MEMORIA DE PROYECTO

Estudio de las mutaciones generadas como consecuencia de tumores en diferentes órganos, su relación con la tipología de pacientes y el impacto en la supervivencia

1. JUSTIFICACIÓN DEL PROYECTO

En el contexto del Bootcamp de Análisis de Datos organizado por The Bridge se pide la realización de un proyecto en el que desarrollen las habilidades adquiridas hasta el momento en esta materia. En mi caso parto de una necesidad real por parte de una profesional de la investigación que trata de estudiar las modificaciones del sistema inmune inducidas por las mutaciones en los órganos en los que se encuentra un tumor. El presente proyecto se centrará en el punto intermedio entre el propio tumor y las modificaciones del sistema inmune. Es decir, analizaremos las mutaciones encontradas en los pacientes y trataremos de descubrir si existen conexiones entre estas mutaciones (tipo y cantidad) y a tipología de paciente (sexo, edad, etnia) y relacionar esto con la supervivencia de dichos pacientes.

2. OBTENCIÓN DE DATOS

En primer lugar, obtuvimos los datos de qué mutaciones podemos encontrar en los diferentes órganos. Se seleccionaron dos de estos órganos (páncreas y pulmones) y para cada uno de ellos se eligieron las mutaciones más frecuentes.

A continuación, para cada órgano seleccionado y a partir de las bases de datos de dominio público existentes, se escogieron las características de interés de los pacientes (en nuestro caso, edad, sexo y etnicidad) y se descargaron las tablas con dicha información. Por otra parte, se obtuvo la información genética sobre el tumor de cada paciente.

3. LIMPIEZA YADECUACIÓN DE LOS DATOS

De manera sorprendente varios de los pacientes no tenían asociada ninguna mutación por lo que fueron descartados para el análisis. Del mismo modo se desecharon aquellos pacientes con datos nulos. Para facilitar el manejo de datos y el entendimiento de los procesos se asignó un código alfabético a cada tipo de mutación. A continuación, se añadió una columna de creación propia donde se recoge, por tipos, todas las mutaciones de cada paciente para ver la variabilidad de combinación de dichas mutaciones y así estudiar si cada tipo de asociación impacta de manera diferente en algunas de las variables de análisis.
Por otra parte, se generó una matriz que fuera susceptible de ser analizada mediante una matriz de correlación. Para ello se sustituyeron las posiciones de cada mutación por el valor 1 y las posiciones en las que no hay mutación con -1. Se trata de generar una matriz lo más simétrica posible.

Finalmente se procedió a la unión de las tablas de pacientes con las tablas de sus mutaciones para proceder al análisis.

4. ANÁLISIS DE LOS DATOS

En primer lugar, se procedió a estudiar los *outliers* para cada *dataframe* y a eliminar los que se consideraron perjudiciales para el análisis. En el presente estudio solamente se eliminaron aquellos valores que:

- Estuvieran aislados
- Estuvieran en posiciones muy lejanas del Q75

Esto se debe a que el objeto del estudio consiste en valorar diferencias de procesos en función de valores altos vs valores bajos. Si eliminamos muchos valores altos, que además se encuentran en posiciones sucesivas y próximas a Q75, se temía estar perdiendo una información valiosa.

A continuación, se obtuvieron las matrices de correlación que no mostraron valores significativos para ninguno de los órganos.

El proceso posterior del análisis consistió en comparar variables dos a dos y mostrarlas en función del sexo para ver si este factor era relevante en la supervivencia de pacientes bajo determinadas condiciones de dichas variables de estudio. Solamente obtuvimos indicios para futuros estudios más minuciosos en el páncreas.

Por último, se llevó a cabo análisis multifactorial por cada tipo de asociación de mutaciones, por género. Aquí se apreciaron algunos datos interesantes para unas pocas de esas asociaciones, en las que la supervivencia por género mostraba considerables diferencias.

Sería interesante realizar posteriores estudios específicos sobre dichas asociaciones de mutaciones.

Lamentablemente, la nula calidad de la información sobre etnicidad de los pacientes hizo imposible tener en cuenta este factor en el estudio realizado.

5. CONCLUSIONES

No se aprecian correlaciones significativas en las matrices. Sin embargo, cuando realizamos algunas de las comparativas dos a dos, se aprecian algunos valores para los que podría haber alguna diferencia en la supervivencia en función del género.

Por otra parte, el análisis multivariable muestra diferencias en el promedio de supervivencia en función del género para algunas de las asociaciones de mutaciones.