Identificación de variables importantes y redundantes

El objetivo de este trabajo es identificar entre todas las variables que tenemos inicialmente cuales se deben considerar de aquí en adelante y cuales variables simplemente no aportan.

Variables con muchos datos faltantes

Tabla 1: Varibles con mayor número de faltantes

Variables	**Cantidad de faltantes**
Recaídas	131
PD-L1	191
Área ocupada por los TILs estromales (total)	139

En análisis descriptivos anteriores encontramos que las variables con mayor cantidad de datos faltantes son **PD.L1 (191)**, **Área ocupada por los TILs estromales % total** (139), **recaídas** (131), dado que estas variables presentan una gran cantidad de valores faltantes serán removidas de nuestras variables ya que hacer un método de imputación nos llevará a tener grandes sesgos en nuestros análisis futuros y en la construcción de modelos , ademas dejar estas variables también significaría perder una gran cantidad de información de las demás variables.

Recategorización de variables

En el conjunto de variables tenemos variables que están categorizadas de dos o tres formas diferentes, estas variables son:**edad**,**estrato**,**educación**,**tipo_histologico**,**estadio**,**EUR**,**NAM**,**AFR**, lo que nos interesa saber con que categorización de dichas variables es la que representa mejor la información y así reducir la cantidad de variables de nuestro conjunto.

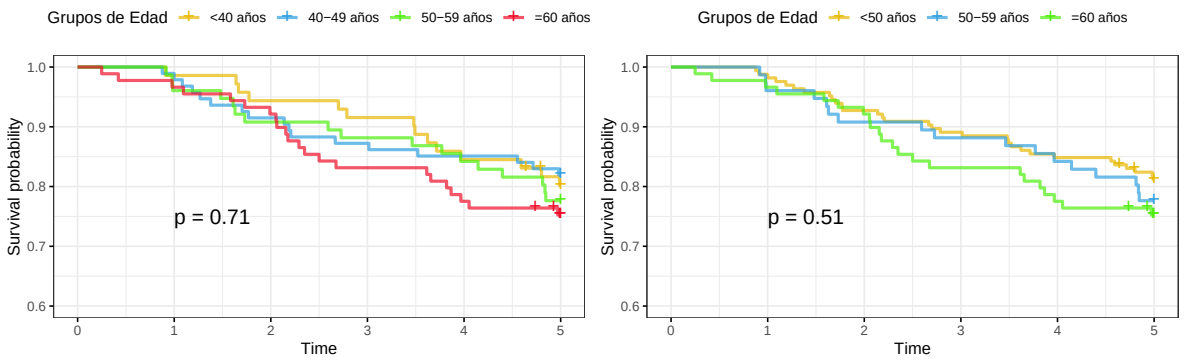
Para lograr esto vamos a usar el método del log-rank test para saber en que categorización resulta con curvas de supervivencia estadisticamente diferentes en por lo menos dos categorías de la variable.

La prueba del Log-Rank test

En el análisis de supervivencia, la prueba del Log-Rank test se usa para comparar las curvas de dos o mas grupos. Esta prueba se basa en la hipótesis nula de que no hay diferencia entre las funciones de supervivencia de los grupos comparados.

- La hipótesis nula (H_0): No hay diferencia en las curvas de supervivencia entre los grupos.
- La hipótesis alternativa (H_1): Al menos un grupo tiene una función de supervivencia diferente.

