Resumen En el presente trabajo se hizo un análisis estadístico de una base de datos que expone información detallada de 1904 pacientes diagnosticados con cáncer de mama. Para el tratamiento de los datos, inicialmente se eliminaron aquellos pacientes que tuvieran información insuficiente dado cierto umbral, y se rellenaron los datos de algunos pacientes con la media de dicha variable. Posteriormente, se hicieron diferentes pruebas de normalización y regresión utilizando pruebas de hipótesis, para determinar qué variables de la base de datos predecían de mejor manera la variable elegida, la cual fue el tiempo de supervivencia promedio. Se concluyó que las variables que cumplen lo anterior son los nodos linfáticos examinados positivos, el índice de pronóstico de nottingham, el tamaño del tumor, la terapia hormonal, las celularidades alta y moderada.

1. Introducción

El cáncer es una de las principales enfermedades que afecta tanto a mujeres como a hombres alrededor del mundo. Entre las enfermedades cancerosas, el cáncer de mama es el que tiene mayor tasa de mortalidad, según lo reporta la sociedad americana del cáncer, ya que una de cada ocho mujeres sufrirá de esta enfermedad en el transcurso de su vida[2].
Las etiologías del cáncer de mama son desconocidas, debido a que existe una proliferación anormal en el crecimiento de las células, lo que indica que la enfermedad tiene relación con la alteración en el ADN, llevando a que algunas células muten y se desarrolle el cáncer, que en muchos casos tienden a multiplicarse y disiparse en otras regiones aparte del tejido mamario formando metástasis. Por lo tanto, no existe la forma de prevenir el cáncer, pero la detección temprana permite el tratamiento antes de que se propague a otras zonas del cuerpo. El diagnóstico temprano aumenta la probabilidad de supervivencia, evitando de esta manera tratamientos invasivos y prolongados. Este tipo de tratamientos dependen de la fase de desarrollo en la que se encuentra la enfermedad, y consisten en extirpaciones quirúrgicas, radioterapia, medicación con hormonas y quimioterapia, ayudando a evitar la prolongación del cáncer y mejorando la calidad de vida de los pacientes. El análisis de datos clínicos y genéticos aplicados al uso de técnicas computacionales ha sido de gran ayuda para la toma de decisiones al momento de realizar estimaciones en el tiempo de supervivencia de los pacientes, y de esta manera, evitar tratamientos y procedimientos innecesarios. En este trabajo se busca realizar un análisis exploratorio de datos para hacer una estimación del cálculo de supervivencia de los pacientes utilizando algoritmos matemáticos[1].

2. Metodología

La base de datos utilizada en este proyecto es METABRIC (Consorcio Internacional de Taxonomía Molecular del Cáncer de Mama) tomada de la plataforma Kaggle, que contiene 1980 muestras con cáncer de mama con 1904 pacientes y 693 variables entre atributos clínicos y genéticos. Por consiguiente se realizo un análisis exploratorio y descriptivo de los datos para comprender los datos que se utilizaran en el modelo, mediante técnicas simples de resumen de datos, estadística descriptiva y pruebas de hipótesis con las variables a tomadas, con el fin de conocer cada una de ellas. Como parte importante del proceso, también se identificaron los valores nulos y se eliminaron los pacientes con el $20\,\%$ además de reemplazar con la media los datos nulos para tratar los datos faltantes.

Posterior a esto, se seleccionaron las variables categóricas, ya que permiten hacer un análisis descriptivo del modelo. Una vez se conoce el estado de los datos, se convierten las variables categóricas en "dummies" o variables ficticias; de esta manera, se podrán analizar estas variables de manera numérica y realizar los respectivos análisis estadísticos, como determinar si existe normalidad en las variables utilizadas, determinar la correlación que se tiene entre las diferentes variables independientes e indicar cuáles de esta tienen significancia estadística en el modelo que se esté utilizando.

