Modelos de clasificación de personas ocupadas que cotizan pensión en Colombia

*MiguelContreras1/Trabajo-Final-Big-Data-Machine-Learning (github.com)

1st Juan Camilo Martinez
dept.de Economia
Universidad de los Andes
jc.martinezr@uniandes.edu.co

2nd Miguel Contreras
dept.de Economia
Universidad de los Andes
ja.trujillom@uniandes.edu.co

3nd Arturo Trujillo
dept.de Economia
Universidad de los Andes
ja.trujillom@uniandes.edu.co

4th Jesus Cepeda LLanes
dept.de Economia
Universidad de los Andes
jd.cepedal@uniandes.edu.co

***Abstract*—El presente artículo muestra los resultados de un estudio que busca clasificar de forma adecuada a las personas ocupadas que cotizan a pensión en Colombia. Para ello, se utilizan diversos modelos de Machine Learning de clasificación, y los datos utilizados provienen de la Gran Encuesta Integrada de Hogares (GEIH) a corte de septiembre de 2022. El objetivo es entrenar un modelo que, basándose en diversas variables que muestran las condiciones de los individuos, pueda servir para identificar aquellos que cotizan a pensión y los que no lo hacen. La clasificación a través de información de variables puede servir para diseñar políticas públicas y planes de acción focalizados, que ayuden a que más personas logren hacer aportes a pensión en el país. En total se eligieron 7 variables, y entre los 6 modelos implementados, aquel que mejor Accuracy tuvo fue el Random Forest (0,92 dentro de muestra y 0,89 fuera de muestra).**

***Index Terms*—Big Data, clasificacion, prediccion, estimacion, Pobreza**

I. Introduccion

La pregunta que se busca resolver con este trabajo es ¿cuáles son los factores que determinan que una persona cotice al sistema de pensiones en Colombia? Esta investigación es relevante pues da luces acerca de cuáles son las variables más influyentes para lograr cotizar a pensión, pues en el país una gran porción de la población no logra pensionarse. Se busca que los resultados sirvan de base para implementar políticas públicas que focalicen los esfuerzos con la finalidad de que más colombianos ocupados aporten al sistema pensional, y con ello puedan tener una vejez en la cual no queden desprotegidos por falta de ingresos.

En Colombia el Sistema General de Pensiones tiene como objetivo garantizar a la población, el amparo contra las contingencias derivadas de la vejez, invalidez o muerte, mediante el reconocimiento de una pensión y prestaciones determinadas en la Ley. Este se encuentra compuesto por el Régimen Solidario de Prima Media (RPM) y el Régimen de Ahorro Individual (Rais) (Ministerio de Salud y Protección Social, 2012).

A pesar de reconocerse la importancia de las pensiones para una vejez digna, actualmente en el país el 75% de los colombianos no accede a la pensión de vejez (Fasecolda, 2019), a pesar de que cerca de 25 millones de personas se encuentran afiliadas al sistema pensional. Unos 18,2 millones de personas están vinculados en alguna de las cuatro Administradoras de Fondos de Pensiones (AFP) del Rais, mientras que 6,78 millones están en Colpensiones en el RPM (La República, 2022).

Para determinar adecuadamente cuáles son los factores más relevantes para que las personas coticen a algún fondo y más adelante logren cumplir con la meta de pensionarse, se trata de encontrar las variables asociadas a la condición de la persona y de su entorno, que influyen en la probabilidad de cotizar a pensiones. Nuestra serie de análisis (los cuales incluyen 6 modelos de Machine Learning de clasificación y 7 variables explicativas) nos indican que las variables elegidas son significativas a la hora de clasificar adecuadamente a los individuos que cotizan a pensiones, y se puede resaltar que los modelos que mejor clasifican son los el Random Forest y el XGBoots.

Entre los referentes teóricos consultados cabe mencionar que, algunas aproximaciones a la solución de nuestra pregunta pueden encontrarse en el trabajo de Franco, quien entrenó un modelo Probit para establecer el efecto de las variables determinantes al Sistema Pensional Colombiano. El autor encontró que la probabilidad de afiliarse está determinada principalmente por el nivel educativo, los ingresos y por el pertenecer al sector laboral formal (Franco, 2012).

