Decisión en clasificación binaria

Resultado del proceso de predicción: ¿qué haremos con nuestras predicciones en un caso real?

	observacion	prob
2902	2902	0.002253886
1685	1685	0.191795764
1093	1093	0.675927032
1112	1112	0.607122317
728	728	0.860130218
1636	1636	0.241558307
937	937	0.880905823
840	840	0.804763614
1987	1987	0.062973749
104	104	0.992507514
2521	2521	0.019578365
107	107	0.990536413
2870	2870	0.001515387
1438	1438	0.457457862
2149	2149	0.020783208

Conocimiento del contexto en el que trabajamos

1) campaña de emailing o llamadas telefónicas a observaciones con alta probabilidad de y=1.

2) app móvil para clasificar imágenes

3) Triaje en un ingreso hospitalario

4) Se estudia si una reclamación de seguros es fraudulenta.

5) En un programa de antivirus se marca un mensaje como spam o no, derivándolo a otras carpetas o acciones sucesivas.

6) En medicina, dependiendo de la probabilidad de y=1, se realiza un diagnóstico concreto (clasificación dura) o bien se pasa a realizar otras pruebas.

7) Se predice un resultado en tenis, y las apuestas son uno a uno (si se acierta, se obtiene exactamente la cantidad que se apuesta).

La decisión del punto de corte es una **decisión** del investigador

Esta decision tiene que tener en cuenta la **estructura numérica del problema**:

muestra A: aquella con la que se contruye el modelo

muestra B: futura muestra desconocida sobre la cual aplicaremos el modelo en la práctica

Ejemplo 1:

Aplicación masiva PCR en aeropuertos:

- incidencia de 1/1000, en 50.000 individuos habría 50 positivos
- PCR tiene especificidad de 98%, sensitividad 85%
- Se detectarían 43 positivos y se darían 999 falsos positivos y 7 falsos negativos.

Ejemplo 2:

Test antígenos aplicado en residencias:

- El test de antígenos tiene sensitividad (detección de 1), de un 93% pero en pacientes que llevan más de 5 días de evolución desde el inicio de los síntomas produce demasiados falsos negativos (baja la sensitividad).
- En estos ultimos casos habría que hacer PCR, no test de antígenos SARS.
- Realmente al final se ha llegado a la conclusion de que en residencias, si el test de antígenos da negativo, es prudente hacer un test PCR posterior.

Aspectos Relativos a la muestra A:

1) El número de observaciones con y=1 en la muestra
2) El número de observaciones total en la muestra
3) Aspectos relativos al desempeño del algoritmo predictivo:
    3.1 Número de variables input
    3.2 Auc obtenido en validación cruzada; Accuracy o tasa de fallos obtenida en validación cruzada
    3.3 Número de observaciones clasificadas como y=1 con punto de corte básico 0.5
    3.4 Sensitividad, especificidad con punto de corte básico 0.5, FP,FN
4) Porcentaje de y=1 en la muestra

Aspectos Relativos a la futura muestra B:

5) Si se conoce, número de observaciones a predecir en una muestra B, en la aplicación práctica futura del algoritmo. Puede ser una cifra aproximada.
El número aproximado de observaciones y=1 que se espera en esta muestra B también es una cifra importante.
En algunos contextos las predicciones no se van a aplicar sobre una muestra B de casos, sino en observaciones aisladas una a una. Entonces no aplica lo dicho en este punto 5).

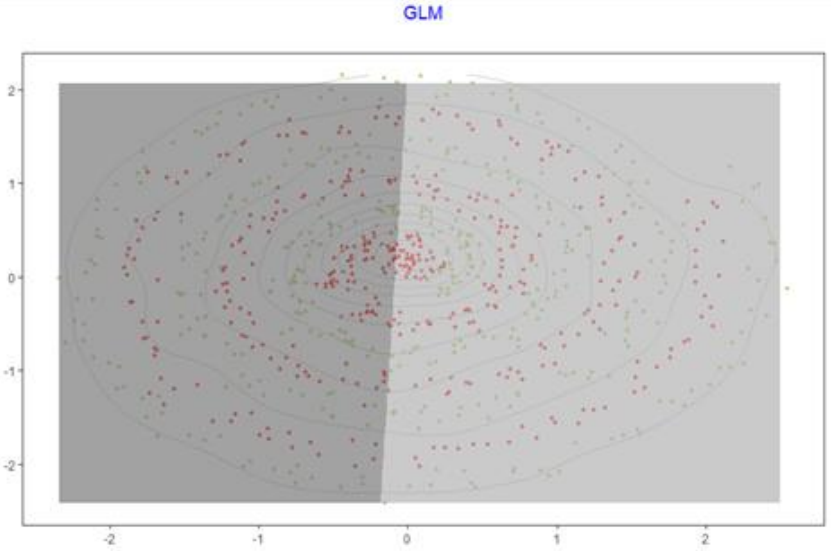
Ejemplos artificiales básicos

Archivo graficos tema decision.R

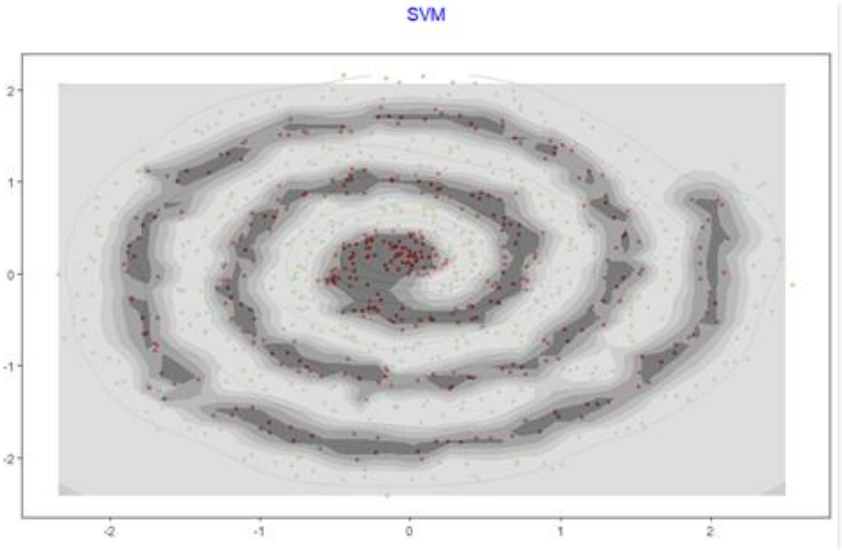
Sensitividad=Probabilidad de que la predicción sea 1, dado que la observación es realmente 1= capacidad de detectar positivos =VP/(VP+FN)=0.65. También se le puede llamar ***recall***.
Especificidad=Probabilidad de que la predicción sea 0, dado que la observación es realmente 0=capacidad de detectar negativos=VN/(VN+FP)=0.80

1

2



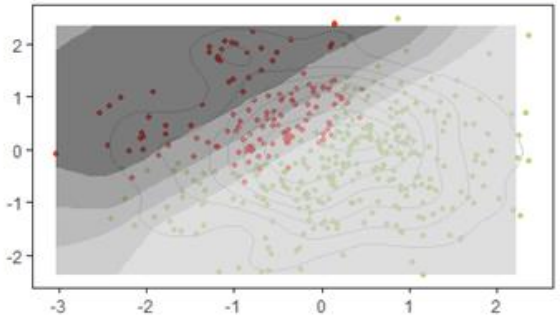
obs: 803
obs clase minoritaria: 401
% clase minoritaria: 50 %
tasa fallos: 0.5554172
auc: 0.5670897
sensitivity: 0.3715711
specificity: 0.5174129
numero de Yes predichos: 343
FP: 194
FN: 252
VP: 149
VN: 208



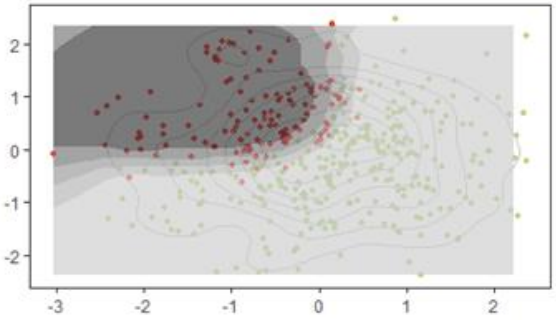
obs: 803
obs clase minoritaria: 401
% clase minoritaria: 50 %
tasa fallos: 0.06351183
auc: 0.9697708
sensitivity: 0.9376559
specificity: 0.9353234
numero de Yes predichos: 402
FP: 26
FN: 25
VP: 376
VN: 376

