Entrega final

POR:

Alejandro Sarasti Sierra

MATERIA:

Modelos y simulación I

PROFESOR:

Raul Ramos Pollan

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

MEDELLÍN

Tabla de contenido

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.

1.1. 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 que se pueden ver en el anexo 1.

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

1.2. Métrica

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:

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

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.

2. Exploración

Para comprender el problema lo primero que se hizo fue hacer una exploración minuciosa de las variables objetivo, categóricas y numéricas. Una dificultad que presentaba el dataset era su amplia cantidad de variables, por lo que una parte importante del trabajo fue conocer cada una de las variables que representaba y que buscaba expresar.

2.1. Variable objetivo (Niveles de pobreza)

El total de personas en ‘train’ que pertenecen a cada uno de los niveles de pobreza está representado por la siguiente tabla:

	count
No vulnerable	5996
Vulnerable	1597
Moderadamente pobre	1209
Extremadamente pobre	755

Se puede observar que los datos obtenidos por parte de los hogares no vulnerables es mayor comparado a los otros niveles, por lo que a la hora de crear nuestros modelos debemos tener cuidado con este hecho.

2.2. Variables Categóricas

Se detectaron más de 90 variables categóricas, cuyo análisis completo se encuentra en el notebook, en esta sección los limitaremos a referencia cual fue el proceso que se hizo para la exploración, debido a que se agruparon estas variables (solo con motivos descriptivos) en función a que tipo de preguntan respondían, de la siguiente manera:

1. ¿Qué tipo de hacinamiento tiene? 'hacdor' y 'hacapo'

2. ¿Que tienen los hogares? 'v14a', 'refrig', 'v18q', 'computer', 'television', 'mobilephone'

3. ¿De que estan hechos los hogares?

    3.1 Paredes: 'paredblolad', 'paredzocalo', 'paredpreb', 'pareddes', 'paredmad', 'paredzinc', 'paredfibras', 'paredother'

    3.2 Pisos: 'pisomoscer', 'pisocemento', 'pisoother', 'pisonatur', 'pisonotiene', 'pisomadera'

    3.3 Techos: 'techozinc', 'techoentrepiso', 'techocane', 'techootro'

4. ¿Como acceden a los servicios publicos?

    4.1 Agua: 'abastaguadentro', 'abastaguafuera', 'abastaguano'

    4.2 Electricidad: 'public', 'planpri', 'noelec', 'coopele'

    4.3 Saneamiento: 'sanitario1', 'sanitario2', 'sanitario3', 'sanitario5', 'sanitario6'

    4.4 Energía para cocinar: 'energcocinar1', 'energcocinar2', 'energcocinar3', 'energcocinar4'

    4.5 Gestión de residuos: 'elimbasu1', 'elimbasu2', 'elimbasu3', 'elimbasu4', 'elimbasu5', 'elimbasu6'

5. ¿Cuál es el estado de la vivienda?

    5.1 Paredes: 'epared1', 'epared2', 'epared3'

    5.2 Techo: 'etecho1', 'etecho2', 'etecho3'

    5.3 Piso: 'eviv1', 'eviv2', 'eviv3'

6. ¿Cuál es la composición de las familias en las viviendas?

    6.1 ¿En el hogar hay personas con discapacidad? 'dis'

6.2 ¿Cuál es el género de las personas de la vivienda? 'male', 'female'

6.3 ¿Cuál es el estado civil de las personas de la vivienda? 'estadocivil1', 'estadocivil2', 'estadocivil3', 'estadocivil4', 'estadocivil5', 'estadocivil6', 'estadocivil7'

7. ¿Cuál es el nivel educativo de las personas de la vivienda? 'instlevel1', 'instlevel2', 'instlevel3', 'instlevel4', 'instlevel5', 'instlevel6', 'instlevel7', 'instlevel8', 'instlevel9'

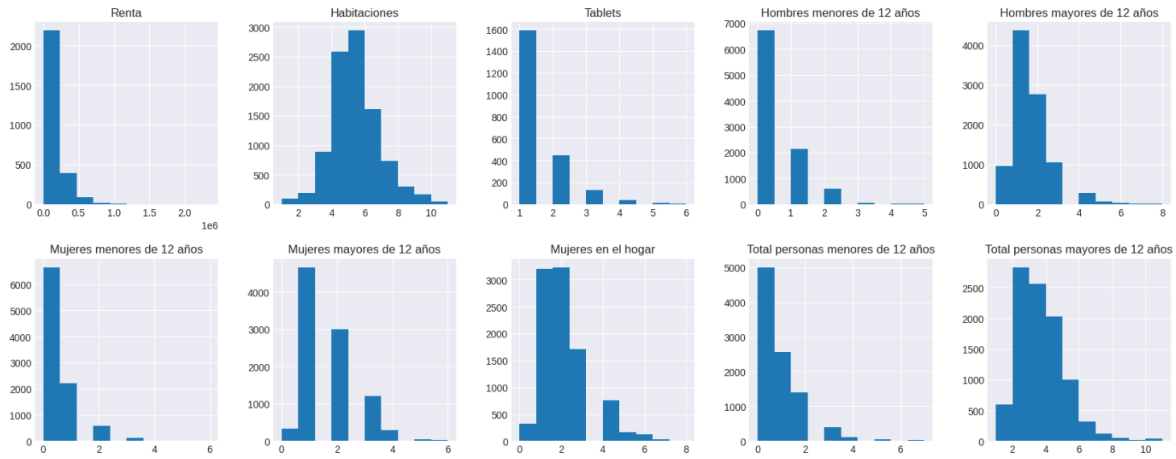
8. ¿La vivienda es propiedad de la familia? 'tipovivi1', 'tipovivi2', 'tipovivi3', 'tipovivi4', 'tipovivi5'

9. ¿Cuál es la ubicación de la vivienda? 'lugar1', 'lugar2', 'lugar3', 'lugar4', 'lugar5', 'lugar6'

9.1 ¿Es una ubicación urbana o rural? 'area1', 'area2'

2.3. Variables numéricas

En el caso de este tipo de variables se hizo una exploración básica que debe ser profundizada en las siguientes entregas, se reviso la distribución de las variables y su correlación de forma gráfica, se evidencia que hay al menos dos de estas variables que presentan la misma información, por lo que se debe revisar que se debe hacer con ella para que no afecte negativamente a los modelos.



