Identificación de factores de riesgo en pacientes de diabetes post pabellón.

Matías F. Rebolledo G.

Martes, 8 de Octubre de 2019

Objetivos:

• Identificar factores de riesgo asociados a la reincidencia de pacientes en un periodo de 30 posterior a intervención en pabellón.

Estrategias:

• Muestra hecha a partir de un cuestionario que contiene variables altamente colineales

$\Longrightarrow$ Modelo de regularización en la función de verosimilitud

• Muestra imbalanceada con una proporción de casos de un 0.04% versus de controles de 0.96%

$\Longrightarrow$ Optimización de costo de sensitividad usando ponderación

$\Longrightarrow$ Modelo con remuestreo para muestra imbalanceada

• Número total de pacientes de 284.

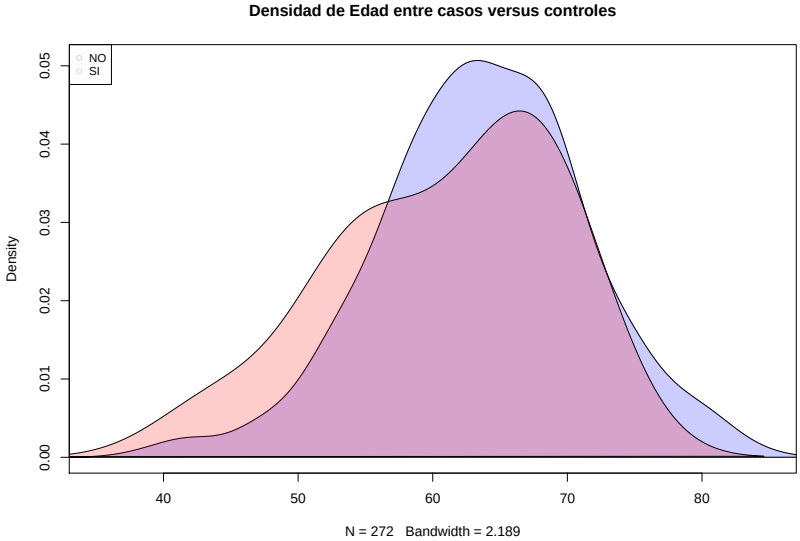
$\Longrightarrow$ Modelo con validación cruzada

Etapas:

- Etapa análisis descriptivo
  - Grupo incluido variables factor
    DIABETES, BIOP.PREV,
    VOL.PROST,
    ANTIB.PROFI, BIOPSIA
  - Grupo excluido variables factor
    EPOC, NUM.DIAS.POST.BIOP,
    FIEBRE, TIPO.CULTIVO,
    AGENTE.AISLADO, PATR.RESIST
    $\Longrightarrow$ Combinación de fórmulas desde 1 ... 3080
  - Grupo opcional variables factor
    HOSP.ULT.MES, CUP, ITU
- Etapa de tipificación de usando k-means
- Etapa modelación para identificar variables significativas
  - Identificar modelos con métricas más elevadas
  - Usar otras métricas que representen mejor el imbalance, tales como, Precision, Recall, F1 y Precision-Recall-AUC o punto de corte óptimo
- Etapa de configuración de modelos para controlar parámetros específicos
  - Estimación de parámetros usando grillas
  - Cost-Sensitive Training

	cor_x	cor_y	cor	Pval
1	HOSPITALIZACION	FIEBRE	0,9590	0
7	HOSPITALIZACION	DIAS.HOSP.MQ	0,8075	0
8	HOSPITALIZACION	NUM.DIAS.POST.BIOP	0,7871	0
16	HOSPITALIZACION	AGENTE.AISLADO	0,5373	0
17	HOSPITALIZACION	PATR.RESIST	0,5188	0
20	HOSPITALIZACION	TIPO.CULTIVO	0,5136	0

