Modelos de machine learning para la estimación de probabilidad de default

Leonela Tacure P., Victor Villacorta P., Osman Rafael T.

Resumen – El sector bancario peruano se ha vuelto muy competitivo, y cada vez más expuesto a múltiples fuentes de información, muchos bancos están buscando nuevos métodos para reducir el riesgo de crédito, con el fin de mejorar su desempeño. El objetivo de la investigación consiste en estimar la probabilidad de default de los clientes del banco peruano Interbank. Y como objetivo específico encontrar perfiles de riesgo luego del ordenamiento de los clientes en base al score estimado. Para realizar la estimación contamos con múltiples fuentes de información con un periodo de 12 meses (marzo 2017 a febrero 2018) para entrenar el modelo y 3 meses del 2019 para la validación. La estructura de la investigación inicia con un análisis descriptivo de las variables, luego mostramos los resultados de los 3 modelos desarrollados para la estimación (regresión logística, regresión de enlaces asimétricos, ensamblados). Finalmente, se muestra un comparativo del indicador Gini y curva ROC para elegir al mejor modelo.

**Índice de Términos – riesgos, deault, regresión logística, enlaces asimétricos, Gini, curva ROC**

# introduccion

Fue en el sector bancario donde la evaluación del riesgo crediticio se estableció por primera vez a mediados del siglo XX. El Riesgo de Crédito, según la SBS (Superintendencia de Banca, Seguros y AFP’s), se define como “la posibilidad de pérdidas por la incapacidad o falta de voluntad de los deudores, contrapartes, o terceros obligados, para cumplir sus obligaciones contractuales registradas dentro o fuera del balance”.

La Probabilidad de Incumplimiento (Probability of default, PD) se define como la probabilidad de que la contraparte no pueda hacer frente a sus obligaciones contractuales en un determinado horizonte de tiempo -generalmente un año- la cual podría ser expresado a través de esta ecuación en uno de los modelos propuestos para su cálculo:

𝑃𝐷𝑡 = 𝑂𝑝𝑒𝑟𝑎𝑐𝑖𝑜𝑛𝑒𝑠 𝑚𝑜𝑟𝑜𝑠𝑎𝑠 𝑒𝑛 𝑒𝑙 𝑎ñ𝑜 𝑡 𝑑𝑒 𝑠𝑢 𝑣𝑖𝑑𝑎

𝑇𝑜𝑡𝑎𝑙 𝑑𝑒 𝑜𝑝𝑒𝑟𝑎𝑐𝑖𝑜𝑛𝑒𝑠 𝑞𝑢𝑒 ℎ𝑎𝑛 𝑣𝑖𝑣𝑖𝑑𝑜 ℎ𝑎𝑠𝑡𝑎 𝑒𝑙 𝑎ñ𝑜 𝑡

En el Perú la entidad que regula a las metodologías del cálculo del scoring de aplicación, scoring de comportamiento y scoring de severidad es la SBS.

Por lo cual la investigación propone desarrollar un modelo que estime la probabilidad de default con un sustento estadístico y también con un sustento normativo en cuanto a algunas definiciones, tales como la definición de incumplimiento dadas por la Entidad Reguladora.

# Procedimiento desarrollo del trabajo

## Análisis exploratorio

**1.Fuentes de información:** A continuación, se muestran las 4 fuentes de información usadas para estimar la probabilidad de default de los clientes del banco Interbank.

TABLA I

Fuentes de información

|  |  |
| --- | --- |
| Nombre | Descripción |
| rcc\_train | Información del historial creditico de los clientes |
| censo\_train  sunat\_train  censo\_train | Información censal del Perú año 2017  Información fiscal de los clientes  Información censal del Perú año 2017 |
|  |  |

**2. Operacionalización de variables: describimos a las variables por cada fuente de información**

TABLA I

DESCRIPCIÓN BASE SE\_TRAIN

|  |  |
| --- | --- |
| Nombre | Descripción |
| edad | año y mes de la observación |
| sexo  est\_civil  sit\_lab | días de atraso  tipo de crédito  código de la institución financiera |
| cod\_ocu  ctd\_hijos  flag\_sin\_email  ctd\_veh  cod\_ubi  lgr\_vot  prv  dto  rgn  tip\_lvledu | monto de línea  tipo de producto  indicador ordinal del riesgo  indicador categórico del riesgo  código ubigeo  lugar de votación  provincia  departamento  región  tipo de nivel educativo |
|  |  |