Se tomaron como variables 10 atributos clínicos, descritos a continuación:

- Patient id:ID del paciente
- Type of breast surgery: MASTECTOMIA, cirugía para extirpar todo el tejido mamario de una mama como una forma de tratar o prevenir el cáncer de mama. CONSERVADORA DE MAMA, se remueve solo la parte de la mama que tiene cáncer.
- Cancer type detailed:Carcinoma ductal invasivo de mama, Carcinoma ductal y lobulillar mixto de mama, Carcinoma lobular invasivo de mama, Carcinoma mucinoso mixto invasivo de mama y Cáncer de mama metaplásico
- Cellularity: cantidad de células tumorales en la muestra.
- Neoplasm histologic grade: Determinado por patología al observar la naturaleza de las células, se ven agresivas o no (toma un valor de 1 a 3)
- hormone therapy :Si la paciente tuvo tratamiento hormonal o no.
- nottingham prognostic index: Se utiliza para determinar el pronóstico después de una cirugía por cáncer de mama. Su valor se calcula utilizando tres criterios patológicos el tamaño del tumor el número de ganglios linfáticos afectados y el grado del tumor.
- Overall survival:Variable objetivo si el paciente está vivo o muerto.
- tumor size :Tamaño del tumor medido por técnicas de imagen.
- death from cancer: si la muerte del paciente se debió a cáncer u otras causas.

Una vez definidas las variables de estudio, se procedió a realizar un análisis exploratorio y descriptivo de los datos, para comprender los datos que se utilizaron en el modelo, mediante técnicas simples de resumen de datos, estadística descriptiva y pruebas de hipótesis. Al realizar la exploración de datos se eliminaron los pacientes con el $20\,\%$ de datos nulos y se identificaron las variables categóricas para reemplazarlas por datos cuantitativos.

3. Análisis y Resultados

Inicialmente, se busca saber si los datos siguen una distribución normal (hipótesis nula). Para ello, se utilizó la variable tamaño del tumor en un gráfico Q-Q y se puede observar que los datos no se ajustan a la normalidad. Para corroborar esta hipótesis, se utilizo el Test K^2 de D'Agostino en todas las variables y se obtuvo que el p-value es menor que el valor de significancia 0.05, por ende se rechaza la hipótesis nula y se asume que los datos tanto en la variable del tamaño del tumor como en las demás no tienen normalidad.

Para la siguiente hipótesis se utiliza la prueba Chi-cuadrado en la que se asume como hipótesis nula que no existe una relación entre las variables y para el otro caso es la alternativa, en este caso el valor de significancia es 0.05 entre las variables tumor size y overall survival, y se tiene un p-valor mayor a la significancia por lo tanto se acepta la hipótesis nula ya que no existe relación entre estas variables.

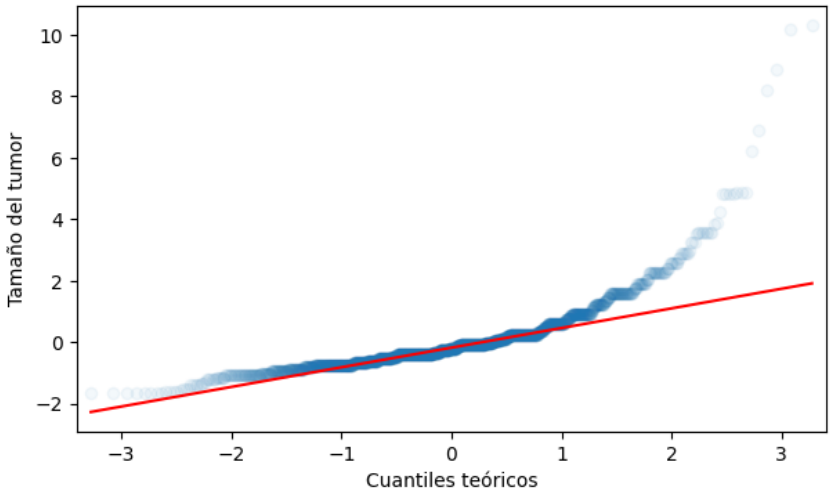

Figura 1: Gráfica Q-Q del tamaño del tumor

Al analizar la matriz de correlación se puede evidenciar que en la mayoria de las variables no existe casi ninguna relación, sin embargo, entre las variables Overall survival y death from cancer Living existe una relación positiva muy fuerte ya que la supervivencia promedio se calcula con base en las personas que hayan logrado sobrevivir a la enfermedad.