Otro de los hallazgos relevantes en dicho trabajo es que la variable con mayor incidencia en la probabilidad de estar o no afiliado al Sistema Pensional es la Salud. Este resultado evidencia que si se hace parte de la población ocupada del sector formal se espera la cotización en los dos sistemas.

En la revisión bibliográfica también se encontró que Cardona y Toro elaboraron un modelo Logit el cual concluyó que, a mayor edad la probabilidad máxima de cotizar en el régimen de prima media es mayor. Además, se observa que para con educación entre los 0 y 5 años (nivel educativo equivalente a básica primaria) la mediana de la probabilidad de que no cotice a pensión se encuentra por encima del 90 %. Mientras que para todos los individuos que tienen acceso a educación superior (17 años o más) la mediana de la probabilidad de no cotizar cae a niveles del 5

Identify applicable funding agency here. If none, delete this.

Consideramos relevante en nuestro estudio tener en cuenta los análisis ya realizados en este campo, pero, además decidimos incluir nuevas variables que podrían llegar a favorecer la precisión, y también optamos por entrenar una mayor cantidad de modelos de distintos grados de complejidad, con el fin de que el estudio tenga resultados más confiables.

Así las cosas, nuestros datos indican que la probabilidad de estar afiliado a un fondo de pensiones es menor para las personas que: tienen mayores ingresos, reciben subsidio de transporte y tienen mayor nivel de educación.

Este artículo está compuesto por la presente introducción, una sección donde se describen los datos utilizados, una donde se explican los diversos modelos elaborados, otra donde se muestran los resultados de los modelos, cerrando con la sección de conclusiones. Al final del documento se detalla el sitio web donde están disponibles los datos y códigos utilizados.

II. Datos

Como insumos de los modelos de clasificación se utilizaron datos de la GEIH de septiembre de 2022. Esta encuesta es elaborada por la Dirección de Metodología y Producción Estadística del DANE. Esta puede ser consultada en el sitio web: https://microdatos.dane.gov.co//catalog/728/get_microdata
Concretamente se usaron las bases de datos "Características generales seguridad social en salud y educación" la cual contiene 75.383 observaciones de 74 variables, y la base de "Ocupados" que contiene 31.867 observaciones de 200 variables. Procedimos entonces a realizar un merge de ambas bases por el ID único para cada persona, dado que necesitábamos variables de ambas para los modelos.

También se depuró eliminando los missing values, así como los valores que reportaban 'no saben, no informan', y se dejaron solamente los datos de os mayores de 18 años. La base con la que trabajamos se nombró 'df' y quedó conformada por 15.905 observaciones y 8 variables. Con esta se sembró una semilla (12345) y se hizo una partición en base de entrenamiento (0,8) y testeo (0,2). Así las cosas, train quedó con 12.724 observaciones y test con 3181 observaciones.

III. Seleccion de Variables

* P6920 ¿Está... cotizando actualmente a un fondo de pensiones?: esta es nuestra variable objetivo que queremos clasificar adecuadamente en los modelos. Es una variable discreta numérica que toma los siguientes valores: 1. Sí, 2. No, 3. Ya es pensionado. Se eliminaron los pensionados y se procedió a dar valores de 0 si no cotiza a pensión y de 1 sí cotiza a pensión.

Para el caso de las variables explicativas "X" del modelo a continuación se explica cuáles se utilizaron y las razones por las cuales se decidió incluir cada una en el modelo:

* P6585s3 ¿Cuál o cuáles de los siguientes subsidios recibió ... el mes pasado: c. Subsidio Familiar? Esta variable discreta numérica toma los siguientes valores: 1. Sí, 2. No, 9. No sabe, no informa. Se incluyó pues consideramos relevante que las personas que reciben subsidios de este estilo pueden pertenecer a la población vulnerable, y ello puede estar asociado con el no tener una estabilidad laboral y de ingresos que le permita cotizar a pensión.

* P6585S2 ¿Cuál o cuáles de los siguientes subsidios recibió ... el mes pasado: b. Auxilio o subsidio de transporte? Esta variable discreta numérica toma los siguientes valores: 1. Sí, 2. No, 9. No sabe, no informa. Esto puede estar asociado con el no tener una estabilidad laboral y de ingresos que le permita cotizar a pensión.

* INGLABO Ingresos Laborales. Esta variable es continua numérica. Consideramos que los ingresos bajos no tendrían posibilidad de cubrir sus gastos en las necesidades básicas y a la vez dejar un porcentaje de su ingreso para destinarlo a ahorro pensional.