3

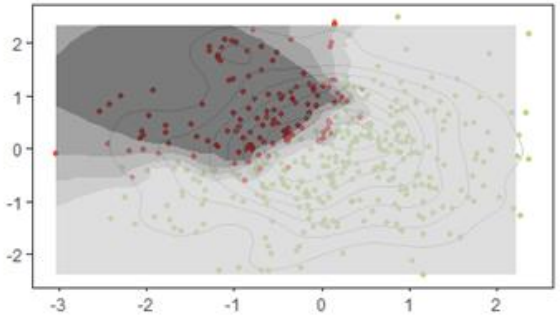
GLM



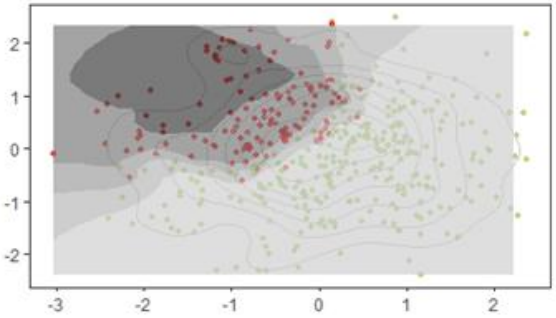
NNET



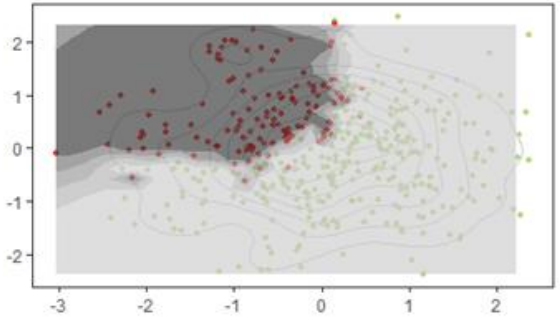
RF



GBM



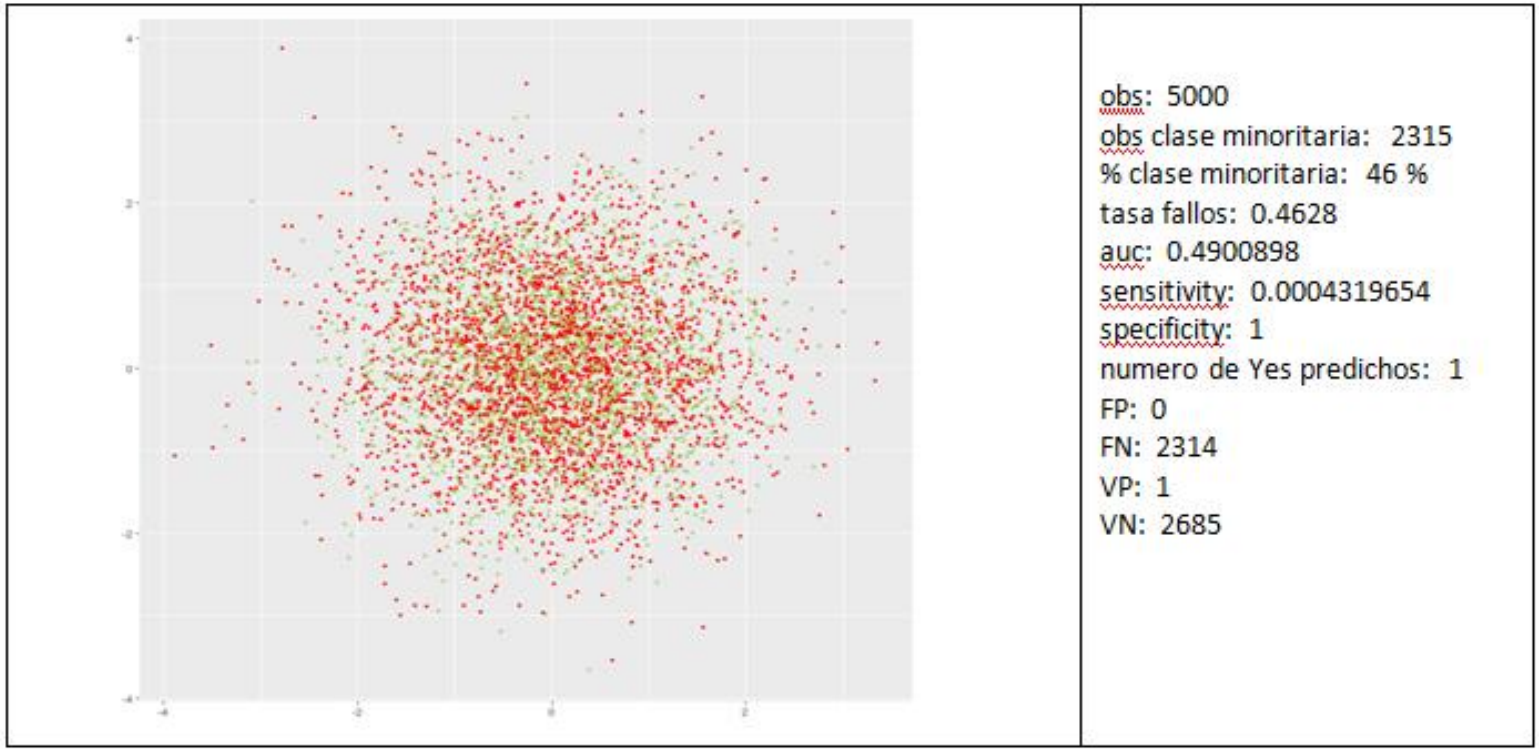
SVM



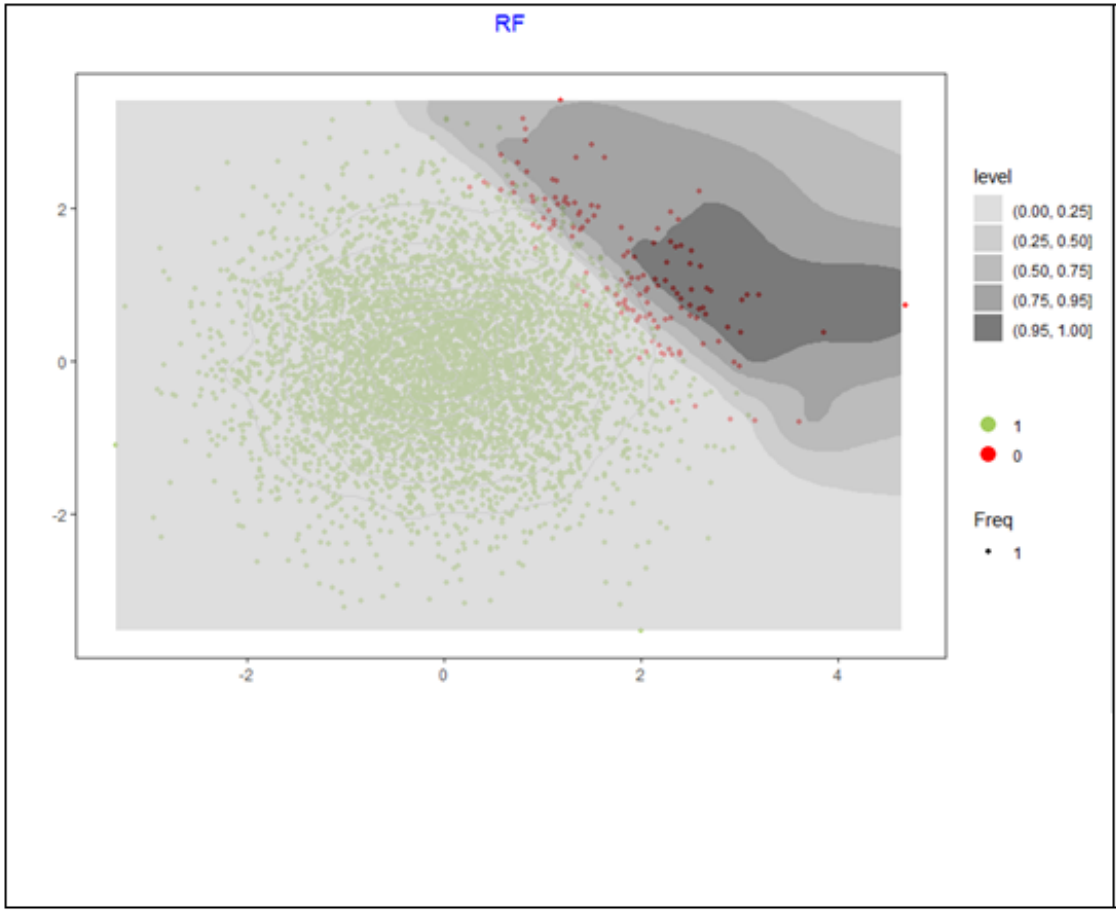
Recordatorio

Sensitividad=Probabilidad de que la predicción sea 1, dado que la observación es realmente 1= capacidad de detectar positivos =VP/(VP+FN)=0.65. También se le puede llamar ***recall***.
Especificidad=Probabilidad de que la predicción sea 0, dado que la observación es realmente 0=capacidad de detectar negativos=VN/(VN+FP)=0.80