	cor_x	cor_y	cor	Pval
2	PATR.RESIST	TIPO.CULTIVO	0,9421	0
3	FIEBRE	NUM.DIAS.POST.BIOP	0,9019	0
4	PATR.RESIST	AGENTE.AISLADO	0,8958	0
5	AGENTE.AISLADO	ITU	0,8714	0
6	PATR.RESIST	ITU	0,8415	0
9	DIAS.HOSP.MQ	AGENTE.AISLADO	0,7781	0
10	AGENTE.AISLADO	TIPO.CULTIVO	0,7781	0
11	DIAS.HOSP.MQ	FIEBRE	0,7744	0
12	TIPO.CULTIVO	ITU	0,7425	0
13	DIAS.HOSP.MQ	NUM.DIAS.POST.BIOP	0,7132	0
14	DIAS.HOSP.MQ	PATR.RESIST	0,6231	0
15	DIAS.HOSP.MQ	ITU	0,6092	0
18	AGENTE.AISLADO	FIEBRE	0,5153	0
19	DIAS.HOSP.UPC	TIPO.CULTIVO	0,5139	0



Identificación de factores de riesgo en pacientes de diabetes post pabellón.

Matías F. Rebolledo G.

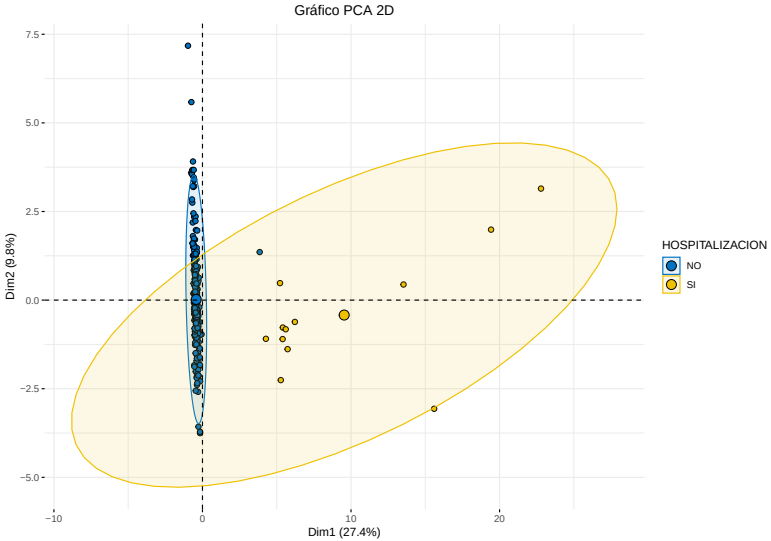
PCA:

- Propósitos:
  - Graficar PCA para visualizar separación de la variable de respuesta
  - Para visualizar separación de otros modelos de clasificación
- Mejoras:
  - Se usaron 2 versiones de matriz de correlaciones: para datos continuos (stats::cor) y para datos mixtos (polycor::hetcor) obteniendo las mismas PCA siempre usando datos centrados
  - Se usó la siguiente fórmula

$$\text{COMP}_{PCA} = \left[\text{EIG(COV)}^T * \text{DATOS}^T\right]^T$$

$$(n \times p) = \left[(p \times p)^T * (p \times n)\right]^T = (n \times p)$$

- Se usó una librería especializada para graficar las PCA, pero no soporta data.frame ni matrix, solo objetos prcomp. La reemplazamos por la función base.
  - Usamos la nube base PCA (PC1 versus PC2) para visualizar la clasificación de otros modelos



Identificación de factores de riesgo en pacientes de diabetes post pabellón.

Matías F. Rebolledo G.

Modelo k-means:

- Mejoras:
  - distancia euclideana $\Longrightarrow$ distancia gower
  - K-means $\Longrightarrow$ K-prototypes
  - Se ejecutaron 450 corridas usando 50 semillas diferentes

generadas con distribución uniforme y para valores de k $= \{2,\ldots,10\}$

- Coeficiente silhouette se mantuvo
- Se ejecutarán 50 corridas adicionales en cuanto a variables significativas
- Todas las variables numéricas están centradas
- Usaremos matriz de confusión y algunas métricas en todos los modelos