* P3042 ¿Cuál es el mayor nivel educativo alcanzado y el último grado o semestre aprobado por . . . ..? Esta variable continua numérica toma los siguientes valores: 1. Ninguno, 2. Preescolar, 3. Básica primaria (1o - 5o), 4. Básica secundaria (6o - 9o), 5. Media académica (Bachillerato clásico), 6. Media técnica (Bachillerato técnico), 7. Normalista, 8. Técnica profesional, 9. Tecnológica, 10. Universitaria, 11. Especialización, 12. Maestría, 13. Doctorado 99. No sabe, no informa. La razón de la inclusión en el modelo es que, la exploración inicial de los datos nos mostró que los mayores niveles de desocupación se encontraban en la población con menor nivel educativo, y si no están ocupados se puede asociar, aunque no en todos los casos, con que la persona no tenga aportes a pensión.

* P2057 ¿Usted se considera campesino(a)? Esta variable discreta numérica toma los siguientes valores: 1. Sí, 2. No. La exploración inicial de los datos nos permitió verificar que los mayores niveles de contratos de los ocupados que se consideraron campesinos eran verbales, consideramos que esto puede ser un factor que favorezca el no cotizar. Además, dado que muchos trabajos en la ruralidad son estacionales, ello no favorece para que sean formales.

* P6040 ¿cuántos años cumplidos tiene...? (si es menor de 1 año, escriba 00). Esta variable es continua numérica. La razón para incluirla fue que Bustamante encontró que la edad está directamente relacionada con la probabilidad de estar afiliado al sistema pensional, además de que a partir de los 35 años la probabilidad de afiliación se incrementa el doble (Bustamante, 2006).

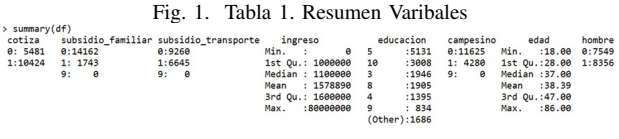
* P3271 ¿Cuál fue su sexo al nacer? Esta variable discreta numérica toma los siguientes valores: 1. Hombre, 2. No. Mujer. Se incluyó pues estudios como el de López muestran que solo el 12% de mujeres en Colombia se pensiona, y además la brecha entre hombres y mujeres es de un 10% (López, 2019). Por ello consideramos que puede ser relevante a la hora de clasificar si la persona cotiza a pensión.

A continuación, se muestra un resumen de las variables elegidas en la base antes de dividirla:

IV. Modelo

Predecir una respuesta cualitativa (de 0 o 1) para una observación puede denominarse clasificar esa observación, ya

Fig. 1. Tabla 1. Resumen Varibales



```
> summary(df)
 cotiza    subsidio_familiar subsidio_transporte    ingreso          educacion    campesino      edad      hombre
 0: 5481   0:14162           0:9260               Min.   :      0   5      :5131   0:11625   Min.   :18.00  0:7549
 1:10424   1: 1743           1:6645               1st Qu.: 1000000   10     :3008   1: 4280   1st Qu.:28.00  1:8356
           9:    0           9:   0               Median : 1100000   3      :1946   9:    0   Median :37.00
                                                  Mean   : 1578890   8      :1905             Mean   :38.39
                                                  3rd Qu.: 1600000   4      :1395             3rd Qu.:47.00
                                                  Max.   :80000000   9      : 834             Max.   :86.00
                                                                     (Other):1686
```

Fig. 2. Example of a figure caption.

que implica asignar la observación a una categoría o clase. Se estableció que el modelo base para el análisis tiene la siguiente estructura:

$$cotiza = \beta_0 + \beta_1 subsidio_familiar + \beta_2 subsidio_transporte + +\beta_3 ingreso + \beta_4 educacion + \beta_5 campesino + \beta_6 edad + \beta_7 hombre + u$$

Con este modelo esperamos que las variables explicativas seleccionadas nos den una adecuada clasificación de las personas que cotizan en el sistema de pensión. Para el ejercicio, la base de datos se dividió en train y test.