TABLA II

DESCRIPCIÓN BASE RCC\_TRAIN

|  |  |
| --- | --- |
| Nombre | Descripción |
| Codmes | año y mes de la observación |
| condicion  tipo\_credito  cod\_inst\_financiera | días de atraso  tipo de crédito  código de la institución financiera |
| saldo  producto  riesgo\_directo  cod\_clasifi\_deudor | monto de línea  tipo de producto  indicador ordinal del riesgo  indicador categórico del riesgo |
|  |  |

TABLA III

DESCRIPCIÓN BASE SUNAT\_TRAIN

|  |  |
| --- | --- |
| Nombre | Descripción |
| tipcontribuyente | tipo de contribuyente sunat |
| tippersona  ciiu  ubigeo | tipo de persona sunat  codigo de actividad económica  ubigeo |
| condiciondomicilio  estadocontribuyente  fecalta  fecbaja  codvia  codzona  contabilidad  facturación  domiciliado  comercioexterior  cargorele  codentidadtributo  estadotributo | condicion del domicilio fiscal  estado del contribuyente  fecha alta  fecha baja  codigo de vía  codigo de zona  tipo contabilidad  tipo de facturación  condición de domicilio  marca actividad comercio exterior  cargo representante legal  codigo de entidad deuda  estado deuda tributaria |
|  |  |

**3. Estadísticas descriptivas de variables**

**Target:** clasificación del riesgo de default, está etiquetada como 0 (No tienen riesgo) y 1 (Sí tienen riesgo), con un 84.1% y 15.8% respectivamente.

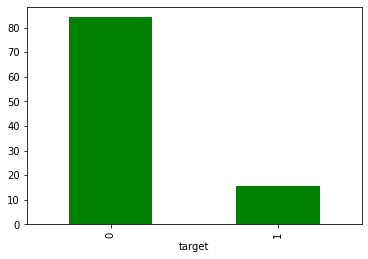


Fig. 1. Distribución de la variable target

**Variables de rcc\_train:** Presentamos mediante gráficos como se distribuyen las variables categóricas de la fuente rcc\_train. Para el caso de código de meses, mostramos los 12 meses contando clientes únicos. Para el caso de la variable tipo de crédito el 85% se concentra en los tipos de crédito 12,11,10. Para el caso de la clasificación de deudor el 80% se concentra en el código 0. Para la variable condición el 70% se concentra en el código 0. Para riesgo directo más del 90% se concentra en los códigos 1 y -1. Finalmente, para el caso de la variable producto la cardinalidad es alta, sobresaliendo el producto código 1 y 0.

