Primera entrega de proyecto

POR:

Alejandro Sarasti Sierra

MATERIA:

Modelos y simulación I

PROFESOR:

Raul Ramos Pollan

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

MEDELLÍN 2023-2

1. Planteamiento del problema

Los programas de ayuda humanitaria ya sean gubernamentales o de organismos internacionales, a menudo enfrentan limitaciones en la cantidad de recursos disponibles para su distribución. Por lo tanto, es de vital importancia asegurar que estos recursos lleguen a las familias, personas y comunidades que se encuentran en mayor situación de vulnerabilidad. Sin embargo, identificar quiénes son exactamente los más vulnerables puede resultar un desafío, especialmente en poblaciones de bajos ingresos que carecen de registros financieros sólidos. Para abordar este problema, se ha desarrollado un método conocido como 'prueba de medios indirectos' (Proxy Means Test), el cual se basa en las características de las viviendas y terrenos de los hogares para estimar su nivel de vulnerabilidad. En la actualidad, este método se apoya en enfoques econométricos simples, pero se están explorando técnicas de inteligencia artificial que podrían mejorar la precisión de esta evaluación.

El objetivo de este proyecto es predecir el nivel de vulnerabilidad en el que se encuentra una familia, que se divide en cuatro categorías: 1. Pobreza extrema, 2. Pobreza moderada, 3. Hogares vulnerables y 4. Hogares no vulnerables. Como se mencionó anteriormente, esta predicción se realizará en función de las características de la vivienda y la composición de la familia que reside en ella.

2. Dataset

El dataset a utilizar proviene de la competencia de kaggle “Costa Rican Household Poverty Level Prediction” que fue patrocinada por el IBD, misma organización de donde originalmente provienen los datos originales. La base de datos contiene 142 columnas, de las cuales aproximadamente el 71% son categóricas. Además, está dividida en dos archivos .csv *train* y *test*, los cuales tienen las siguientes columnas:

1. Id, identificador de la fila.
2. v2a1, Pago mensual de alquiler
3. hacdor, =1 Hacinamiento por habitaciones
4. rooms, número de todas las habitaciones en la casa
5. hacapo, =1 Hacinamiento por habitaciones
6. v14a, =1 Tiene baño en el hogar
7. refrig, =1 Si el hogar tiene refrigerador
8. v18q, Posee una tablet
9. v18q1, Número de tablets que posee el hogar
10. r4h1, Hombres menores de 12 años de edad
11. r4h2, Hombres de 12 años de edad o mayores
12. r4h3, Total de hombres en el hogar
13. r4m1, Mujeres menores de 12 años de edad
14. r4m2, Mujeres de 12 años de edad o mayores

15. r4m3, Total de mujeres en el hogar
16. r4t1, Personas menores de 12 años de edad
17. r4t2, Personas de 12 años de edad o mayores
18. r4t3, Total de personas en el hogar
19. tamhog, Tamaño del hogar
20. tamviv, Número de personas que viven en el hogar
21. escolari, Años de escolaridad
22. rez_esc, Años de retraso en la escuela
23. hhsize, Tamaño del hogar
24. paredblolad, =1 Si el material predominante en la pared exterior es bloque o ladrillo
25. paredzocalo, =1 Si el material predominante en la pared exterior es zócalo (madera, zinc o asbesto)
26. paredpreb, =1 Si el material predominante en la pared exterior es prefabricado o cemento
27. pareddes, =1 Si el material predominante en la pared exterior es material de desecho
28. paredmad, =1 Si el material predominante en la pared exterior es madera
29. paredzinc, =1 Si el material predominante en la pared exterior es zinc
30. paredfibras, =1 Si el material predominante en la pared exterior es fibras naturales
31. paredother, =1 Si el material predominante en la pared exterior es otro
32. pisomoscer, =1 Si el material predominante en el piso es mosaico, cerámica, terrazo
33. pisocemento, =1 Si el material predominante en el piso es cemento
34. pisoother, =1 Si el material predominante en el piso es otro
35. pisonatur, =1 Si el material predominante en el piso es material natural
36. pisonotiene, =1 Si no hay piso en el hogar
37. pisomadera, =1 Si el material predominante en el piso es madera
38. techozinc, =1 Si el material predominante en el techo es lámina de metal o zinc
39. techoentrepiso, =1 Si el material predominante en el techo es de fibrocemento o entrepiso
40. techocane, =1 Si el material predominante en el techo es fibras naturales
41. techootro, =1 Si el material predominante en el techo es otro
42. cielorazo, =1 Si la casa tiene cielo raso
43. abastaguadentro, =1 Si el suministro de agua está dentro de la vivienda
44. abastaguafuera, =1 Si el suministro de agua está fuera de la vivienda
45. abastaguano, =1 Si no hay suministro de agua
46. public, =1 Electricidad de CNFL, ICE, ESPH/JASEC
47. planpri, =1 Electricidad de planta privada
48. noelec, =1 Sin electricidad en la vivienda
49. coopele, =1 Electricidad de cooperativa
50. sanitario1, =1 Sin inodoro en la vivienda
51. sanitario2, =1 Inodoro conectado a alcantarillado o fosa séptica
52. sanitario3, =1 Inodoro conectado a tanque séptico
53. sanitario5, =1 Inodoro conectado a pozo negro o letrina