Cabe recordar que, en este caso la regresión lineal no es la más apropiada para ser utilizada, pues buscamos una respuesta cualitativa. Inicialmente se corrió un modelo de regresión logística (Logit), el cual usando una función nos da salidas entre 0 y 1 (en nuestro caso, estos representan que no cotiza, o que sí hace). La función logística producirá una curva en forma de S, por lo que independientemente del valor, se podrá obtener una predicción más sensata. También este modelo es más capaz de capturar el rango de probabilidades que la regresión lineal.

Posteriormente se corrió un modelo Probit. En la práctica, los modelos Probit y Logit generalmente arrojan probabilidades predichas muy similares. La distribución logística tiende a dar mayores probabilidades a Y = 1 cuando Xb es extremadamente pequeño (y menores probabilidades a Y = 1 cuando Xb es muy grande). Dado que es difícil justificar la elección de una u otra distribución sobre bases teóricas, se corrieron ambos modelos para ver su nivel de ajuste.

Se pretendía hacer un Análisis Discriminante Lineal (LDA), el cual es básicamente un enfoque empírico de Bayes, que asume las densidades como normales, y que las varianzas son iguales, y hacer un Análisis Cuadrático Discriminante (QDA), el cual proporciona una alternativa para cuando se asumen no linealidades. Sin embargo, al tener la variable 'Educación' que es categórica esto no permitió implementar estos modelos. Entonces se llevó a cabo un modelo de K vecinos más cercanos (KNN), el cual se fija en la cercanía de los observables para predecir la probabilidad condicional.

Buscando mejorar el rendimiento mediante la agregación de árboles se corrió un modelo de Random Forest. Este permite que se reduzca la correlación entre los árboles, y tiene la particularidad de que, si hay p predictores, en cada partición usa solo m ¡ p predictores, elegidos al azar. Esto evita el problema de que, si hay un predictor fuerte, los diferentes árboles sean muy similares entre sí.

Con el fin de aprender de los propios errores, se entrenó el modelo de Adaptative Boosting AdaBoost. Esta es la implementación para clasificación del Boosting, el cual con árboles cortos (poco profundos) puede hacer buenas predicciones, y les da más peso a los errores.

Por último, se entrenó un modelo Extreme Gradient Boost XGBoots, el cual es una implementación rápida de los Boosting trees, que tiene un rendimiento y velocidad óptimos. Este suele funcionar bien en datos tabulares, como es el caso de la GEIH que se utilizaron en este trabajo. El modelo corta los árboles de forma óptima, agregando un término de penalidad por número de hojas y por el score de la hoja.

Tras correr todos los modelos y realizar las curvas ROC, nos basamos en la métrica de Accuracy para elegir el mejor, y en este caso el ganador fue el Random Forest.

V. Rsultados

A continuación, se muestra la tabla resumen con las métricas de los principales modelos:

Modelo	Evaluación	Accuracy	Precision	Recall	F1
XGboost	Dentro de Muetsra	0.91	0.94	0.93	0.93
XGboost	Fuera de Muestra	0.89	0.92	0.92	0.92
Random Forest	Dentro de Muetsra	0.92	0.95	0.92	0.94
Random Forest	Fuera de Muetsra	0.89	0.92	0.91	0.92
KNN	NA	0.92	0.93	0.94	0.94
Adaboost	Dentro de Muestra	0.90	0.94	0.91	0.92
Adaboost	Fuera de Muestra	0.89	0.93	0.90	0.92
LOGIT	Dentro de Muestra	0.90	0.94	0.91	0.92
LOGIT	Fuera de Muestra	0.89	0.93	0.91	0.92
PROBIT	Dentro de Muestra	0.90	0.95	0.90	0.92
PROBIT	FUERA de Muestra	0.89	0.94	0.90	0.92

TABLE I
FUENTE: GEIH COLOMBIA
HTTPS:MICRODATOS.DANE.GOV.COCATALOG728GET_MICRODATA

Los coeficientes del modelo granador (el Random Forest) nos indican que e Accuracy fue del 0,92 dentro de muestra y del 0,89 fuera de muestra, logrando clasificar correctamente a gran mayoría de personas cotizantes al sistema de pensiones.

Este modelo tiene la ventaja de que permite que no se usen todas las variables, como hace Bagging, sino que elige un subconjunto de esas variables. A diferencia de los árboles, el Random Forest busca que no sea tan profundo.