	Obs	
Pred	SI	NO
SI	TP	FP
NO	FN	TN

	Obs	
Pred	NO	SI
NO	TN	FN
SI	FP	TP

$$Acc = \frac{(TP + TN)}{(TP + FN + FP + TN)}$$

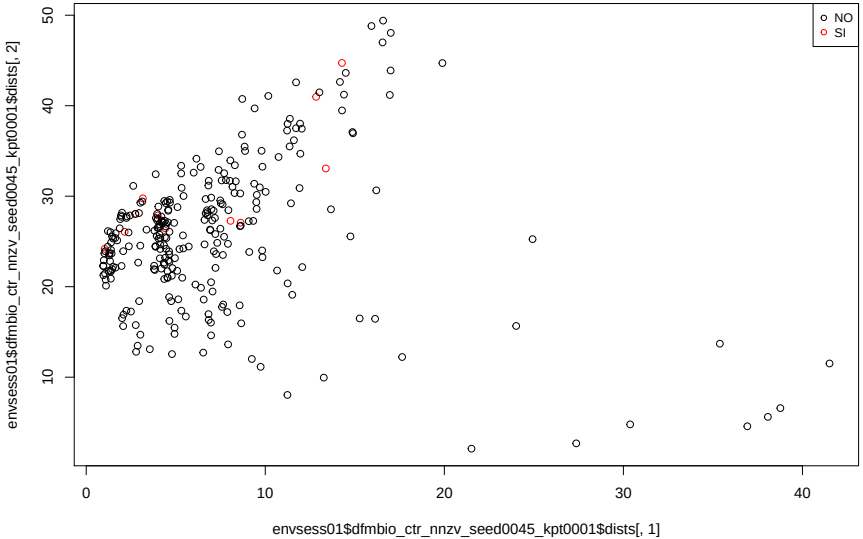
$$Sens = (TP)/(TP + FN)$$

$$Spec = (TN)/(TN + FP)$$

- Se exhibe un patrón donde Sens es bajo mientras que el de Spec y Acc son altos
- Se exhibe un patrón donde Sens es mediano mientras que el de Spec y Acc son también medianos
- Se exhibe un patrón donde Sens es alto mientras que el de Spec y Acc son bajos

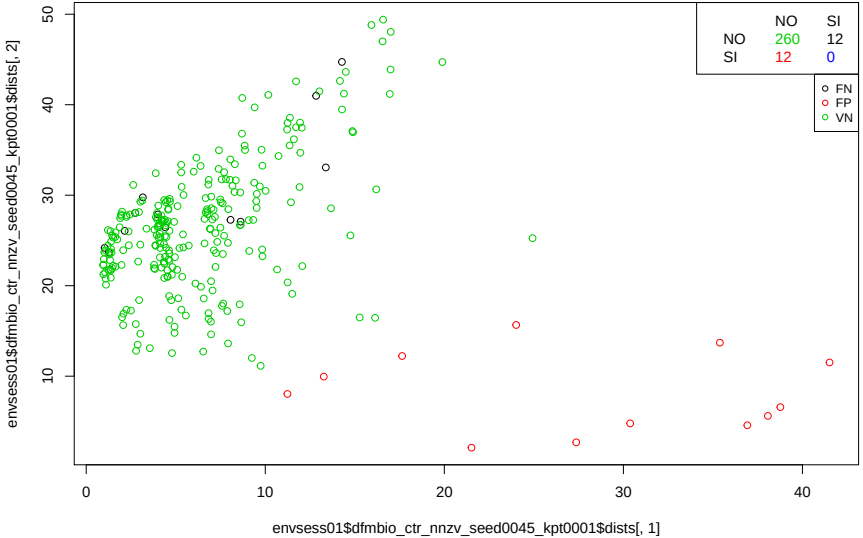
Semilla	k	Acc	Sens	Spec
seed0045	2	0,9155	0,0000	0,9559
seed0033	2	0,8979	0,0000	0,9375
seed0020	2	0,8732	0,1667	0,9044
seed0001	2	0,8310	0,2500	0,8566
seed0031	2	0,7676	0,0833	0,7978
seed0044	2	0,7183	0,2500	0,7390
seed0041	2	0,7042	0,0833	0,7316
seed0049	2	0,6725	0,1667	0,6949
seed0015	2	0,6655	0,1667	0,6875

Semilla	k	Acc	Sens	Spec
seed0013	2	0,3345	0,8333	0,3125
seed0003	2	0,3275	0,8333	0,3051
seed0014	2	0,1408	0,8333	0,1103
seed0007	2	0,5035	0,7500	0,4926
seed0008	2	0,4965	0,7500	0,4853
seed0050	2	0,4859	0,7500	0,4743