54. sanitario6, =1 Inodoro conectado a otro sistema
55. energcocinar1, =1 Sin fuente principal de energía para cocinar (sin cocina)
56. energcocinar2, =1 Fuente principal de energía para cocinar: electricidad
57. energcocinar3, =1 Fuente principal de energía para cocinar: gas
58. energcocinar4, =1 Fuente principal de energía para cocinar: leña o carbón
59. elimbasu1, =1 Si la disposición de basura es principalmente por camión cisterna
60. elimbasu2, =1 Si la disposición de basura es principalmente en agujero o enterrada
61. elimbasu3, =1 Si la disposición de basura es principalmente por quemado
62. elimbasu4, =1 Si la disposición de basura es principalmente arrojada en un espacio no ocupado
63. elimbasu5, =1 Si la disposición de basura es principalmente arrojada en río, arroyo o mar
64. elimbasu6, =1 Si la disposición de basura es principalmente otra
65. epared1, =1 Si las paredes son malas
66. epared2, =1 Si las paredes son regulares
67. epared3, =1 Si las paredes son buenas
68. etecho1, =1 Si el techo es malo
69. etecho2, =1 Si el techo es regular
70. etecho3, =1 Si el techo es bueno
71. eviv1, =1 Si el piso es malo
72. eviv2, =1 Si el piso es regular
73. eviv3, =1 Si el piso es bueno
74. dis, =1 Si hay una persona con discapacidad
75. male, =1 Si es hombre
76. female, =1 Si es mujer
77. estadocivil1, =1 Si tiene menos de 10 años de edad
78. estadocivil2, =1 Si está libre o en unión de hecho
79. estadocivil3, =1 Si está casado(a)
80. estadocivil4, =1 Si está divorciado(a)
81. estadocivil5, =1 Si está separado(a)
82. estadocivil6, =1 Si es viudo(a)
83. estadocivil7, =1 Si está soltero(a)
84. parentesco1, =1 Si es jefe(a) del hogar
85. parentesco2, =1 Si es cónyuge o pareja
86. parentesco3, =1 Si es hijo(a)
87. parentesco4, =1 Si es hijastro(a)
88. parentesco5, =1 Si es yerno(a)
89. parentesco6, =1 Si es nieto(a)
90. parentesco7, =1 Si es madre/padre
91. parentesco8, =1 Si es suegro(a)
92. parentesco9, =1 Si es hermano(a)
93. parentesco10, =1 Si es cuñado(a)
94. parentesco11, =1 Si es otro miembro de la familia
95. parentesco12, =1 Si es otro miembro no familiar
96. idhogar, Identificador a nivel de hogar