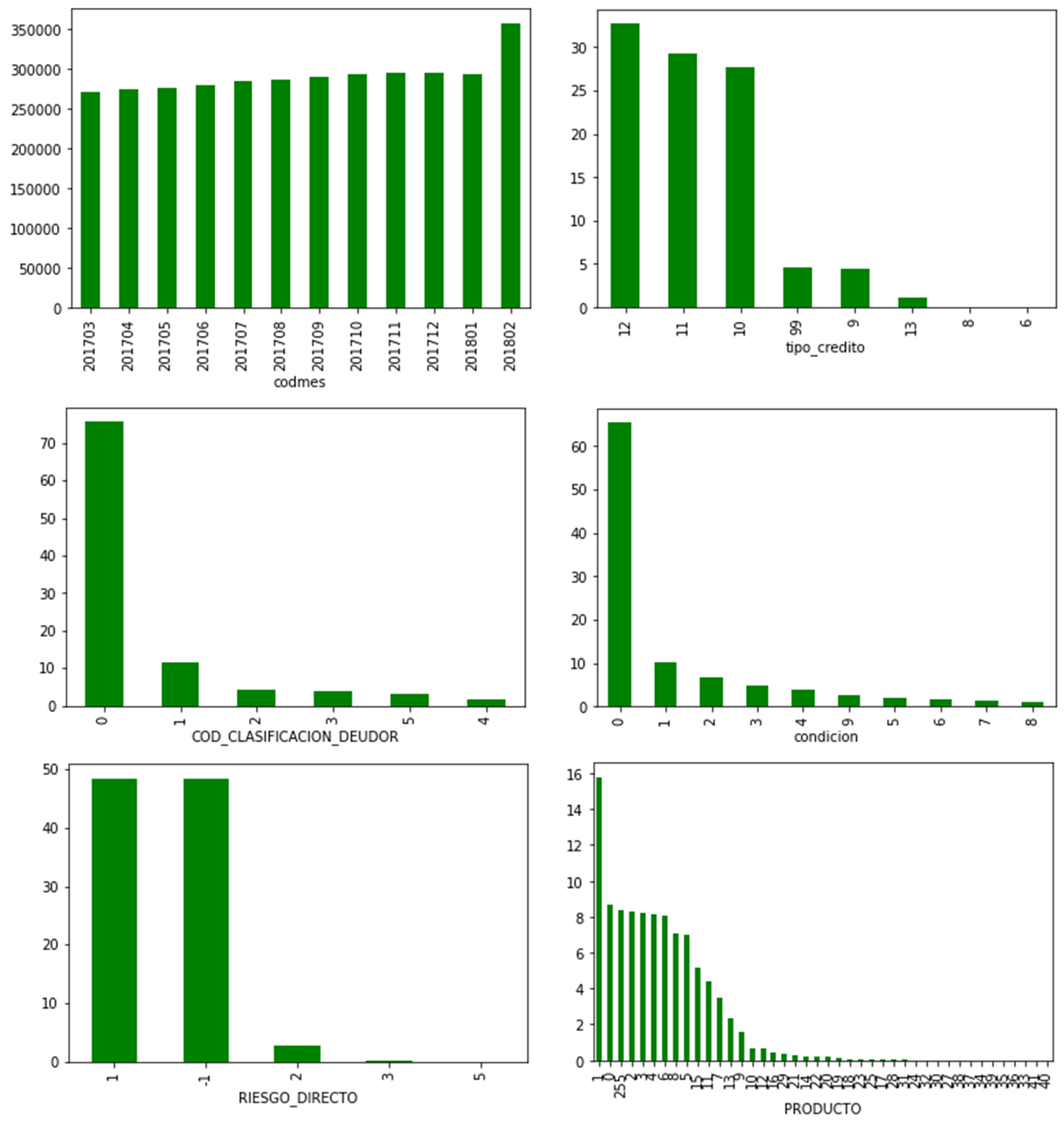


Fig. 2. Distribución de las variables categóricas fuente rcc\_train

Mostramos la variable saldo cruzada con la variable target en un BoxPlot, se ha calculado la suma del saldo y me media del saldo para cada grupo del target. El gráfico nos muestra la presencia de outliers para la categoría 0, pero aún más marcada en la categoría 1.