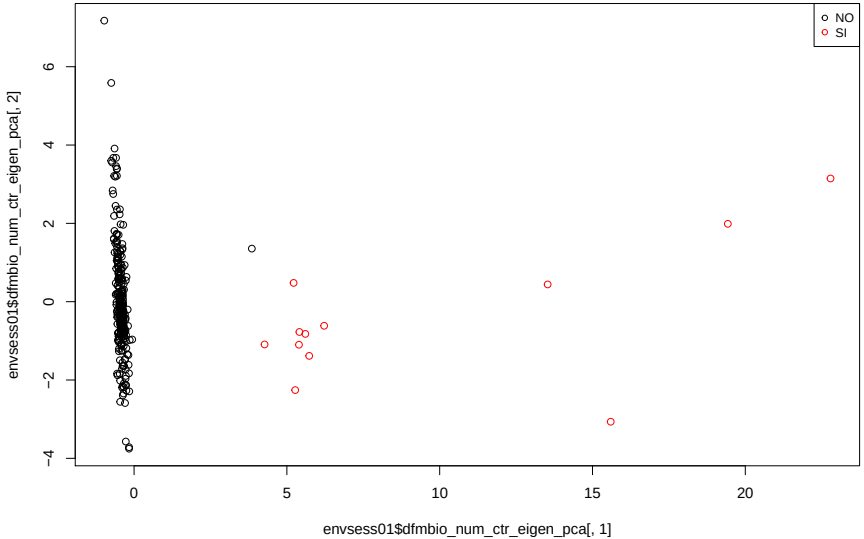
Identificación de factores de riesgo en pacientes de diabetes post pabellón.

Matías F. Rebolledo G.



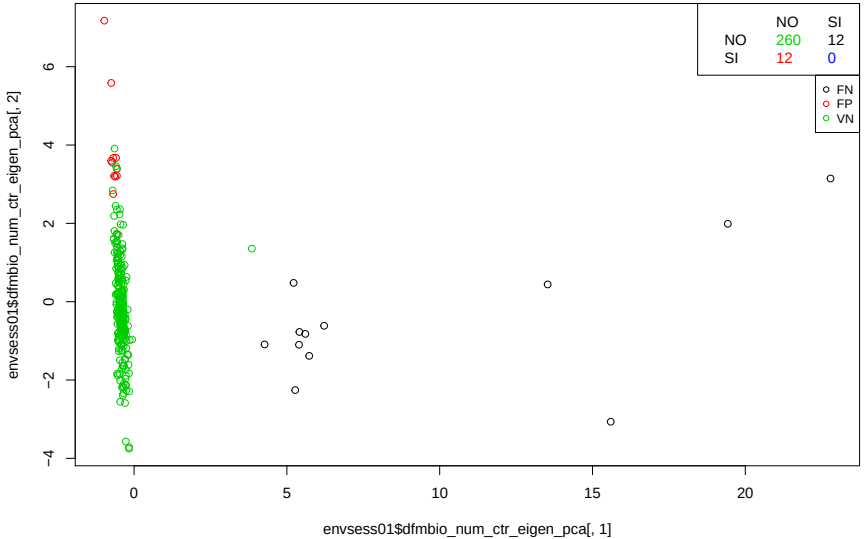
Identificación de factores de riesgo en pacientes de diabetes post pabellón.

Matías F. Rebolledo G.



Identificación de factores de riesgo en pacientes de diabetes post pabellón.

Matías F. Rebolledo G.



Identificación de factores de riesgo en pacientes de diabetes post pabellón.

Matías F. Rebolledo G.