4



5

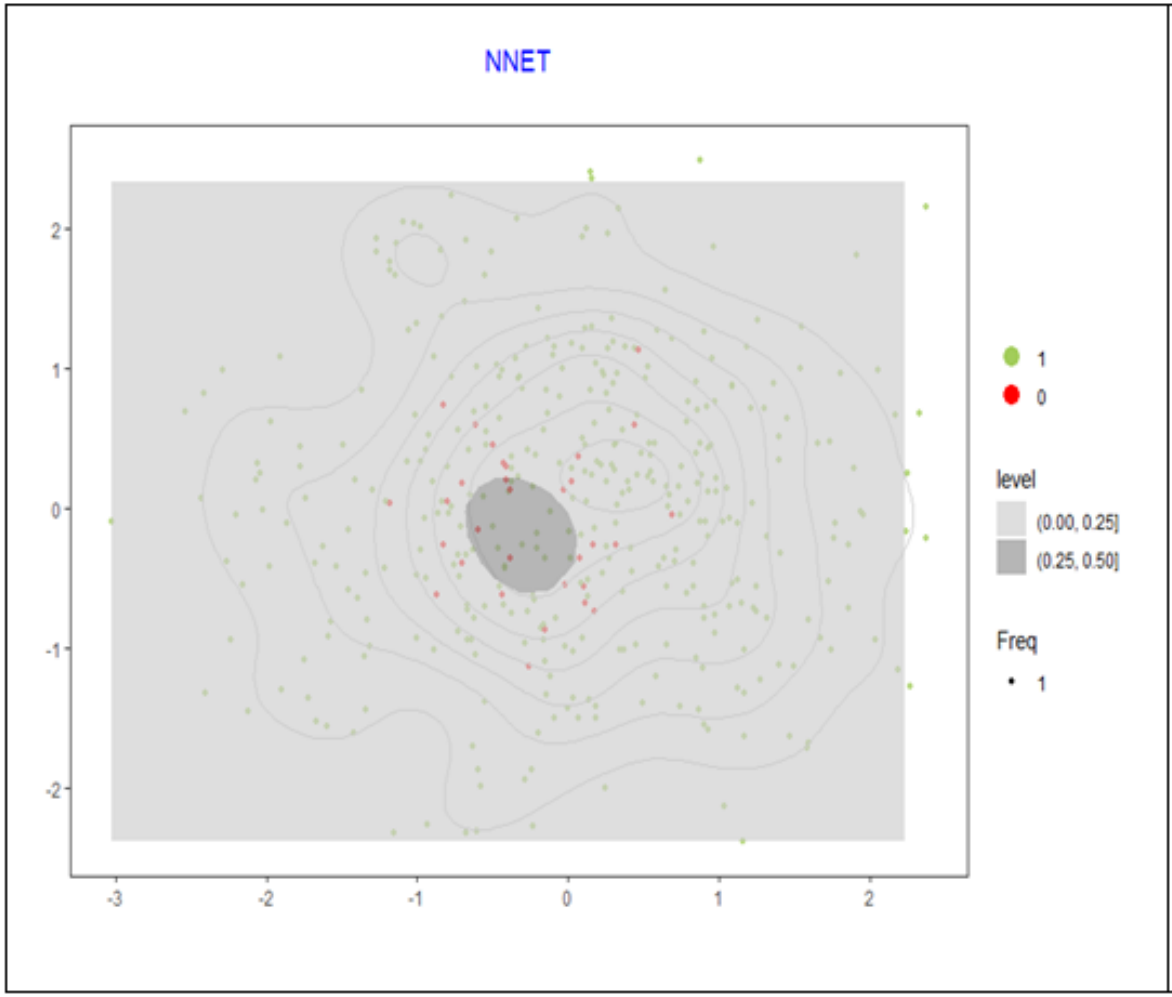


obs: 5000
obs clase minoritaria: 138
% clase minoritaria: 3 %
tasa fallos: 0.0136
auc: 0.987026
sensitivity: 0.7028986
specificity: 0.9944467
numero de Yes predichos: 124
FP: 27
FN: 41
VP: 97
VN: 4835

corte 0.4

obs: 5000
obs clase minoritaria: 138
% clase minoritaria: 3 %
tasa fallos: 0.0138
auc: 0.987026
sensitivity: 0.7681159
specificity: 0.99239
numero de Yes predichos: 143
FP: 37
FN: 32
VP: 106
VN: 4825

7



obs: 400
obs clase minoritaria: 31
% clase minoritaria: 8 %
tasa fallos: 0.0775
auc: 0.8009441
sensitivity: 0
specificity: 1
numero de Yes predichos: 0
FP: 0
FN: 31
VP: 0
VN: 369

corte 0.2

obs: 400
obs clase minoritaria: 31
% clase minoritaria: 8 %
tasa fallos: 0.1775
auc: 0.8009441
sensitivity: 0.4516129
specificity: 0.8536585
numero de Yes predichos: 68
FP: 54
FN: 17
VP: 14
VN: 315

Paquete visualpred

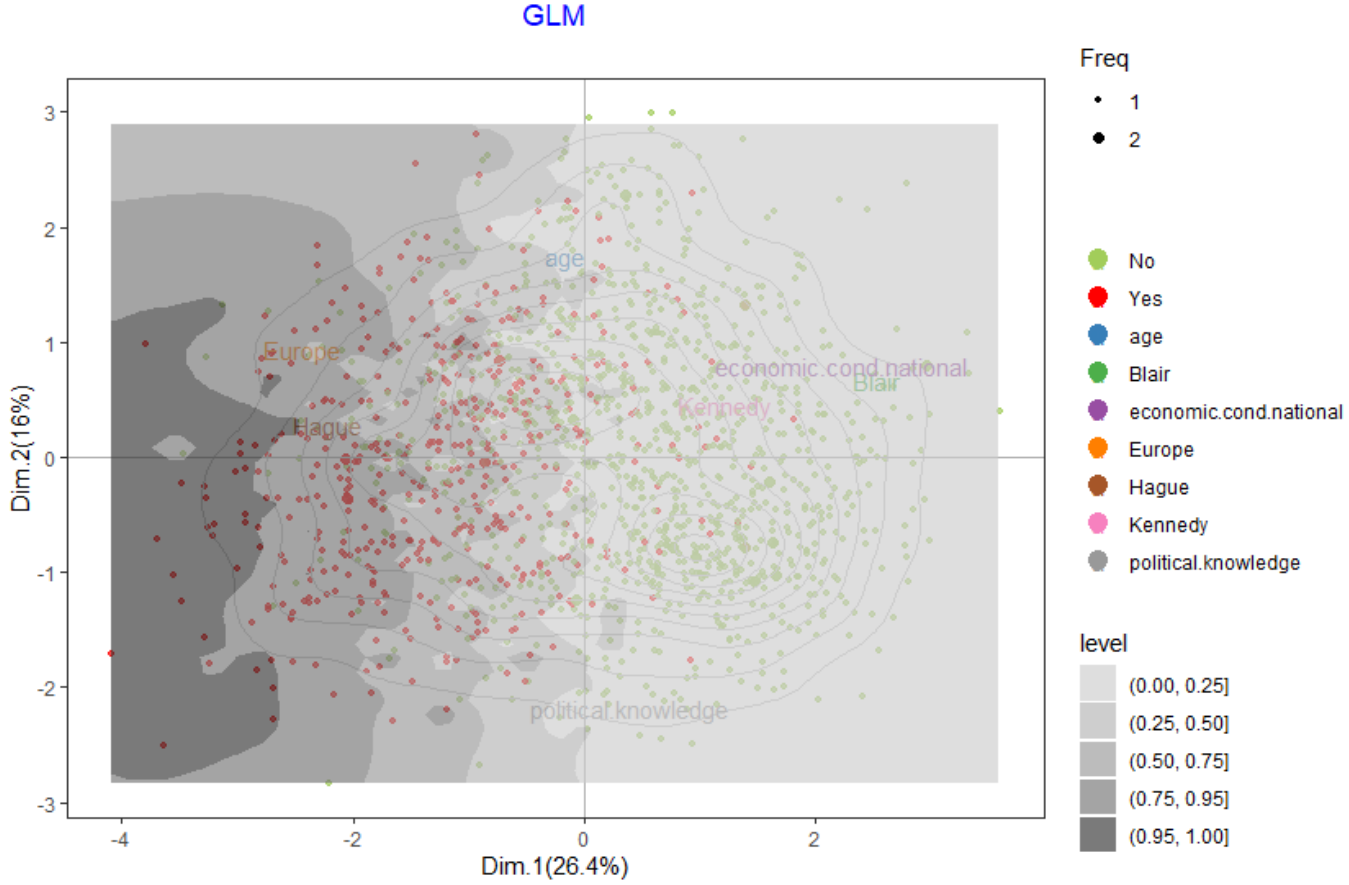
1) MCA o FAMD para reducir el espacio de variables input en factores

2) Se presentan las observaciones sobre los dos primeros factores coloreadas según la variable objetivo

3) Se aplican algoritmos predictivos que dan lugar a una predicción de probabilidad para cada observación. Permite glm, rf,gbmy svm radial.

4) Se construye una rejilla de valores en el rango de los valores que toman los dos primeros factores, y se da valores a esa rejilla por interpolación, tomando como referencia las predicciones de probabilidad obtenidas en 3.

5) Con los valores interpolados de probabilidades obtenidas en 4 para esa rejilla, se superpone un gráfico de contour sobre los gráficos básicos de puntos.