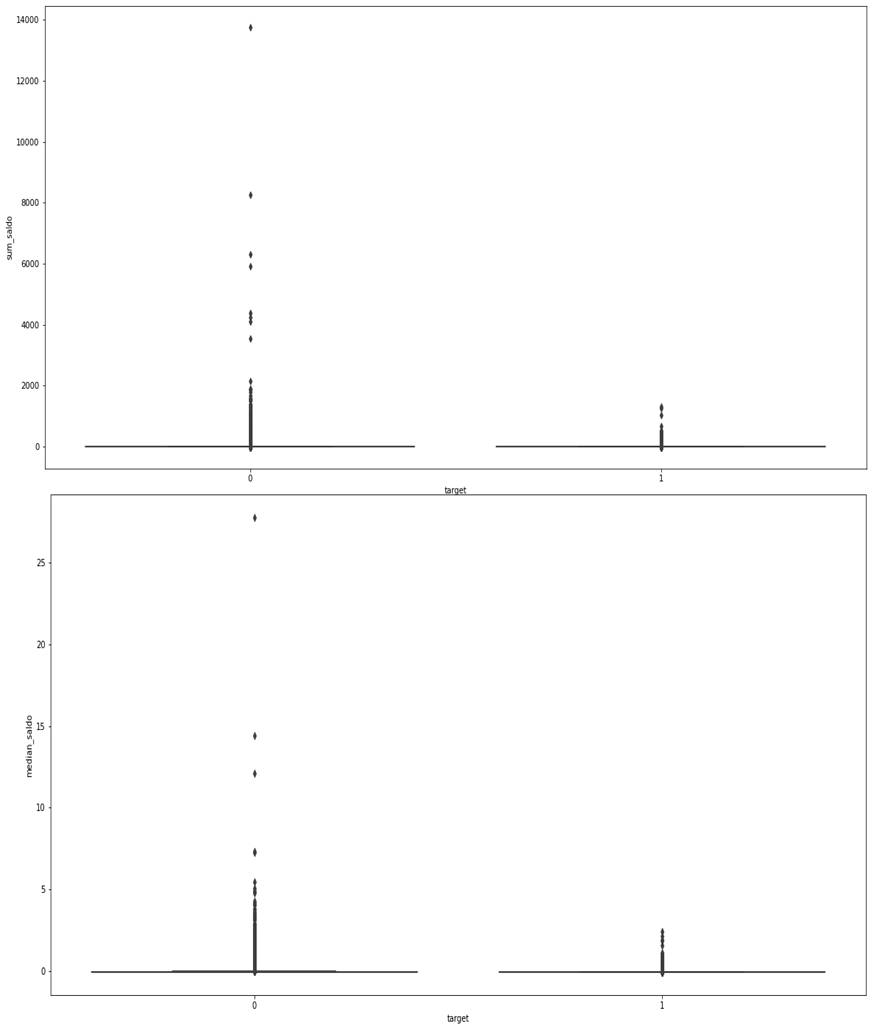


Fig. 3. Distribución de las variables continuas fuente rcc\_train

**Variables de se\_train:** Presentamos mediante gráficos como se distribuyen las variables categóricas de la fuente rcc\_train cruzada con la variable target. Para el caso de situación laboral hay una mayor proporción de casos de éxito el código 3. Para el caso de sexo la proporción de casos de éxito es homogénea. Para la variable estado civil los códigos que tienen mayor proporción de casos de éxito es 2 y 4. Para el caso de cantidad de hijos hay un ligero incremento de casos positivos a partir de 5 hijos a más. Respecto a si tiene mail o no, tenemos menos casos de éxito en la target cuando el cliente sí tiene mail. Por ultimo respecto a la variable nivel educativo, no tenemos muchas casos de éxito en el nivel educativo 1 y a mayor cantidad de casos de éxito están en el nivel 7.