Modelo glm:

- Mejoras:
  - Se separaron las variables en tres grupos en términos de modelamiento
  - Se definieron hasta ahora 3080 fórmulas (objetos) diferentes incluyendo interacciones de tipo numéricas y factor
  - Se eliminó el intercepto en todas las fórmulas debido a que las variables numéricas están centradas
  - Creamos una variable auxiliar llamada ponderador que nos permite plantear un modelo ponderado para cada fórmula (total 6160 modelos)

```
dfmpnd$ponderador [ dfmbio$HOSPITALIZACION == "SI" ] <-
( 1 - proporcion ) = 0.9577
dfmpnd$ponderador [ dfmbio$HOSPITALIZACION == "NO" ] <-
( proporcion ) = 0.0423
```

  - Hay otras opciones que deben ser mejoradas configurando los modelos, relacionado con, punto de corte, parámetros de ajuste, remuestreo (sub y sobre) y validación cruzada.

Modelo	Dev	Acc	Sens	Spec	ac_y	se_y	sp_y	NC	NS
modglmfrw2584	267,03	0,77	0,92	0,76	0,73	1,00	0,72	33	33
modglmfrw2596	279,21	0,75	0,92	0,75	0,81	0,92	0,80	30	30
modglmfrw2642	312,72	0,71	0,92	0,71	0,73	0,92	0,72	31	31
modglmfrw2670	327,95	0,70	0,92	0,69	0,65	1,00	0,64	30	30
modglmfrw1298	330,99	0,69	0,92	0,68	0,73	0,92	0,72	28	28

Modelo	Dev	Acc	Sens	Spec	ac_y	se_y	sp_y	NC	NS
modglmfrw0627	11,79	0,84	0,92	0,84	0,83	1	0,82	27	0
modglmfrw0609	12,48	0,84	0,92	0,83	0,82	1	0,81	24	0
modglmfrw1743	14,08	0,83	0,92	0,83	0,81	1	0,81	26	0
modglmfrw2575	12,65	0,83	0,92	0,82	0,83	1	0,82	32	0
modglmfrw0615	12,81	0,82	0,92	0,82	0,82	1	0,81	27	0

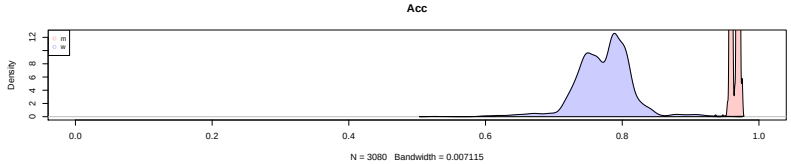
Modelo	Dev	Acc	Sens	Spec	ac_y	se_y	sp_y	NC	NS
modglmfrm2671	504,61	0,98	0,42	1,00	0,72	0,83	0,71	30	30
modglmfrm3006	576,70	0,97	0,42	1,00	0,68	0,92	0,67	32	32
modglmfrm2547	576,70	0,97	0,33	1,00	0,70	0,92	0,69	30	30
modglmfrm2557	648,79	0,97	0,42	0,99	0,62	1,00	0,61	29	29
modglmfrm2559	648,79	0,97	0,33	1,00	0,85	0,75	0,85	30	30

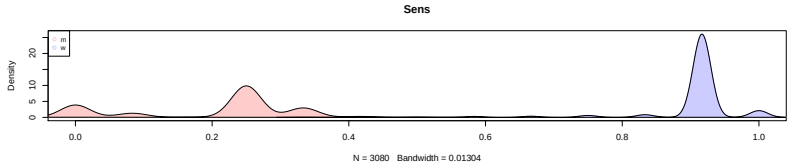
Modelo	Dev	Acc	Sens	Spec	ac_y	se_y	sp_y	NC	NS
modglmfrm2569	53,15	0,98	0,42	1	0,86	0,83	0,86	29	2
modglmfrm0067	58,03	0,97	0,33	1	0,78	0,92	0,78	24	0
modglmfrm0627	50,68	0,97	0,33	1	0,85	0,92	0,85	27	0
modglmfrm0004	65,31	0,97	0,25	1	0,83	0,83	0,83	20	0
modglmfrm1743	58,80	0,96	0,25	1	0,92	0,75	0,93	26	0

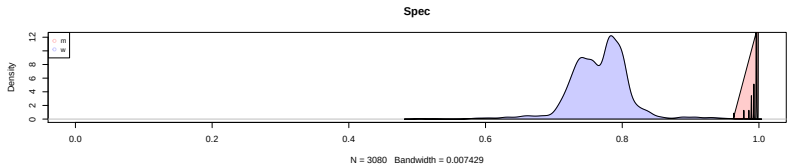
Pond	Acc.xbar	Acc.med	Acc.sd	Acc.n
m	0,9660	0,9683	0,0053	3.080
w	0,7731	0,7782	0,0411	3.080