Información y descarga:

https://cran.r-project.org/web/packages/visualpred/index.html

Vignettes=ejemplos

https://cran.r-project.org/web/packages/visualpred/vignettes/Basic_example.html
https://cran.r-project.org/web/packages/visualpred/vignettes/Comparing.html
https://cran.r-project.org/web/packages/visualpred/vignettes/Outliers.html
https://cran.r-project.org/web/packages/visualpred/vignettes/Advanced.html

Manual:

https://cran.r-project.org/web/packages/visualpred/visualpred.pdf

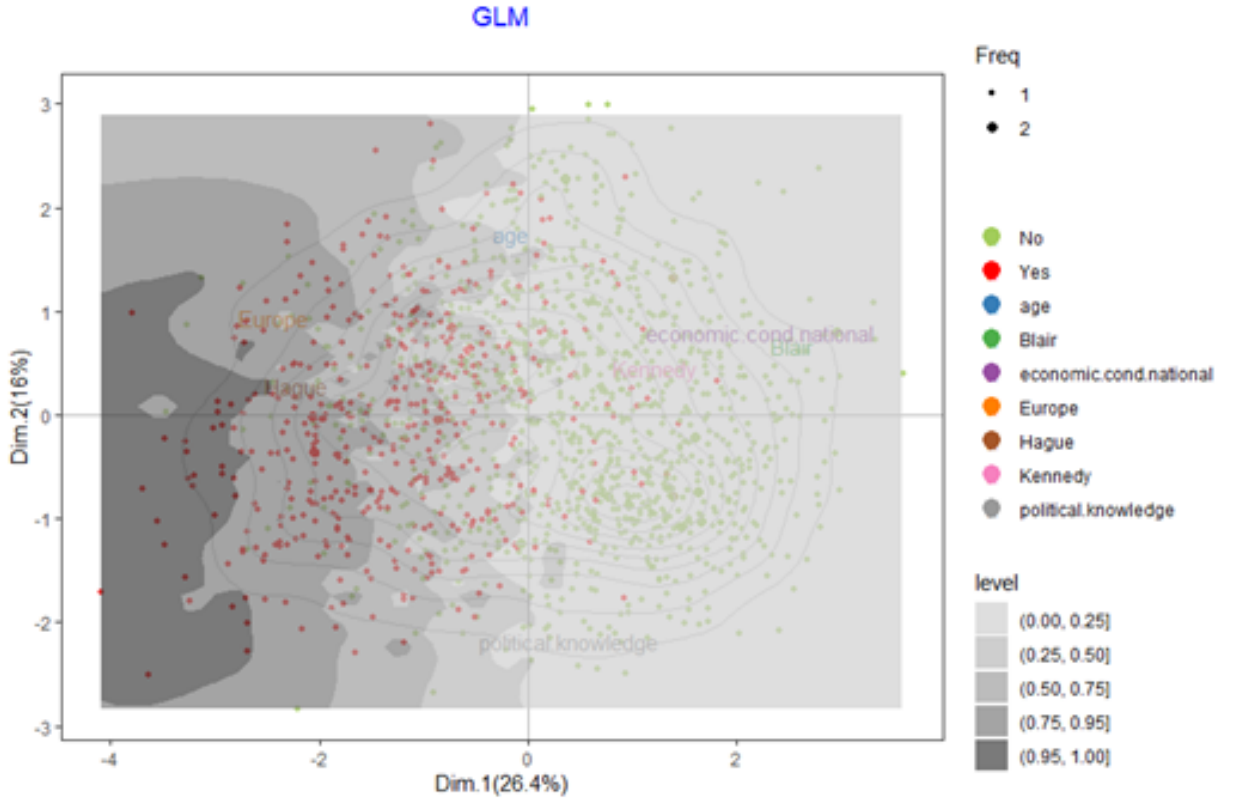
Ejemplo sintaxis

```
library(visualpred)
load("BEPS2.Rda")
archivo<-BEPS
listconti<-c("age", "economic.cond.national", "economic.cond.household",
             "Blair", "Hague", "Kennedy", "Europe", "political.knowledge")
listclass<-c("gender")
vardep<-"y"

result<-famdcontour(dataf=archivo,listconti=listconti,listclass=listclass,vardep=vardep,
                    title="rf",title2="
",Dime1="Dim.1",Dime2="Dim.2",selec=1,modelo="rf",classvar=0,mtry=5)
result[[2]]
result[[3]]
result[[4]]
result[[4]]
result[[5]]
result[[6]]
```

Ejemplos datos reales

BEPS

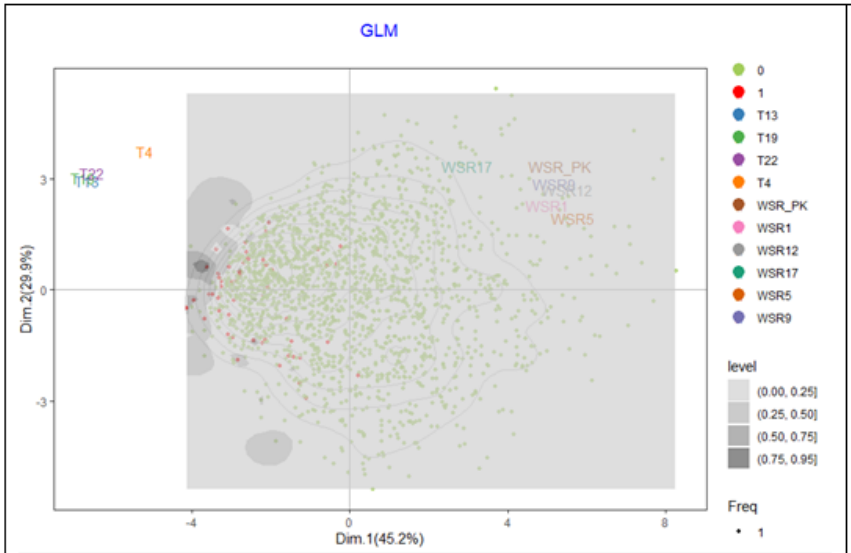


obs: 1525
obs clase minoritaria: 462
% clase minoritaria: 30 %
tasa fallos: 0.1606557
auc: 0.8882217
sensitivity: 0.6709957
specificity: 0.9125118
numero de Yes predichos: 403
FP: 93
FN: 152
VP: 310
VN: 970

corte 0.4

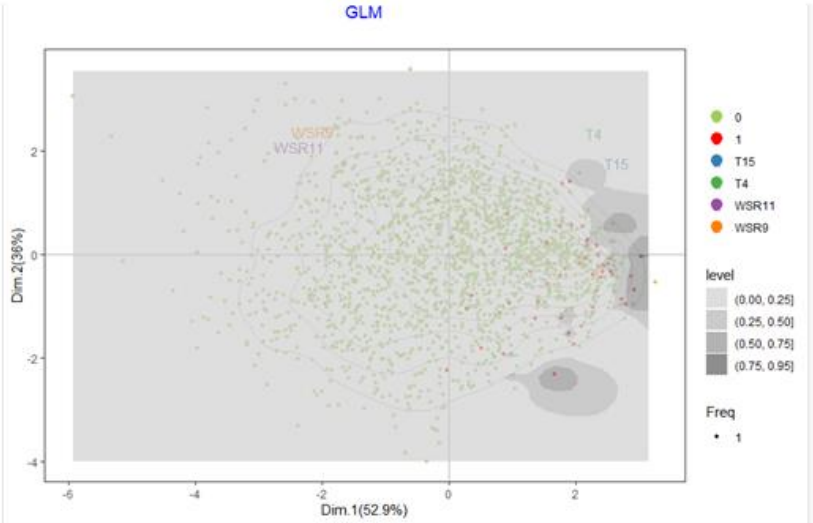
obs: 1525
obs clase minoritaria: 462
% clase minoritaria: 30 %
tasa fallos: 0.16
auc: 0.8882217
sensitivity: 0.7575758
specificity: 0.8758231
numero de Yes predichos: 482
FP: 132
FN: 112
VP: 350
VN: 931

OZONO



obs: 1848
obs clase minoritaria: 57
% clase minoritaria: 3 %
tasa fallos: 0.02922078
auc: 0.9295307
sensitivity: 0.1929825
specificity: 0.9955332
numero de Yes predichos: 19
FP: 8
FN: 46
VP: 11
VN: 1783

10 variables



obs: 1848
obs clase minoritaria: 57
% clase minoritaria: 3 %
tasa fallos: 0.02922078
auc: 0.9207735
sensitivity: 0.1052632
specificity: 0.998325
numero de Yes predichos: 9
FP: 3
FN: 51
VP: 6
VN: 1788

corte=0.03

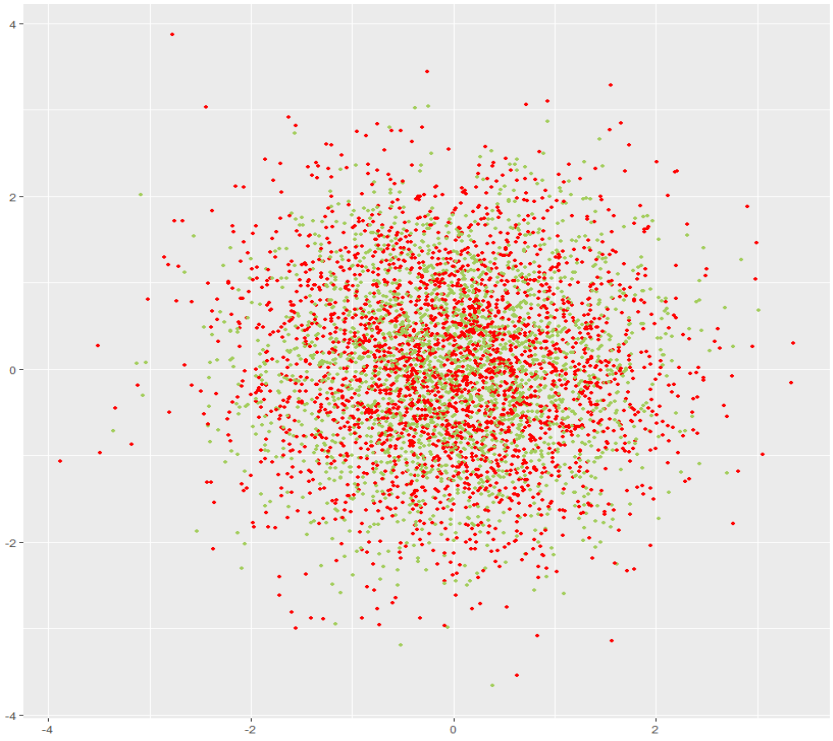
obs: 1848
obs clase minoritaria: 57
% clase minoritaria: 3 %
tasa fallos: 0.1774892
auc: 0.9207735
sensitivity: 0.8421053
specificity: 0.8218872
numero de Yes predichos: 367
FP: 319
FN: 9
VP: 48
VN: 1472