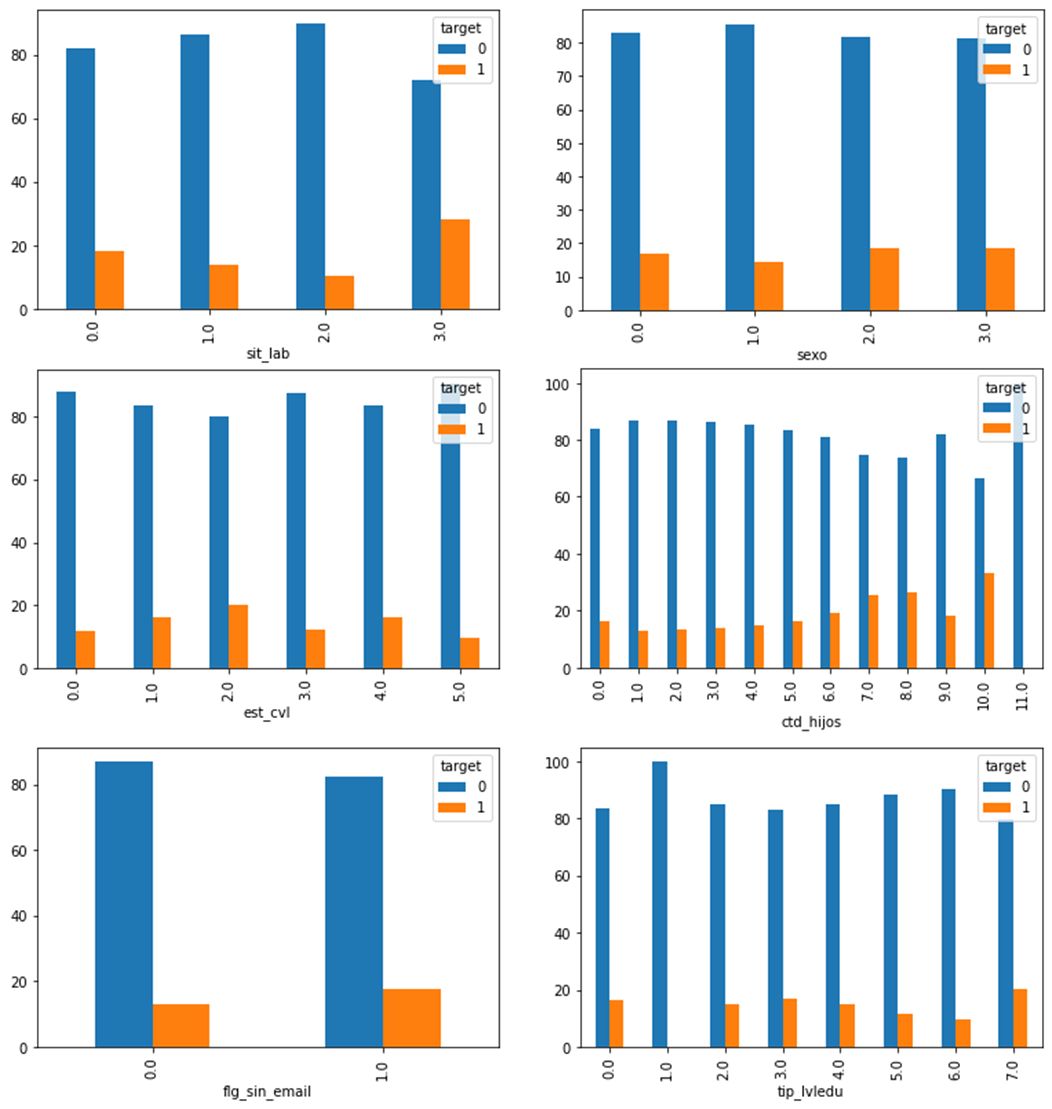


Fig. 4. Distribución de las variables categóricas fuente se\_train

**Variables de sunat\_train:** Mostramos mediante gráficos las variables categóricas de la fuente sunat\_train

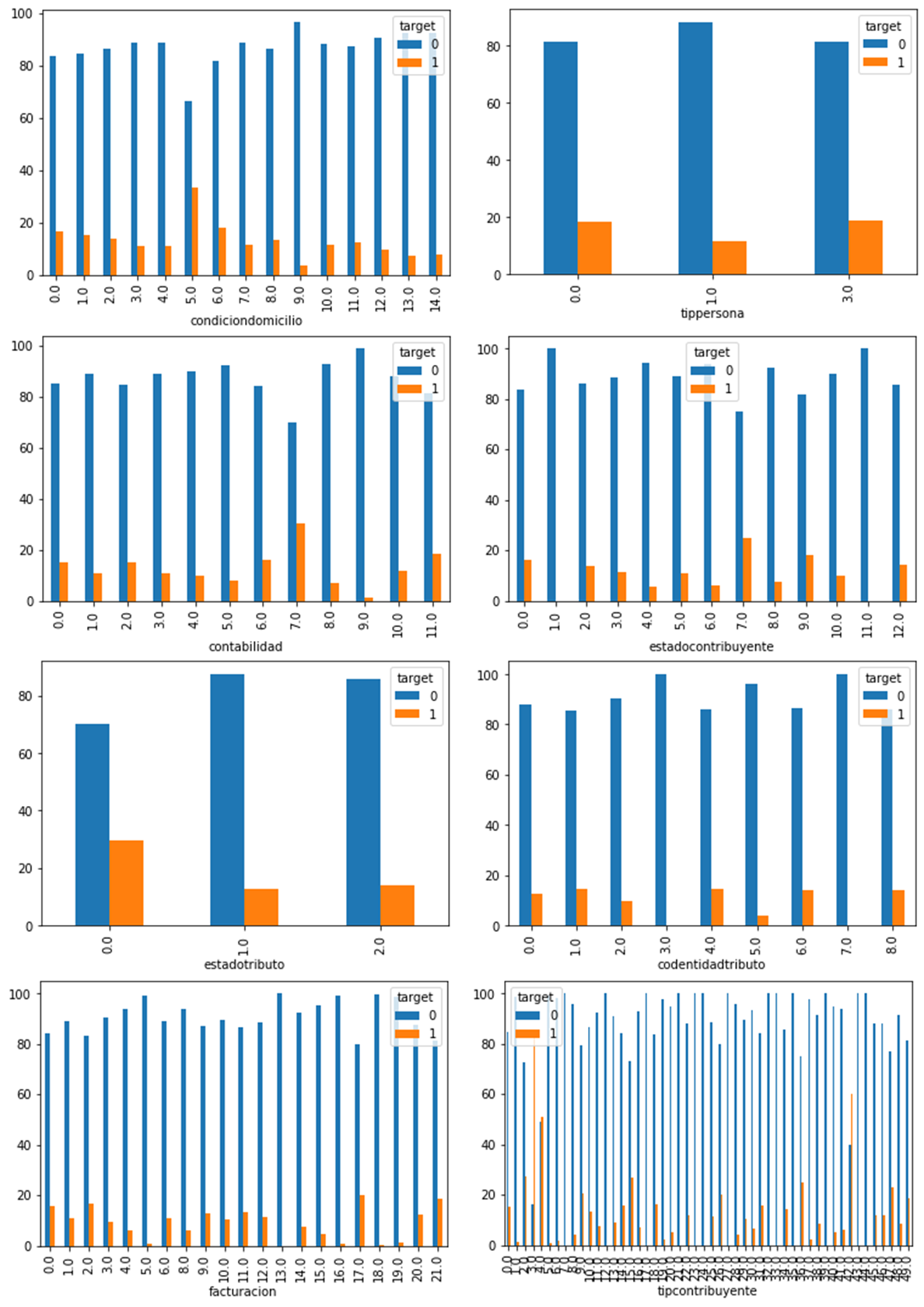


Fig. 4. Distribución de las variables categóricas fuente se\_train

## Tratamiento de las variables

# marco teórico

# resultados de los modelos de machine learning usados

# Discusión

# Conclusiones

Una sección de conclusión no es necesaria. Sin embargo esta puede repasar los puntos principales del artículo, no repita el resumen como conclusión. Una conclusión se elabora con base en la importancia del trabajo realizado o en las aplicaciones y extensiones sugeridas.

Apéndice

Los apéndices, si son necesarios, aparecen antes del reconocimiento.

Reconocimiento

Use el título singular aún cuando tenga que hacer muchos reconocimientos. Evite las expresiones como “Uno de nosotros (S.B.A.) gustaría agradecer....” En cambio, escriba “F. A. agradecimientos del autor....” los reconocimientos a un patrocinador y de apoyo financiero se ponen en la nota a pie de página de la primera página sin numerar.

