Actividad - Método Supervisados

Javier Eduardo Jaimes Velasquez[[1]](#footnote-1) Brandon Valencia Murillo[[2]](#footnote-2)

2024-07-11

Tabla de Contenido

[1.1 Requerimientos iniciales 2](#_Toc171607879)

[1.2 Conjunto de Datos 2](#_Toc171607880)

[1.3 Definicion de Modelos 3](#_Toc171607881)

[1.4 Comparacion de Modelos 4](#_Toc171607882)

[1.5 Conclusion 8](#_Toc171607883)

A continuación, se detalla a manera de informar las actividades solicitadas como parte del módulo de métodos supervisados. En el siguiente documento se busca dar una respuesta a las siguientes preguntas:

* Manipulación y carga del conjunto de datos *attrition*, parte de la librería *modeldata*.
* Definir un modelo de Random Forest. (RF) a partir de la columna *attrition*.
* Definir un modelo GLM, a partir del modelo y conjuntos de datos anterior mencionado.
* Definir un modelo Supervised Vector Machine (SVM) a partir del módulo y conjuntos de datos anterior mencionado.
* A partir de lo anterior: ¿Cual modelo sería el mejor?
* Obtener las matrices de confusión de cada modelo.

## 1.1 Requerimientos iniciales

Se cargan las librerias necesarias: *magrittr, factoextra, readr, tidyverse, caret, class, modeldata, glmnet, ranger*.

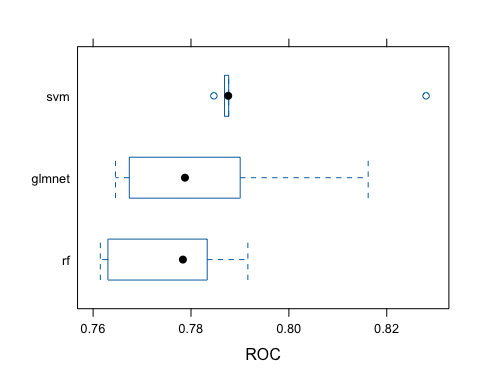
## 1.2 Conjunto de Datos

A continuacion se describe el conjunto de datos, **attrition**, utilizando el método data de R.

Al resumir los datos, validaremos que la columna *Attrition* sea de tipo factor. Además de las demás variables presente los modelos deberán predecir el número de clientes que dejarán de usar el servicio de la empresa.

## 1.3 Definición de Modelos

A continuación, se definen los siguientes modelos: RF, GLM y SVM. De acuerdo con como se explicó en las clases. Y se genera la gráfica de comparación de caja y bigotes en donde se aprecia que su comportamiento en términos generales es muy similar para los 3 modelos.



1.4.1 Matrix de Confuision por Modelo

A continuación se imprimen las matrices de confusion por cada uno de los modelos

1.4.1.1 Random Forest

En este modelo vemos una precision de con una desecersion de clientes del .

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 363 57  
## Yes 6 14  
##   
## Accuracy : 0.8568   
## 95% CI : (0.8206, 0.8882)  
## No Information Rate : 0.8386   
## P-Value [Acc > NIR] : 0.1657   
##   
## Kappa : 0.2548   
##   
## Mcnemar's Test P-Value : 2.988e-10   
##   
## Sensitivity : 0.9837   
## Specificity : 0.1972   
## Pos Pred Value : 0.8643   
## Neg Pred Value : 0.7000   
## Prevalence : 0.8386   
## Detection Rate : 0.8250   
## Detection Prevalence : 0.9545   
## Balanced Accuracy : 0.5905   
##   
## 'Positive' Class : No   
##

1.4.1.2 GLM

En este modelo vemos una precision de con una desecersion de clientes del .

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 362 50  
## Yes 7 21  
##   
## Accuracy : 0.8705   
## 95% CI : (0.8354, 0.9004)  
## No Information Rate : 0.8386   
## P-Value [Acc > NIR] : 0.03728   
##   
## Kappa : 0.3664   
##   
## Mcnemar's Test P-Value : 2.651e-08   
##   
## Sensitivity : 0.9810   
## Specificity : 0.2958   
## Pos Pred Value : 0.8786   
## Neg Pred Value : 0.7500   
## Prevalence : 0.8386   
## Detection Rate : 0.8227   
## Detection Prevalence : 0.9364   
## Balanced Accuracy : 0.6384   
##   
## 'Positive' Class : No   
##

1.4.1.3 SVM

En este modelo vemos una precision de con una desecersion de clientes del .

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 361 51  
## Yes 8 20  
##   
## Accuracy : 0.8659   
## 95% CI : (0.8305, 0.8963)  
## No Information Rate : 0.8386   
## P-Value [Acc > NIR] : 0.0655   
##   
## Kappa : 0.3442   
##   
## Mcnemar's Test P-Value : 4.553e-08   
##   
## Sensitivity : 0.9783   
## Specificity : 0.2817   
## Pos Pred Value : 0.8762   
## Neg Pred Value : 0.7143   
## Prevalence : 0.8386   
## Detection Rate : 0.8205   
## Detection Prevalence : 0.9364   
## Balanced Accuracy : 0.6300   
##   
## 'Positive' Class : No   
##

## 

## 1.5 Conclusión

Los modelos se comportan de forma muy similar, y su varianza en la precisión es muy poco haciendo que cualquier de los modelos una posible solución del problema. Tal vez se deberían integrar otras variables para determinar la elección, de variables que varían de acuerdo con el contexto en el cual se ejecutaran estos modelos.

1. Politécnico Grancolombiano, [jajaimes4@poligran.edu.co](mailto:jajaimes4@poligran.edu.co) [↑](#footnote-ref-1)
2. Politécnico Grancolombiano, [brvalencia6@poligran.edu.co](mailto:brvalencia6@poligran.edu.co) [↑](#footnote-ref-2)