Variable edad



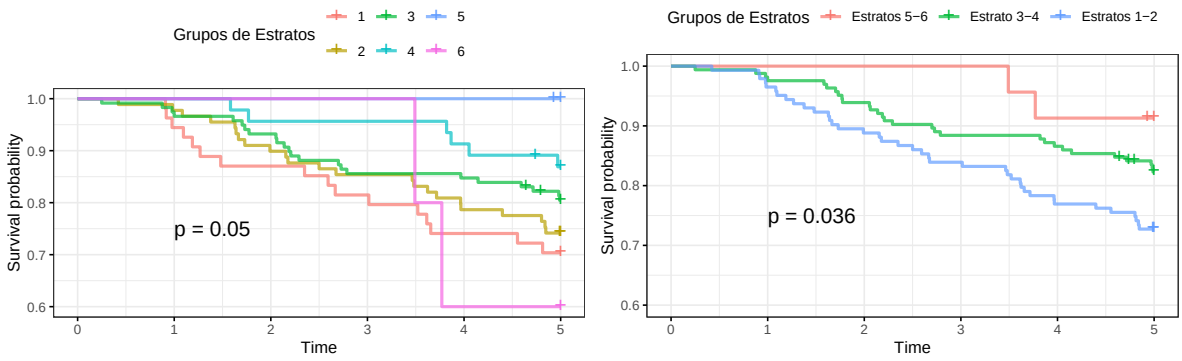
(a) edad_cat

(b) edad_cat2

Figura 1: Curvas de supérvivencia para la variable edad

El P-valor de la prueba es el que aparece dentro de la gráfica, para el caso de la variable edad vemos que las categorizaciones planteadas resultan ser iguales estadisticamente en las diferentes categorias de la variable edad por esta razón decidimos trabajar con la variable numérica de edad.

Variable estrato

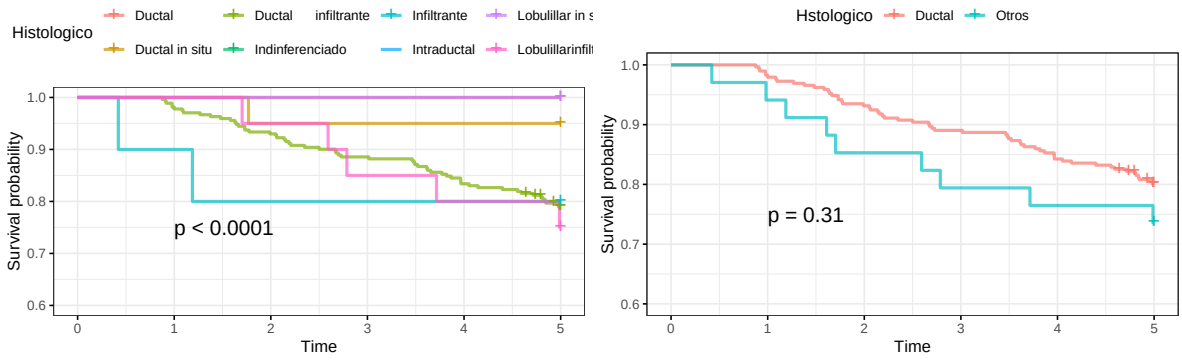


(a) estrato (b) estrato_cat

Figura 2: Curvas de supérvivencia para la variable estrato

La categorización presentada en la variable **estrato_cat** parece agrupar mejor la información, ya que su P-valor resulta ser mas significativo, así que trabajaremos con la variable **estrato_cat**.