References

1. G. O. Young, “Synthetic structure of industrial plastics (Book style with paper title and editor),” in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.
2. W.-K. Chen, *Linear Networks and Systems* (Book style)*.* Belmont, CA: Wadsworth, 1993, pp. 123–135.
3. H. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1985, ch. 4.
4. B. Smith, “An approach to graphs of linear forms (Unpublished work style),” unpublished.
5. E. H. Miller, “A note on reflector arrays (Periodical style—Accepted for publication),” *IEEE Trans. Antennas Propagat.*, to be published.
6. J. Wang, “Fundamentals of erbium-doped fiber amplifiers arrays (Periodical style—Submitted for publication),” *IEEE J. Quantum Electron.*, submitted for publication.
7. C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.
8. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interfaces(Translation Journals style),” *IEEE Transl. J. Magn.Jpn.*, vol. 2, Aug. 1987, pp. 740–741 [*Dig. 9th Annu. Conf. Magnetics* Japan, 1982, p. 301].
9. M. Young, *The Techincal Writers Handbook.* Mill Valley, CA: University Science, 1989.
10. J. U. Duncombe, “Infrared navigation—Part I: An assessment of feasibility (Periodical style),” *IEEE Trans. Electron Devices*, vol. ED-11, pp. 34–39, Jan. 1959.
11. S. Chen, B. Mulgrew, and P. M. Grant, “A clustering technique for digital communications channel equalization using radial basis function networks,” *IEEE Trans. Neural Networks*, vol. 4, pp. 570–578, July 1993.
12. R. W. Lucky, “Automatic equalization for digital communication,” *Bell Syst. Tech. J.*, vol. 44, no. 4, pp. 547–588, Apr. 1965.
13. S. P. Bingulac, “On the compatibility of adaptive controllers (Published Conference Proceedings style),” in *Proc. 4th Annu. Allerton Conf. Circuits and Systems Theory*, New York, 1994, pp. 8–16.
14. G. R. Faulhaber, “Design of service systems with priority reservation,” in *Conf. Rec. 1995 IEEE Int. Conf. Communications,* pp. 3–8.
15. W. D. Doyle, “Magnetization reversal in films with biaxial anisotropy,” in *1987 Proc. INTERMAG Conf.*, pp. 2.2-1–2.2-6.
16. G. W. Juette and L. E. Zeffanella, “Radio noise currents n short sections on bundle conductors (Presented Conference Paper style),” presented at the IEEE Summer power Meeting, Dallas, TX, June 22–27, 1990, Paper 90 SM 690-0 PWRS.
17. J. G. Kreifeldt, “An analysis of surface-detected EMG as an amplitude-modulated noise,” presented at the 1989 Int. Conf. Medicine and Biological Engineering, Chicago, IL.
18. J. Williams, “Narrow-band analyzer (Thesis or Dissertation style),” Ph.D. dissertation, Dept. Elect. Eng., Harvard Univ., Cambridge, MA, 1993.
19. N. Kawasaki, “Parametric study of thermal and chemical nonequilibrium nozzle flow,” M.S. thesis, Dept. Electron. Eng., Osaka Univ., Osaka, Japan, 1993.
20. J. P. Wilkinson, “Nonlinear resonant circuit devices (Patent style),” U.S. Patent 3 624 12, July 16, 1990.
21. *IEEE Criteria for Class IE Electric Systems* (Standards style)*,* IEEE Standard 308, 1969.
22. *Letter Symbols for Quantities*, ANSI Standard Y10.5-1968.
23. R. E. Haskell and C. T. Case, “Transient signal propagation in lossless isotropic plasmas (Report style),” USAF Cambridge Res. Lab., Cambridge, MA Rep. ARCRL-66-234 (II), 1994, vol. 2.
24. E. E. Reber, R. L. Michell, and C. J. Carter, “Oxygen absorption in the Earth’s atmosphere,” Aerospace Corp., Los Angeles, CA, Tech. Rep. TR-0200 (420-46)-3, Nov. 1988.
25. (Handbook style) *Transmission Systems for Communications,* 3rd ed., Western Electric Co., Winston-Salem, NC, 1985, pp. 44–60.
26. *Motorola Semiconductor Data Manual,* Motorola Semiconductor Products Inc., Phoenix, AZ, 1989.
27. (Basic Book/Monograph Online Sources) J. K. Author. (year, month, day). *Title* (edition) [Type of medium]. Volume(issue). Available: <http://www.(URL>)
28. J. Jones. (1991, May 10). Networks (2nd ed.) [Online]. Available: <http://www.atm.com>
29. (Journal Online Sources style) K. Author. (year, month). Title. *Journal* [Type of medium]. Volume(issue), paging if given. Available: <http://www.(URL>)
30. R. J. Vidmar. (1992, August). On the use of atmospheric plasmas as electromagnetic reflectors. *IEEE Trans. Plasma Sci.* [Online]. *21(3).* pp. 876—880. Available: http://www.halcyon.com/pub/journals/21ps03-vidmar

**Biografía Autor(es)** (M'76-SM'81-F'87) y los otros autores pueden incluir las biografías al final de los documentos(papers) regulares. Por favor incluyan nombres y apellidos con los cuales puedan ser identificados al registrar sus artículos (Son registrados en la base de Publindex en Colciencias, entre otras). Si ha enviado documentos antes, no debe suponer que el Comité Editorial conoce o puede decidir cuál es o cuáles son los autores del artículo. *Cada envío debe ser completo en todos sus datos*. El primer párrafo debe contener la filiación institucional (por ejemplo, profesor asociado, Facultad de Ingeniería de Sistemas, Universidad El Bosque). Los grados deben listarse con el tipo de grado, en qué campo, en que institución, ciudad, estado o país.

El segundo párrafo usa el pronombre de la persona (él o ella) y no el apellido o nombre del autor. Lista la experiencia académica y laboral. Se ponen en mayúscula los títulos del trabajo. Pueden listarse cargos anteriores. Información que involucra las publicaciones anteriores puede ser incluida. Intente no listar más de tres libros o artículos publicados. El formato para listar a publicadores de un libro dentro de la biografía es: el título de libro (la ciudad, estado: el nombre del publicador, año) similar a una referencia. Los intereses de investigaciones actuales y anteriores terminan el párrafo.

El tercer párrafo empieza con el título del autor y apellido (por ejemplo, Dr. Smith, Prof. Jones, Sr. Kajor, Ms. Hunter). Finalmente, liste cualquier premio por trabajos y publicaciones. Proporcionar una fotografía es requisito para publicar su artículo: la biografía se dentará alrededor de ella. La fotografía se pone en la esquina superior izquierda de la biografía. Se quitarán las aficiones personales de la biografía.