Como no se evidencia una relación significativa fuerte entre las demás variables se aplicó un modelo de regresión multilineal. Para ello, se toma como variable independiente Overall survival y las otras variables son las independientes para analizar la relación que hay entre ellas. Para ello la hipótesis nula

$$Y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_i x_i \tag{1}$$

es que los coeficientes β_i son iguales a cero y la alternativa que son diferentes de cero. Así, al momento de hacer la regresión, cada coeficiente tendrá asignado un P valor. Si este último valor es menor a 0.05, se rechaza la hipótesis nula; en caso contrario, se rechaza la alternativa y se hace el valor de este coeficiente a cero, eliminando la variable de la ecuación lineal. En este modelo, según el R-cuadrado, las variables independientes explican es su totalidad (100 %) los cambios en la variable dependiente.

Como se puede apreciar en la figura 3, las variables que explican el cambio de la variable dependiente es del 100 %; sin embargo, en el gráfico de cajas se observó que hay varios valores atípicos para el tamaño del tumor y los ganglios linfáticos positivos, por lo que debería haber una falencia en el modelo que no se está detectando o esos valores son los que deben tener en cuenta para hacer estudios posteriores. Para el índice de nottingham, la mediana muestra que aparentemente los datos tienen una distribución mas homogénea, aproximadamente simétrica, y para la supervivencia de los pacientes se tiene que la mediana esta sesgada, por lo que los datos están concentrados en ciertas partes de la distribución.

Del modelo de regresión multilineal se obtuvo la prueba F, basada en la varianza para verificar la igualdad de la media de las diferentes variables. En este caso, la hipótesis nula es que todas las medias de los grupos son iguales, pero como la distribución de los datos no es homogénea y el F es inversamente proporcional al p valor, que para este caso se tiene un F alto, indica que el p valor es mas significativo. Por consiguiente, se rechaza la hipótesis nula ya que hay diferencias significativas entre variables y se concluye que las medias de las variables son diferentes entre sí.