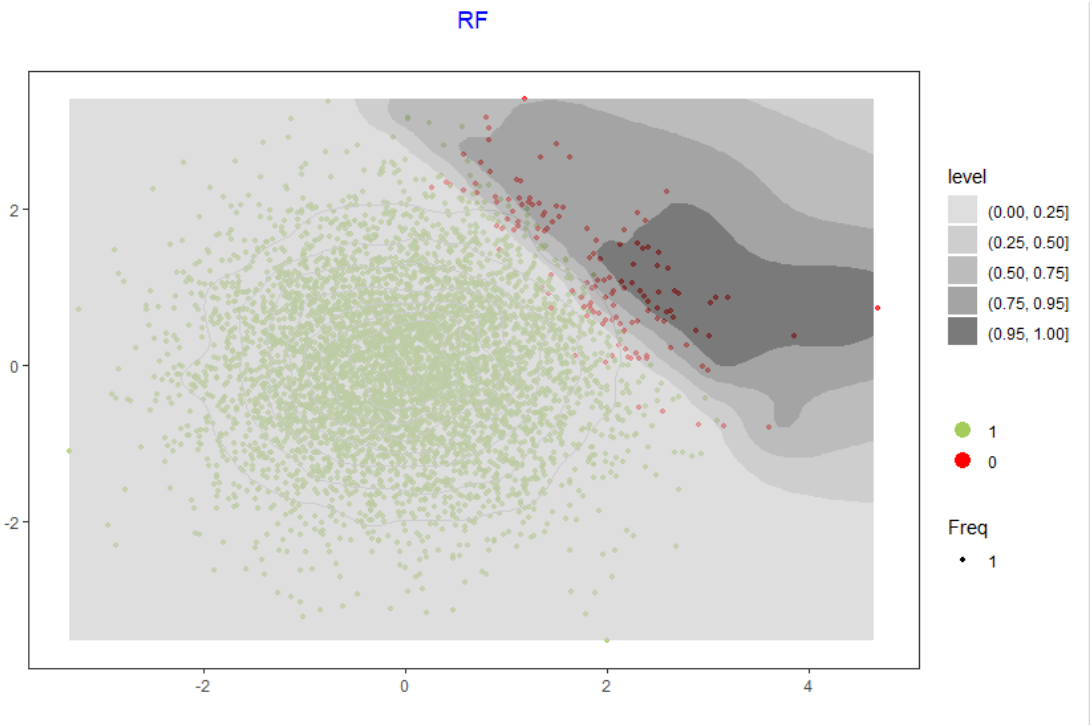
4 variables

La subcultura del “imbalanced dataset”

(sección de opinión)

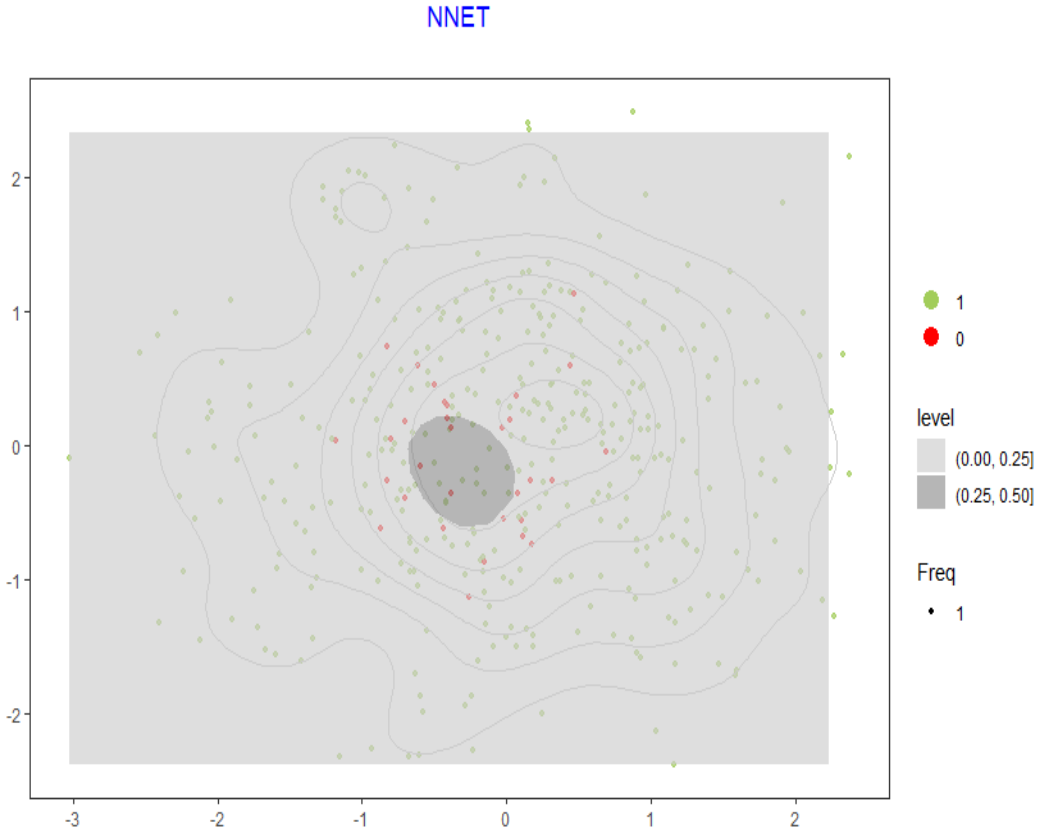
- Izquierda: Clases equilibradas pero mala separabilidad
- Derecha: Clase desequilibradas pero buena separabilidad

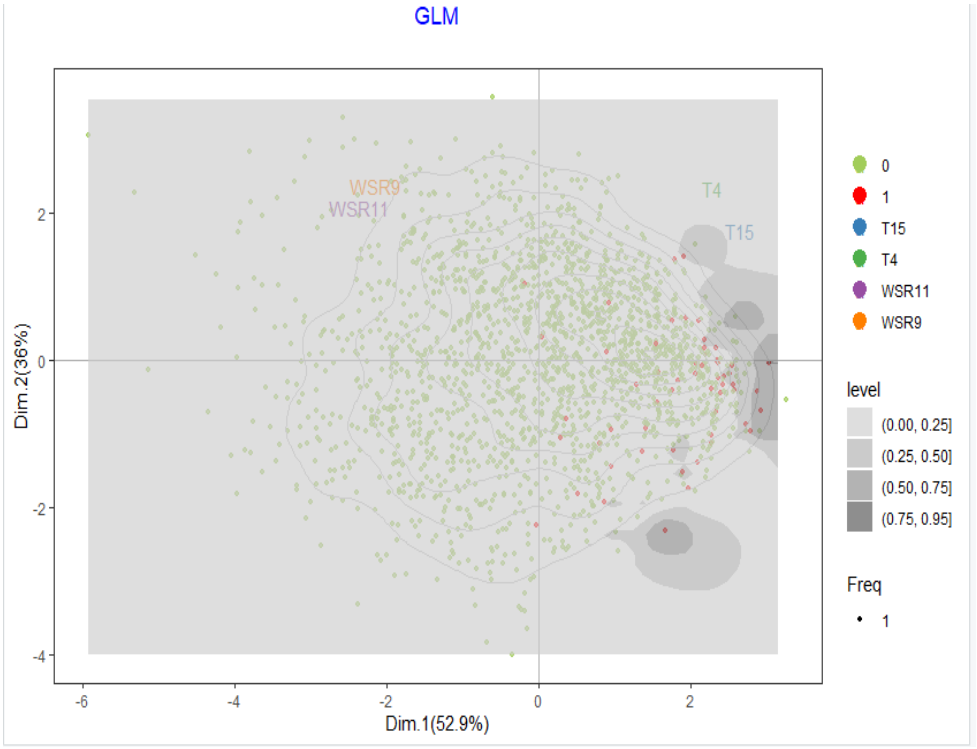




- Los algoritmos detectan bien las regiones y estiman correctamente las probabilidades, tanto en datos desequilibrados como que en todo tipo de datos.

- Es la **presencia o no de variables input útiles** (predictivas) la que determina la buena o mala separabilidad





Undersampling =quedarse con solo una porción de la clase mayoritaria con el objetivo de equilibrar las clases .