97. hogar_nin, Número de niños de 0 a 19 años en el hogar
98. hogar_adul, Número de adultos en el hogar
99. hogar_mayor, Número de individuos de 65 años o más en el hogar
100. hogar_total, Número total de individuos en el hogar
101. dependency, Tasa de dependencia, calculada como (número de miembros del hogar menores de 19 o mayores de 64)/(número de miembros del hogar entre 19 y 64)
102. edjefe, Años de educación del jefe de hogar masculino, basado en la interacción de escolari (años de educación), jefe de hogar y género, sí=1 y no=0
103. edjefa, Años de educación de la jefa de hogar femenina, basado en la interacción de escolari (años de educación), jefa de hogar y género, sí=1 y no=0
104. meaneduc, Años promedio de educación para adultos (18+)
105. instlevel1, =1 Sin nivel de educación
106. instlevel2, =1 Primaria incompleta
107. instlevel3, =1 Primaria completa
108. instlevel4, =1 Nivel secundario académico incompleto
109. instlevel5, =1 Nivel secundario académico completo
110. instlevel6, =1 Nivel secundario técnico incompleto
111. instlevel7, =1 Nivel secundario técnico completo
112. instlevel8, =1 Educación universitaria y superior
113. instlevel9, =1 Educación superior de posgrado
114. bedrooms, Número de habitaciones
115. overcrowding, # personas por habitación
116. tipovivi1, =1 Casa propia y completamente pagada
117. tipovivi2, "=1 Casa propia, pagando en cuotas"
118. tipovivi3, =1 Alquilada
119. tipovivi4, =1 Precaria
120. tipovivi5, "=1 Otro(asignada, prestada)"
121. computer, =1 Si el hogar tiene computadora portátil o de escritorio
122. television, =1 Si el hogar tiene televisión
123. mobilephone, =1 Si tiene teléfono móvil
124. qmobilephone, # de teléfonos móviles
125. lugar1, =1 Región Central
126. lugar2, =1 Región Chorotega
127. lugar3, =1 Región Pacífico Central
128. lugar4, =1 Región Brunca
129. lugar5, =1 Región Huetar Atlántica
130. lugar6, =1 Región Huetar Norte
131. area1, =1 Zona urbana
132. area2, =2 Zona rural
133. age, Edad en años
134. SQBescolari, Escolari al cuadrado
135. SQBage, Edad al cuadrado
136. SQBhogar_total, hogar_total al cuadrado
137. SQBedjefe, edjefe al cuadrado

138. SQBhogar_nin, hogar_nin al cuadrado
139. SQBovercrowding, hacinamiento al cuadrado
140. SQBdependency, dependencia al cuadrado
141. SQBmeaned, cuadrado de la educación promedio de adultos (>=18) en el hogar
142. agesq, Edad al cuadrado

Además, el archivo *train* trae una columna adicional llamada *target* la cual es el nivel real de vulnerabilidad de las familias.

3. Métricas

La métrica de medición principal para el modelo será la medida-F1 (F1-score). Que es la determinación de un valor único ponderado entre la precisión y la exhaustividad de un modelo, y este dado por la formula:

$$F_1 = 2*\frac{\text{presicion*recall}}{\text{presicion} + \text{recall}} = \frac{2\text{tp}}{2\text{tp} + \text{fp} + \text{fn}}$$

Donde:
tp: true positive.
fp: false positive.
fn: false negative.

El "recall" es una medida que evalúa la cantidad de datos correctos predichos por el modelo en relación con el total de datos correctos disponibles. Por otro lado, la "precisión" se refiere al número de datos correctos predichos por el modelo en comparación con el total de datos que el modelo predijo en general.

Esta medida se utiliza porque el objetivo principal es clasificar los hogares de la manera más precisa posible en uno de los niveles de vulnerabilidad. Es crucial para quienes toman decisiones asegurarse de que la ayuda se entregue de manera efectiva a quienes más la necesitan.

4. Desempeño

Se espera que este modelo prediga el nivel de vulnerabilidad de los hogares, con esta información los tomadores de decisiones podrían hacer políticas y entregas de recursos humanitarios enfocados en aquellas familias que realmente los necesitan, es irrelevante la cantidad de hogares que quede en cada uno de los niveles de vulnerabilidad, lo realmente importante es que cada hogar quede correctamente clasificado para que no exista una mala alocación de recursos, es decir se espera que el modelo tenga una métrica-F1 alta (cercana a 1), que le de confianza a los tomadores de decisiones.

5. Bibliografía

- Wood, T. *What is the F-score?* Obtenido de https://deepai.org/machine-learning-glossary-and-terms/f-score
- Fabián Sánchez, Gary Soto, Julia Elliott, Luis Tejerina, Phil Culliton. (2018). Costa Rican Household Poverty Level Prediction. Kaggle. https://kaggle.com/competitions/costa-rican-household-poverty-prediction