Hay que resaltar que, todos los modelos tienen desempeños parecidos y bastante buenos. El que mejor predice es el modelo KNN, sin embargo, como nuestro objetivo es ver a través de qué variables acertar a que coticen a pensión, este no nos sirve de mucho, porque nos predice muy bien pero no nos dice cuáles variables nos ayudan a cambiar esa situación, pues no estima el efecto marginal de las diferentes variables.

Al realizar las evaluaciones fuera de muestra todos los modelos tienen resultados dan muy parecidos en Accuracy y F1. En el caso de las evaluaciones dentro de muestra los dos mejores son Random Forest y XGBoots. Aunque todos

predicen bien, decidimos elegir el Random Forest, porque si bien dentro de muestra el XGBoots tiene un poco más de Recall, el Random Forest tiene gana en Accuracy, Precision y F1. En fuera de muestra están iguales en todo, excepto en Recall donde gana XGBoots.

VI. Conclusiones

La realización de este trabajo permite concluir que las variables más importantes para clasificar a las personas que cotizan a pensión en Colombia son, en su orden: 1. El ingreso de la persona, 2. Si recibe subsidio de transporte, y 3. El nivel de educación máximo alcanzado.

Que el hecho de tener ingresos altos fuera lo más relevante en estos modelos era de esperarse, dado que es una condición sin la cual es muy difícil destinar una parte de dichos ingresos a gastos que no son de primera necesidad como es el caso de la cotización en el sistema pensional. Además, este hallazgo va de la mano con los resultados de otros trabajos consultados en la revisión bibliográfica previa a la elaboración del modelo.

Sin embargo, sorprendió que la segunda variable más importante para que las personas estén en el grupo de los cotizantes a pensión fuera el recibir subsidio de transporte, pues esto es algo en lo que las autoridades pueden incidir a través de políticas públicas, y por ello deberían tenerse en cuenta al momento de diseñarlas.

References

Bustamante, J. (2006). Revista Planeación y Desarrollo. Obtenido de DNP: https://colaboracion.dnp.gov.co/CDT/RevistaPD/2006/pd_vXXXVII

Cardona, J., Toro, G. (2022). Repositorio. Obtenido de EAFIT: https://repository.eafit.edu.co/bitstream/handle/10784/31480/JuanJos ence=2isAllow

Fasecolda. (Julio de 2019). Obtenido de https://fasecolda.com/cms/wp-content/uploads/2019/09/seminario-sistema-pensional.pdf

Franco, L. (Diciembre de 2012). Biblioteca digital. Obtenido de Universidad del Valle: https://bibliotecadigital.univalle.edu.co/bitstream/handle/10893/5957 p.pdf?sequence=1

La República. (Julio de 2022). Obtenido de https://www.larepublica.co/finanzas/afiliados-al-sistema-pensional-llegan-a-25-millones-y-73-estan-en-los-fondos-privados-3400845

López, A. (Agosto de 2019). Escuela de Gobierno. Obtenido de Universidad de Los Andes: https://gobierno.uniandes.edu.co/sites/default/files/books/DT/DT-67.pdf

Ministerio de Salud y Protección Social. (Octubre de 2012). Sistema General de Pensiones. Obtenido de https://www.minsalud.gov.co/proteccionsocial/Paginas/Sistema-general-de-Pensiones.aspx: :text=El%20Sistema%20General%20de%20Pe