Pond	Sens.xbar	Sens.med	Sens.sd	Sens.n
m	0,2016	0,2500	0,1201	3.080
w	0,9060	0,9167	0,0722	3.080

Pond	Spec.xbar	Spec.med	Spec.sd	Spec.n
m	0,9997	1,0000	0,0016	3.080
w	0,7672	0,7721	0,0441	3.080

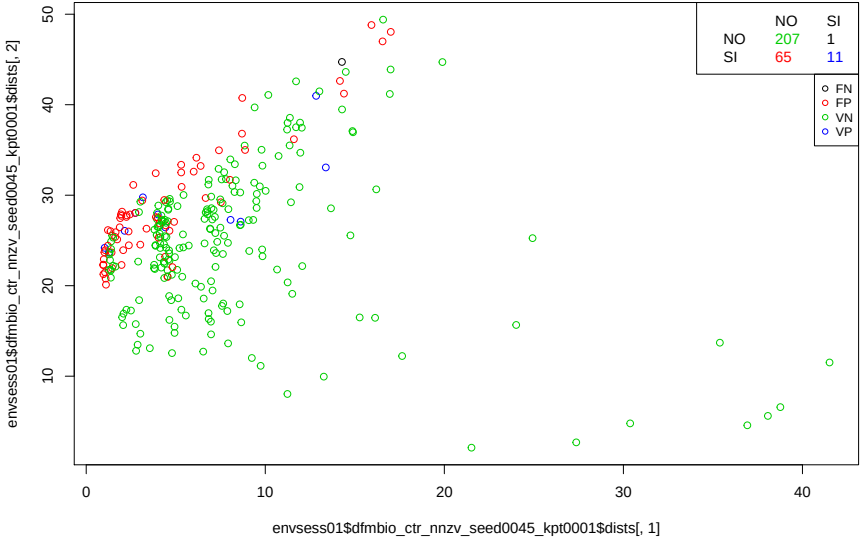






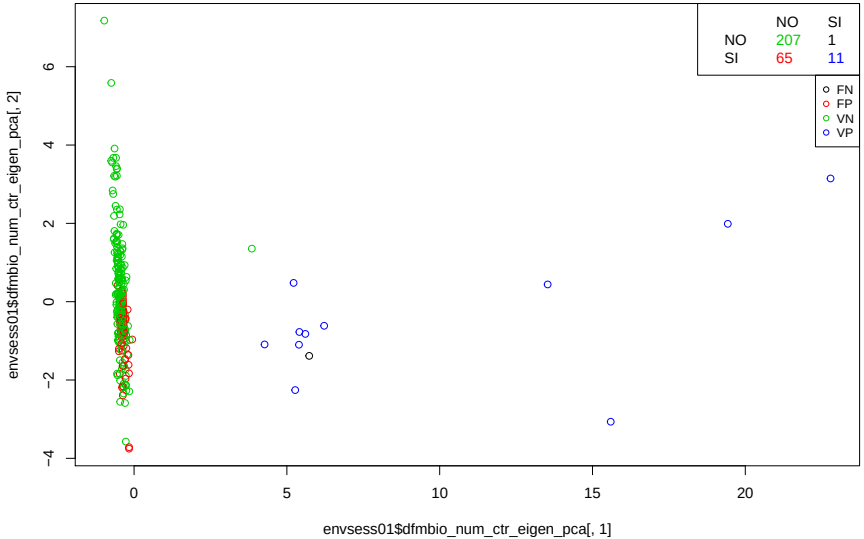
Identificación de factores de riesgo en pacientes de diabetes post pabellón.

Matías F. Rebolledo G.



Identificación de factores de riesgo en pacientes de diabetes post pabellón.

Matías F. Rebolledo G.