En mi opinión, es una estrategia equivocada:

- No se obtienen buenos estimadores de las medidas de diagnóstico y el error puede estar subestimado; las conclusiones no se pueden extrapolar directamente a la población al haber alterado la muestra
- Las regiones detectadas y las probabilidades estimadas no coinciden con la realidad
- Los algoritmos no funcionan “mejor”, funcionan sobre una muestra diferente

"Oversampling" (SMOTE y otros)= generar observaciones artificiales en la clase minoritaria para equilibrar las clases.

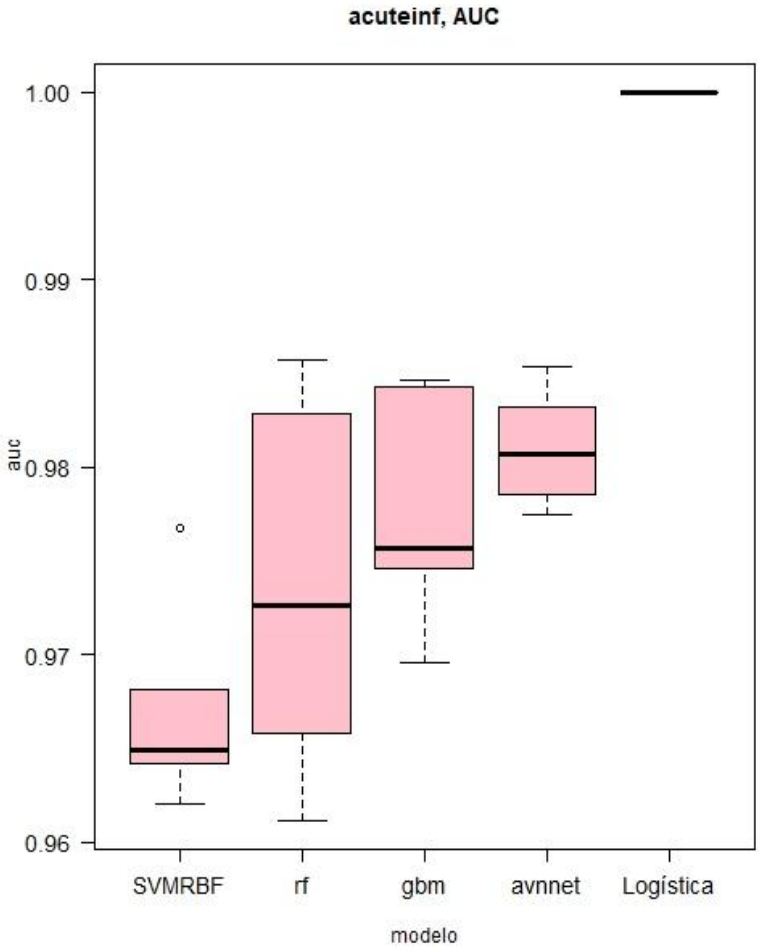
En mi opinión, también es una estrategia equivocada por las mismas razones anteriores:

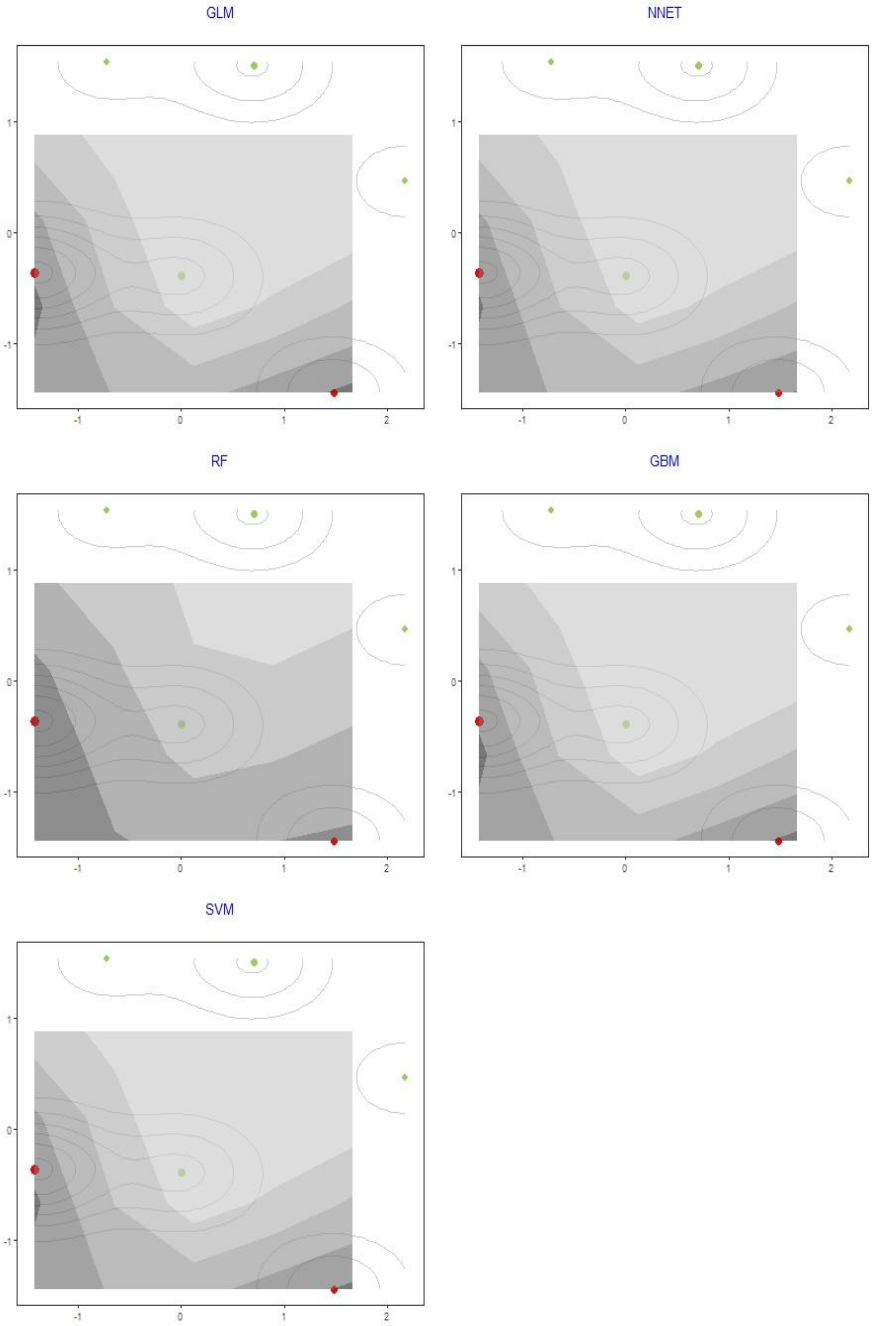
- No se obtienen buenos estimadores de las medidas de diagnóstico y el error puede estar subestimado; las conclusiones no se pueden extrapolar directamente a la población al haber alterado la muestra
- Las regiones detectadas y las probabilidades estimadas no coinciden con la realidad
- Los algoritmos no funcionan “mejor”, funcionan sobre una muestra diferente

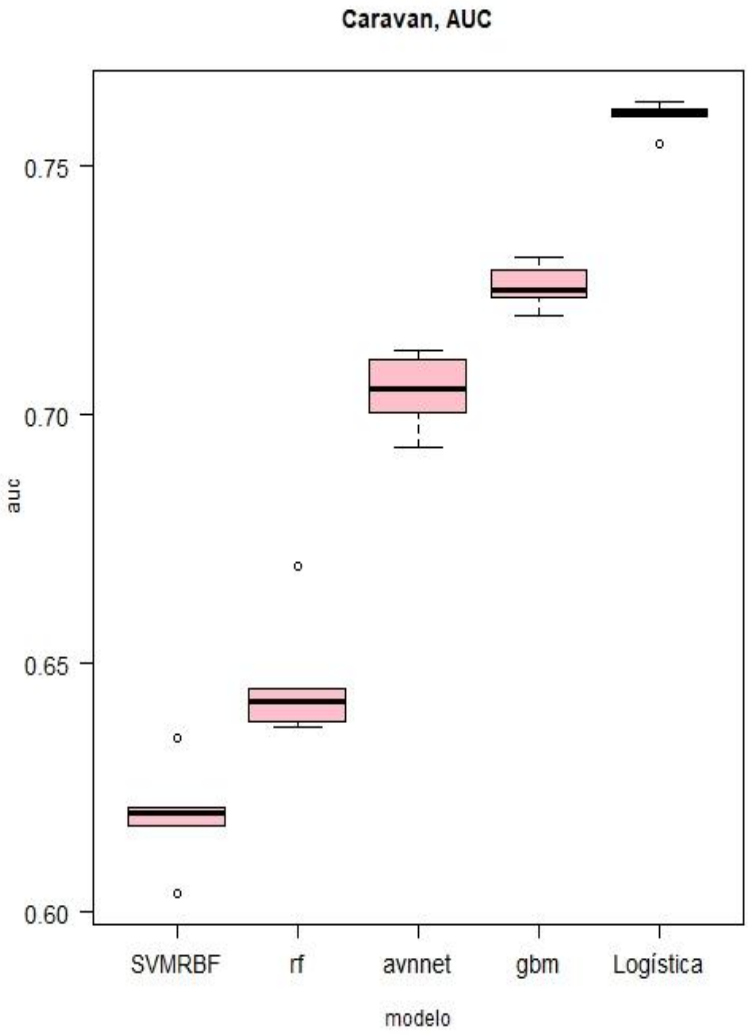
Comportamiento comparativo de algoritmos