Anexos

Fig. 3. Gráfico 1: Curvas ROC separadas

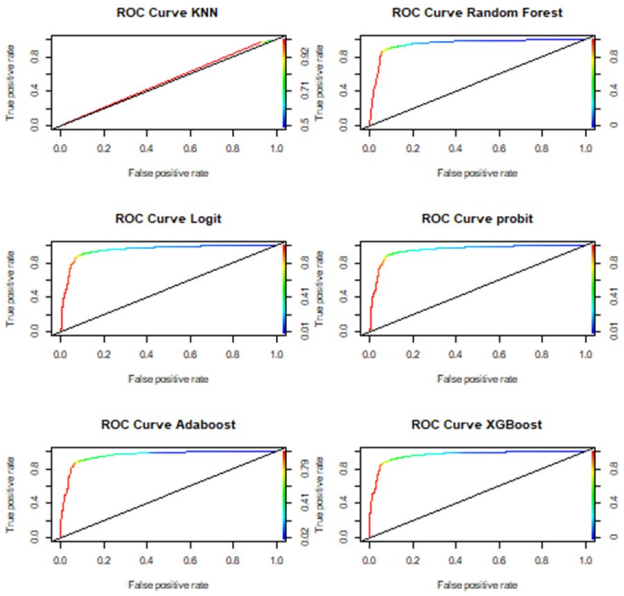


Fig. 4. o

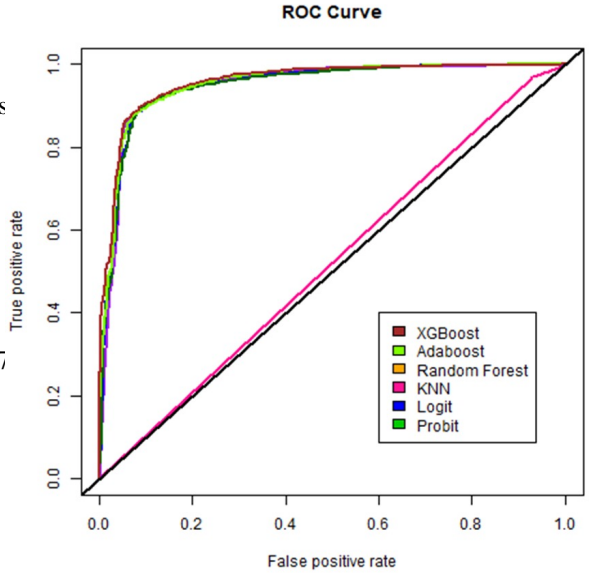


Fig. 5. Gráfico 3: Summary de bases train y test

```
> summary(train)
 cotiza   subsidio_familiar subsidio_transporte    ingreso          educacion    campesino      edad      hombre
 0:4385   0:11319           0:7389               Min.   :      0   5      :4057   0:9273    Min.   :18.00  0:6021
 1:8339   1: 1405           1:5335               1st Qu.: 1000000   10     :2362   1:3451    1st Qu.:28.00  1:6703
          9:    0           9:   0               Median : 1100000   8      :1574   9:   0    Median :37.00
                                                  Mean   : 1578490   3      :1552             Mean   :38.39
                                                  3rd Qu.: 1600000   4      :1129             3rd Qu.:47.00
                                                  Max.   :80000000   9      : 696             Max.   :86.00
                                                                     (Other):1354

> summary(test)
 cotiza   subsidio_familiar subsidio_transporte    ingreso          educacion    campesino      edad      hombre
 0:1096   0:2843           0:1871               Min.   :      0   5      :1074   0:2352    Min.   :18.00  0:1528
 1:2085   1: 338           1:1310               1st Qu.: 1000000   10     : 646   1: 829    1st Qu.:28.00  1:1653
          9:   0           9:   0               Median : 1055000   3      : 394   9:   0    Median :37.00
                                                  Mean   : 1580489   8      : 331             Mean   :38.39
                                                  3rd Qu.: 1600000   4      : 266             3rd Qu.:48.00
                                                  Max.   :40000000   11     : 143             Max.   :81.00
                                                                     (Other): 327
```

Fig. 6. Grafico: 4Matrix de Confusion

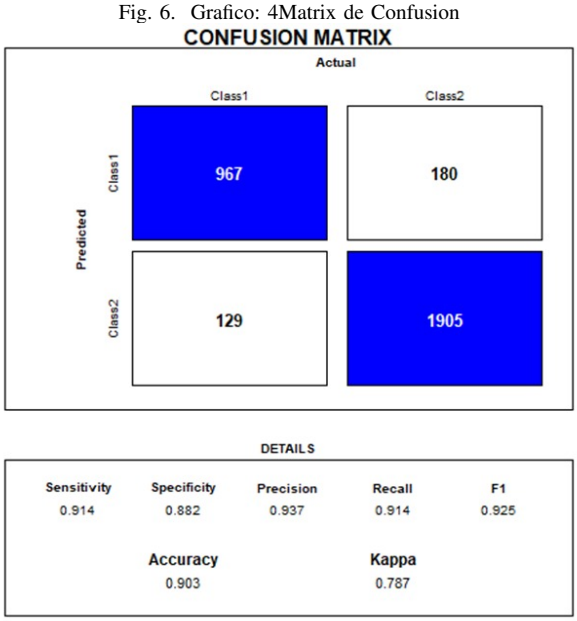


Fig. 7. Gráfico 5: Importancia de las variables