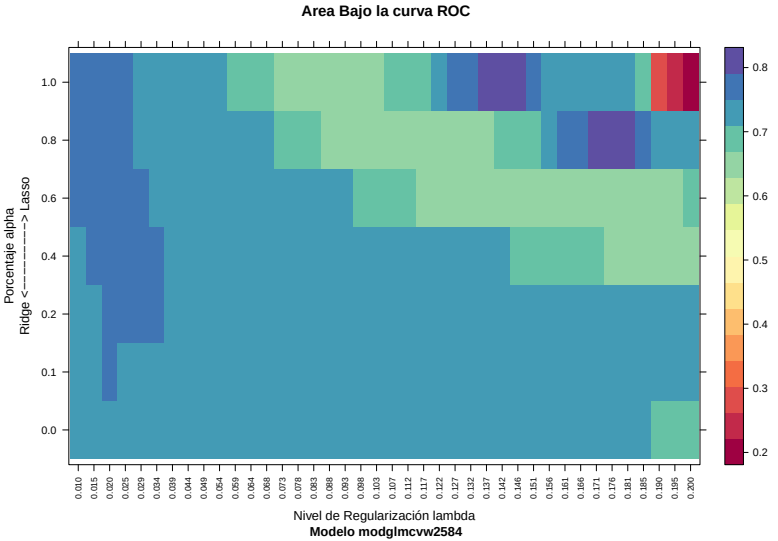
Identificación de factores de riesgo en pacientes de diabetes post pabellón.

Matías F. Rebolledo G.

Modelo logístico penalizado:

- Mejoras:
  - Se obtuvieron predicciones y métricas para 70 modelos
  - El modelo puede ser ridge ($\alpha = 0$), lasso ($\alpha = 1$) o elasticnet ($0 <= \alpha <= 1$)
  - Se usó el valor λ mínimo para generar el vector de predicciones y controlar el grado de penalización
- Por hacer:
  - Configurar valores de corte óptimo de probabilidad donde se maximice la sensitividad y la especificidad en conjunto, así como valores de α y λ

Formula	Dev	Acc	Sens	Spec
frm0037	NA	0.9718310	0.4166667	0.9963235
frm0039	NA	0.9718310	0.4166667	0.9963235
frm0040	NA	0.9718310	0.4166667	0.9963235
frm0061	NA	0.9718310	0.4166667	0.9963235
frm0063	NA	0.9718310	0.4166667	0.9963235
frm0064	NA	0.9718310	0.4166667	0.9963235
frm0004	NA	0.9683099	0.3333333	0.9963235
frm0006	NA	0.9683099	0.3333333	0.9963235
frm0007	NA	0.9683099	0.3333333	0.9963235
frm0019	NA	0.9683099	0.3333333	0.9963235
frm0021	NA	0.9683099	0.3333333	0.9963235
frm0022	NA	0.9683099	0.3333333	0.9963235
frm0031	NA	0.9683099	0.3333333	0.9963235
frm0034	NA	0.9683099	0.3333333	0.9963235
frm0043	NA	0.9683099	0.3333333	0.9963235



Identificación de factores de riesgo en pacientes de diabetes post pabellón.

Matías F. Rebolledo G.

Algunas sugerencias para mantener bajo el nivel de RAM:

- Grabar [parte de] la sesión como bases rdb/rdx indexadas

```
entorno_01 <- new.env()
assign( "modelo_01" , modelo_01 , envir = entorno_01 )
assign( "formula_01" , formula_01 , envir = entorno_01 )
assign( "objeto_01" , objeto_01 , envir = entorno_01 )
tools:::makeLazyLoadDB( entorno_01 , "rdb/entorno_01" )
```

- Se pueden cargar a su propio entorno

```
entorno_01 <- new.env()
base::lazyLoad( "rdb/entorno_01" , envir = entorno_01 )
assign( "modelo_01" , entorno_01$modelo_01 , envir = .GlobalEnv )
```

- Se pueden cargar al entorno global

```
entorno_01 <- new.env()
base::lazyLoad( "rdb/entorno_01" , envir = .GlobalEnv )
```

- Existe la función externa R_lazyLoadDBinsertValue

```
R CMD SHLIB serialize.c
dyn.load( "serialize.so" )
.Call( "R_lazyLoadDBinsertValue" , "rdb/entorno_01" , objeto_02 )
```

Nuevo codigo