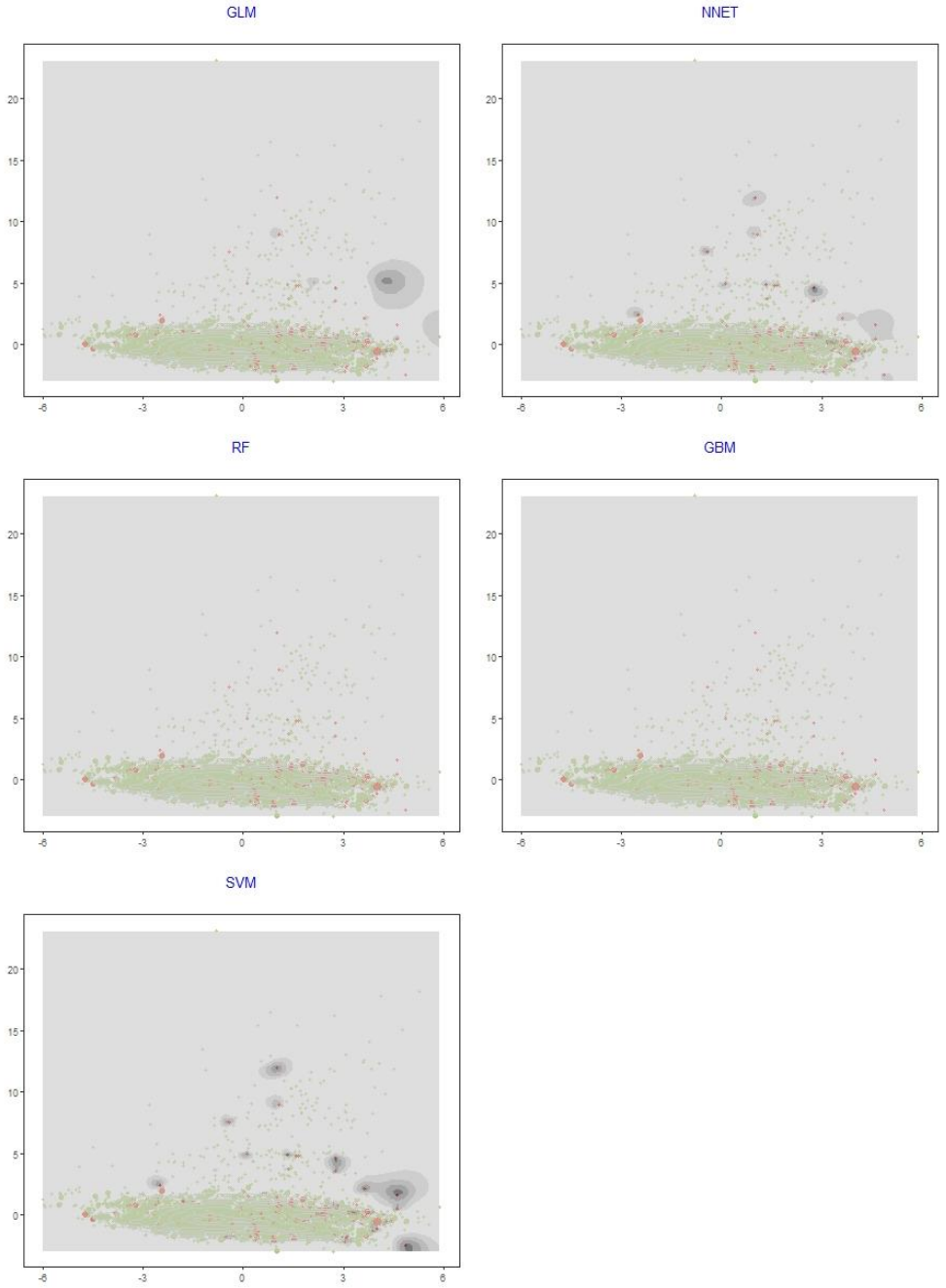
Archivo comparaciones básicas.R

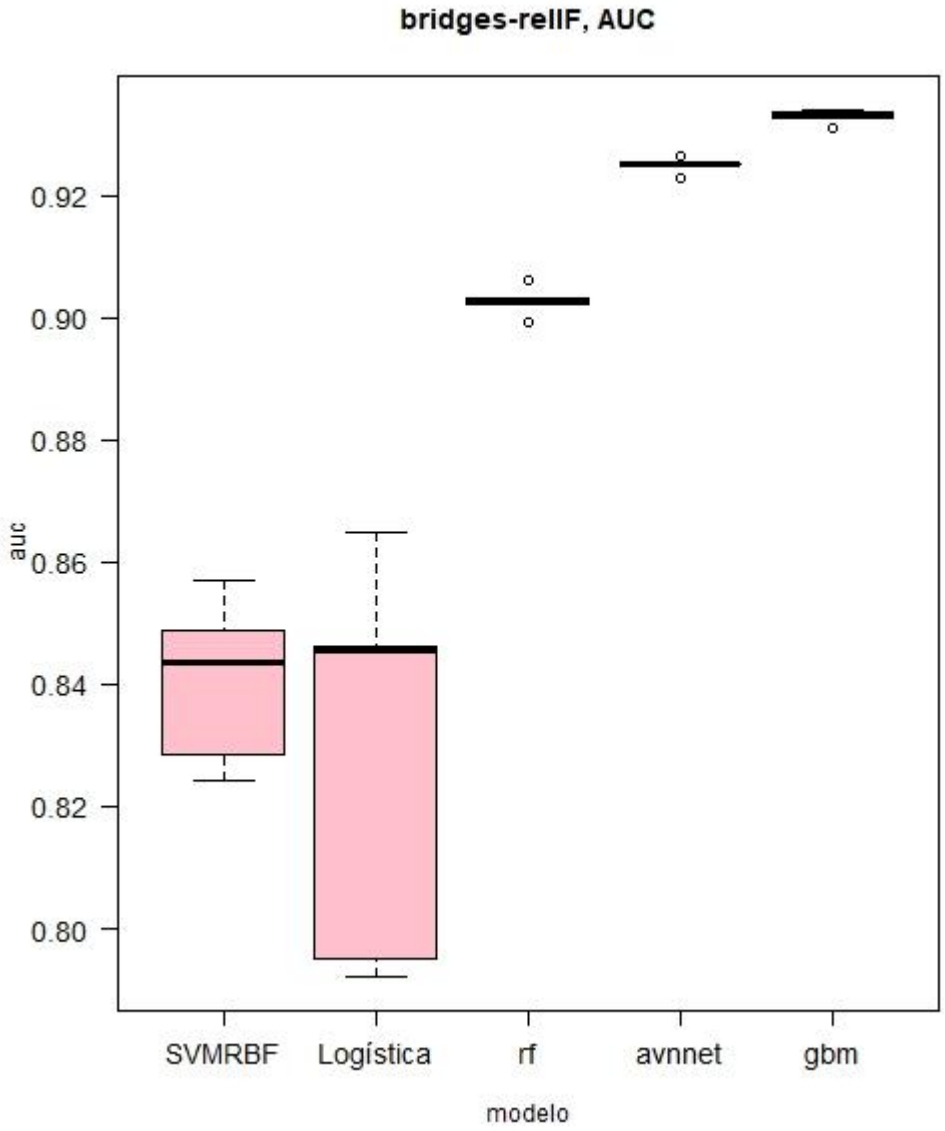
```
# listconti<-c("a1")
# listclass<-c("a2","a3","a4","a5","a6")
```