Variable tipo_histologico

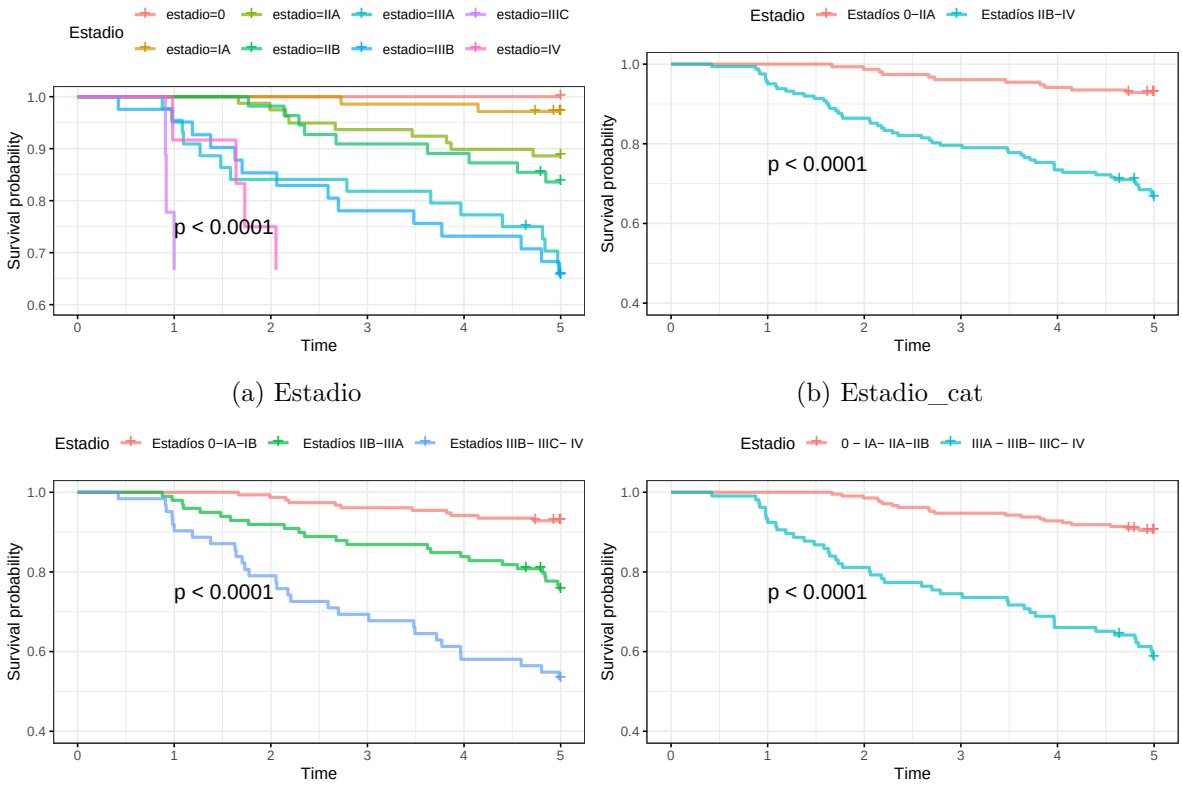


(a) tipo_histologico (b) tipo_histologico_cat

Figura 3: Curvas de supérvivencia para la variable tipo_histologico

Para el caso de la variable **tipo_histologico** notamos que la variable con muchas categorias es la que resulta ser estadisticamente diferente pero para este caso dado que en la variable **tipo_histologico** tiene muchas categorias lo cual no es factible por tanto ns quedaremos con la variable que tiene solo dos categorias.

Variable estadio



(a) Estadio

(b) Estadio_cat

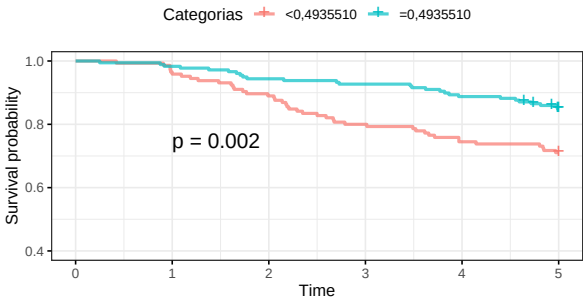
(c) Estadio_cat3

(d) estadio-Early_late

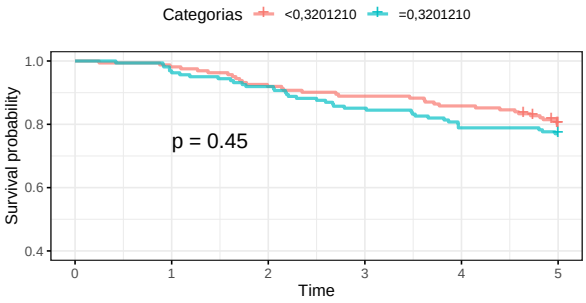
Figura 4: Curvas de supérvivencia para la variable estadio

La variable estadio, es la variable que mas categorizaciones tienen y en cada una de ellas hay al menos dos curvas que son estéticamente diferentes, pero en este caso la variable **estadio_cat3** es la elegida en continuar en el análisis, esta categorización de la variable estadio tiene tres categorias que son las ideales en una variable categórica y ademas cada curva resulta ser estadisticamente diferentes lo que significa que cada categoría tiene información relevante y su supervivencia tiene comportamiento diferente.