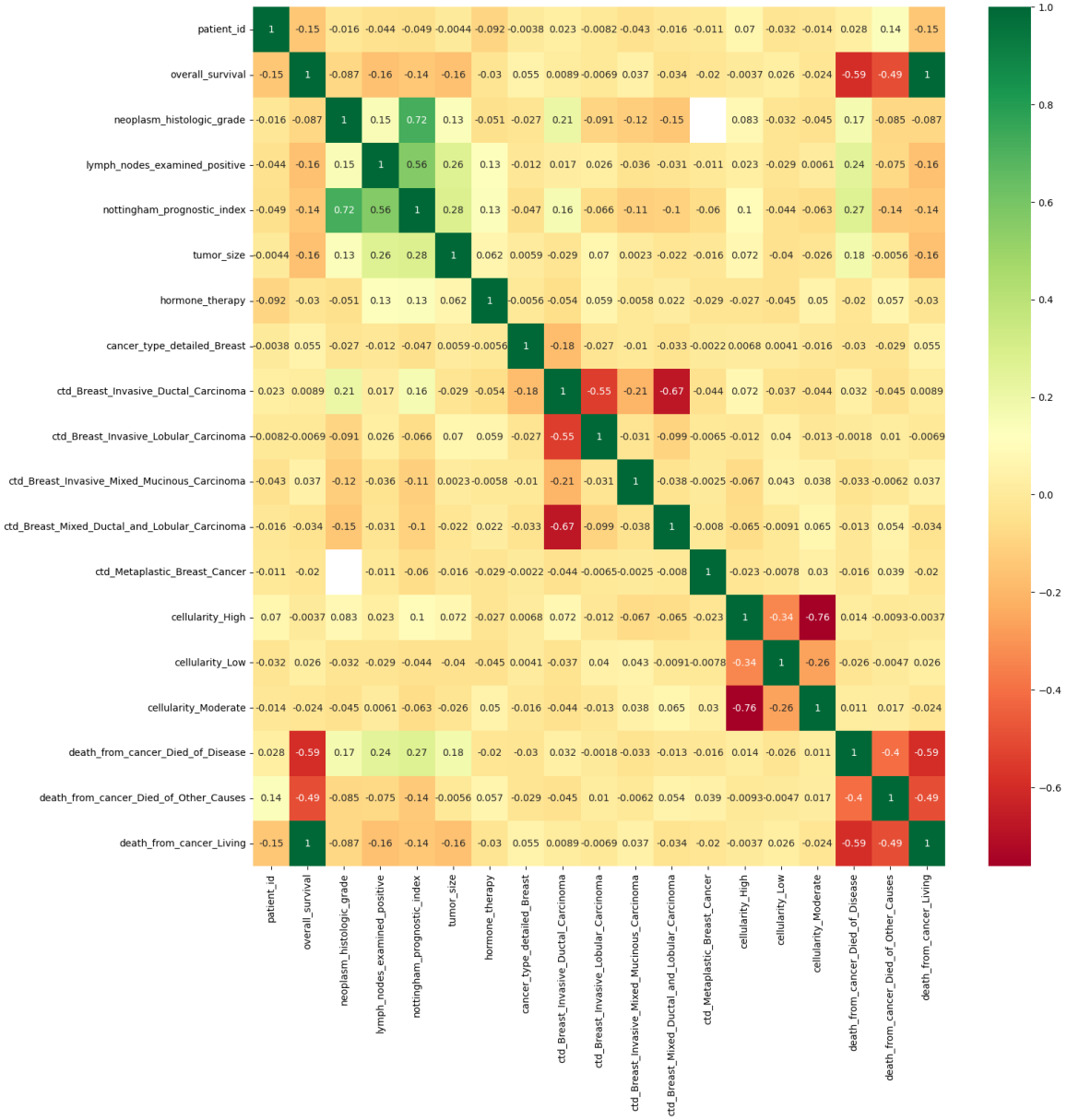

Figura 2: Correlación entre variables

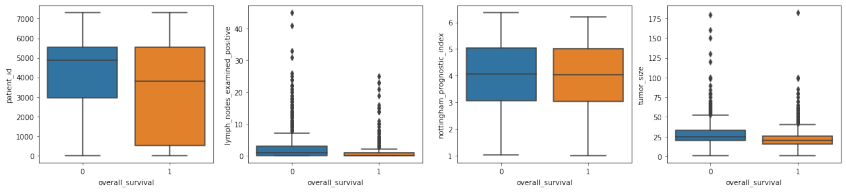

Figura 3: Box-plot entre de las variables utilizadas en la regresión

4. Conclusiones

Principalmente, se concluye que las variables que mejor explican y que pueden predecir la supervivencia promedio de un paciente son los nodos linfáticos examinados positivos, el índice de pronóstico

de nottingham, el tamaño del tumor, la terapia hormonal, las celularidades alta y moderada. Además, probablemente muchas variables no resultaran como era debido ya que no se hizo una normalización de los datos para hacer las regresiones.

Referencias

[1] Baljit Singh Amar Partap Singh Pharwaha. Shannon and non-shannon measures of entropy for statistical texture feature extraction in digitized mammograms. *Member, IAENG*, 2009.

[2] Sergio Vitulano and Andrea Casanova. The role of entropy: Mammogram analysis. *Università di Cagliari*, 2008.