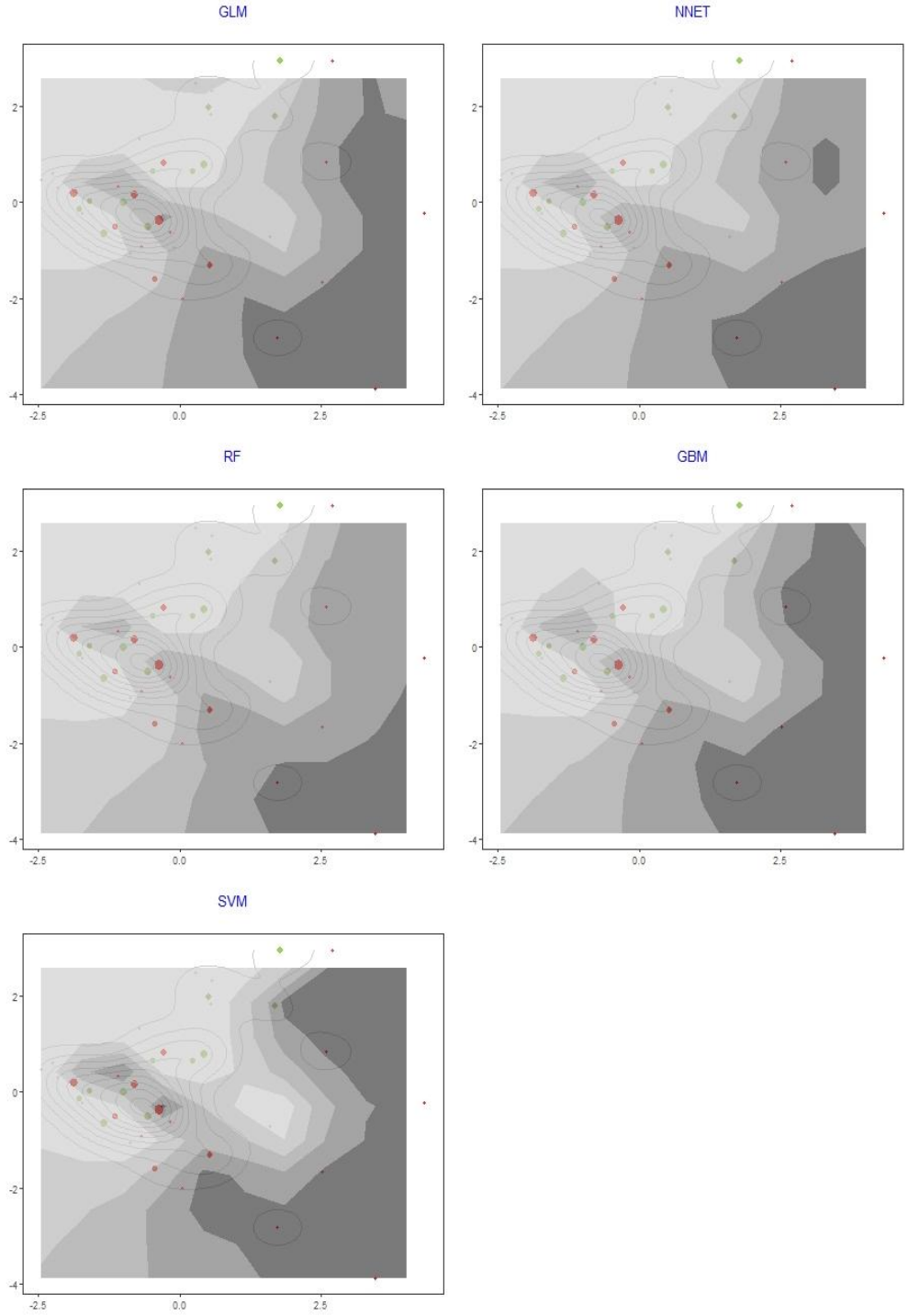


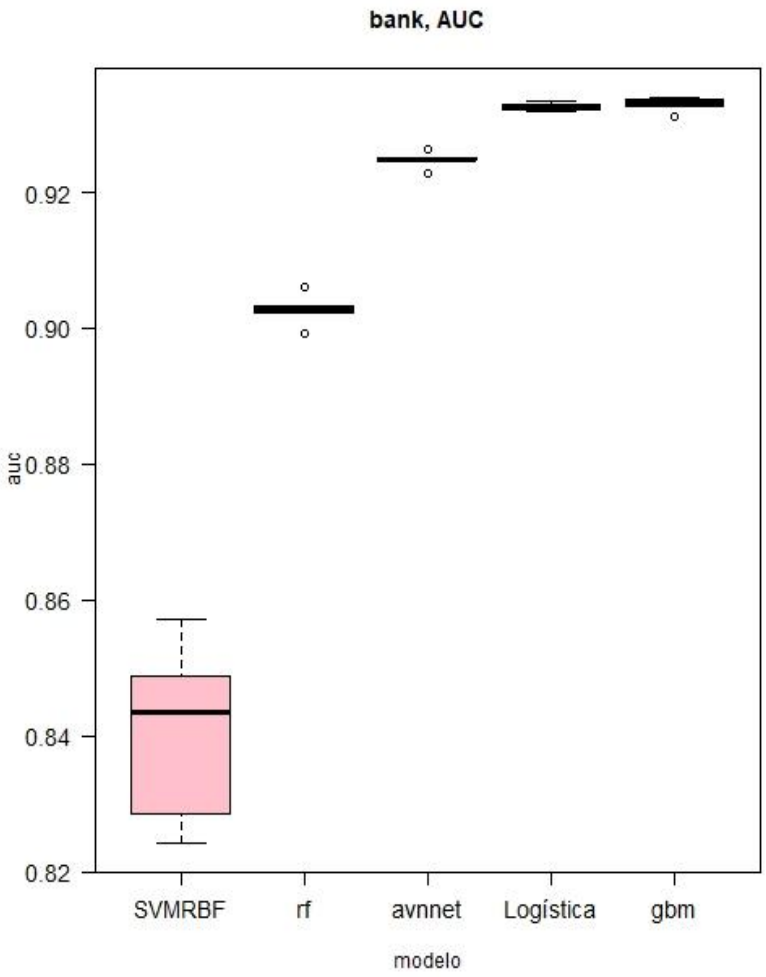


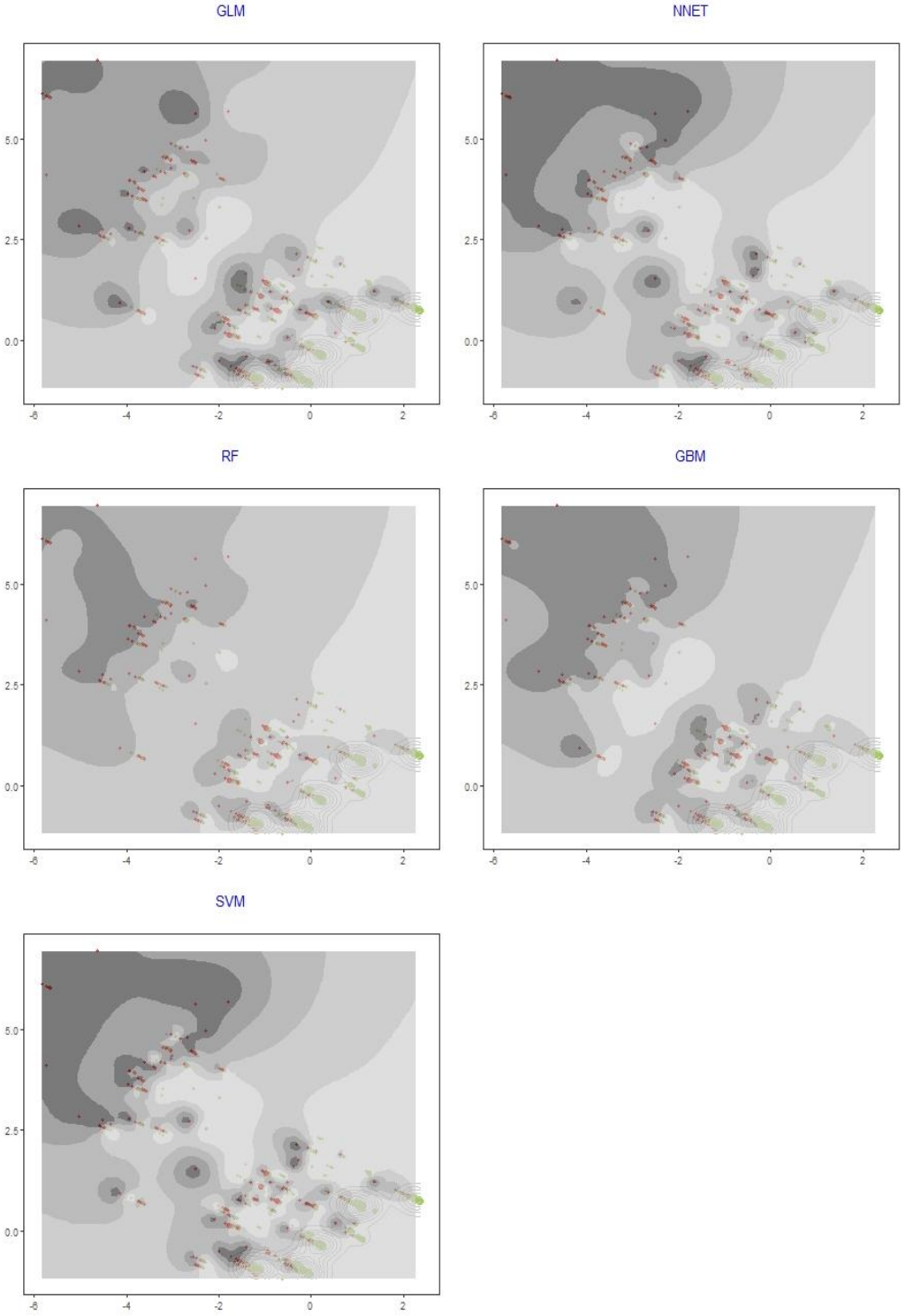

Caravan, AUC
0.75
0.70
0.65
0.60
auc
SVMRBF
rf
avnnet
gbm
Logística
modelo


GLM
NNET
RF
GBM
SVM


bridges-relIF, AUC
0.92
0.90
0.88
0.86
0.84
0.82
0.80
auc
SVMRBF
Logística
rf
avnnet
gbm
modelo


GLM
NNET
RF
GBM
SVM


bank, AUC
auc
0.92
0.90
0.88
0.86
0.84
0.82
SVMRBF
rf
avnnet
Logística
gbm
modelo


GLM
NNET
RF
GBM
SVM