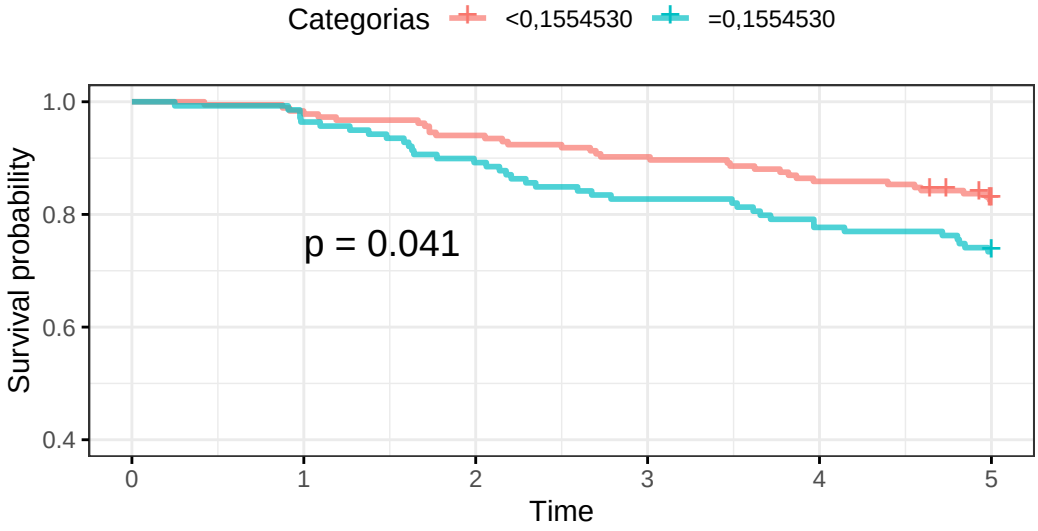
Variable componente Ancestral



(a) Europeo



(b) Nativo Americano



(c) Africano

Figura 5: Curvas de supérvivencia para la variable Componente Ancestral

Para las variables que miden el componente Ancestral de los pacientes Europeo resulta significativa en las dos categorias, de igual forma para el componente ancestral africano, pero la variable nativo americano resulta ser estadisticamente iguales, por eso decidiremos trabajar con la variable numérica y para las demás usaremos las categóricas.

Con este método logramos identificar con que variable categórica trabajar y cuales hacer a un lado hemos reducido el numero de variables, ahora tratemos de identificar variables que no aportan información útil.

Revisaremos las variables restantes de nuestro conjunto de datos y argumentaremos si son relevantes o no.

Dentro de los estudios de análisis de supervivencia es importante la variable que registra si un paciente presentó o no el evento así como también el periodo bajo estudio de los pacientes, en la base de datos tenemos periodos de 2 y 5 años, entre mas tiempo bajo estudio mejor, por este motivo usaremos el periodo de 5 años así que las demás variables que tengan que ver con dos años serán dejadas a un lado, así como también las variables que tienen fechas o años pues ya la variable tiempo de supervivencia la tenemos en días y en años para el caso de a variable grado nuclear y grado histologico segun el diccionario de variables miden lo mismo usaremos grado nuclear ya que tiene menos grados, dicho esto las variables que se van a considerar para la creación de modelos son las siguientes: **estado_vital_5años**,**tiempo_evento_bx_5años**,**ciudad**,**edad**,**estrato_cat**,**educacion_cat** ,**afiliacion**,**lateralidad_cat**,**tipo_histol_cat**,**grado_nuclear**,**gh.gn**,**T**,**N**,**M**,**estadio_cat3**, **ER**,**PR**,**HER2**,**subtipo_molecular_definitivo**,**EUR_cat**,**NAM**,**AFR_cat**, **Interacción_Reg.stage**,**PD.L1.TILs.si.no**,**Missing.Clinical.data**.

Conclusiones

Gracias al análisis descriptivo y al análisis de supervivencia de nuestro conjunto de datos pudimos identificar que 25 variables entre las 54 que teníamos inicialmente son las que se van a considerar para implementación de modelos que se van a desarrollar.

Como primera opción vamos a considerar la implementación de un modelo de **regresión de riesgos proporcionales de cox** ya que este permite analizar el efecto de múltiples variables explicativas sobre el tiempo hasta la ocurrencia de un evento.