En la última iteración se decidió eliminar aquellas variables que entregaban la misma información, además se evidencio que existían tres variables ‘dependency’, ‘edjefe’ y ‘edjefa’, que tenían valores continuos, pero también podían tener dos valores categóricos, ‘si’ o ‘no’.

3. Iteraciones de desarrollo

Primero, se seleccionaron 5 modelos supervisados para tener un acercamiento inicial que nos permitiera conocer un funcionamiento preliminar de ellos, los

modelos seleccionados fueron: regresión logística, árbol de decisiones, bosque aleatorio, k-vecinos y una maquina de vectores de soporte. Antes de efectivamente correr estos modelos se debía manipular los datos para corregir problemas encontrados en la exploración, y transformarlos de manera que puedan ser procesados por la librería ´sklearn´ la cual fue la utilizada para correr estos 5 modelos.

3.1. Importación de datos, corrección y transformación

En la fase inicial del proceso, se procedió a realizar una descarga nuevamente de los conjuntos de datos "train" y "test". Para garantizar la calidad de los datos, se llevó a cabo una exhaustiva revisión y limpieza de estos conjuntos. El primer paso fue identificar aquellas columnas que presentaban una alta proporción de datos nulos, estableciendo un umbral del 70%. Tres columnas superaron este umbral: 'v2a1', 'v18q1' y 'rez_esc'. Por lo tanto, se tomó la decisión de eliminarlas, ya que su contenido carecía significativamente de información útil para el análisis y la predicción.

Posteriormente, se abordó la gestión de valores faltantes en aquellas columnas que contenían menos del 5% de datos ausentes. Dentro de este grupo, se identificaron dos columnas: una era categórica y la otra numérica. Para la columna categórica, se procedió a reemplazar los valores faltantes por la moda de la columna, mientras que en la columna numérica, se optó por sustituir los valores nulos por la media correspondiente.

Además, se enfrentó una situación peculiar con tres columnas que originalmente contenían datos numéricos, pero que también tenían la posibilidad de tomar valores categóricos, específicamente los valores "sí" o "no". En este caso, se realizó una transformación convencional convirtiendo estos valores categóricos en 1 y 0 respectivamente, para facilitar su manipulación en el análisis posterior.

Finalmente, tras esta fase de preparación de datos, se llevó a cabo la división del conjunto de entrenamiento en la variable objetivo ('Target') y las variables independientes. Se tomó la decisión de excluir las columnas "Id" e "idhogar" de las variables independientes, debido a que representaban valores individuales o grupales que no aportaban información relevante al proceso predictivo de clasificación que se buscaba realizar.

3.2. Modelos supervisados de clasificación

En esta etapa, se implementó el método de validación cruzada con 10 "pliegues" para evaluar el rendimiento de los modelos mencionados previamente. La métrica seleccionada para esta evaluación fue el "F1-score", en línea con la métrica utilizada en la competencia de Kaggle. Es importante destacar que, en esta fase, se emplearon los datos de "train". Aunque no se llevó a cabo una división explícita en conjuntos de entrenamiento y prueba, esto se debe al uso del método mencionado anteriormente, el cual internamente realiza divisiones de este tipo durante el

proceso de validación cruzada.

Tras ajustar cada uno de los modelos, se generó un ranking de rendimiento en términos del "F1-score", ordenado de menor a mayor rendimiento. Los resultados fueron los siguientes: la Support Vector Machine (SVM) se ubicó en la posición de menor rendimiento, seguida por la regresión logística, el árbol de decisión, K-vecinos y, finalmente, el bosque aleatorio. Específicamente, el bosque aleatorio obtuvo una métrica cercana a 0.35, siendo el modelo con el mejor desempeño dentro de este conjunto de evaluaciones.

3.3. Modelos supervisados de clasificación

Considerando la posibilidad de que el bajo rendimiento de los modelos estuviera influenciado por discrepancias en los rangos de las variables independientes o por la presencia de sesgos significativos, se optó por implementar dos técnicas de escalado para normalizar los valores de las variables. La primera técnica empleada fue el "StandardScaler" de la biblioteca sklearn, mientras que la segunda fue el "MinMaxScaler", también parte de la misma librería. Ambos métodos demostraron ser eficaces en el contexto de los objetivos planteados, logrando homogeneizar la distribución de la mayoría de las variables, alineándolas a rangos similares y haciéndolas más comparables con distribuciones conocidas.

Tras aplicar esta transformación a los datos, se procedió a ejecutar nuevamente todos los modelos. Se observó que, en ambos casos de escalamiento, el bosque aleatorio mantuvo el mejor rendimiento en comparación con los otros modelos. Sin embargo, se destacó un aspecto significativo: al utilizar el "MinMaxScaler", el rendimiento del bosque aleatorio mejoró en aproximadamente 2 puntos en términos de la métrica seleccionada. Esta mejora indicó una mayor capacidad predictiva y un mejor ajuste del modelo tras la normalización de los datos utilizando este método específico de escalamiento.elegida.

3.4. Importación de datos, corrección y transformación

En esta etapa, se procedió a optimizar los hiperparámetros de nuestro modelo más prometedor, el bosque aleatorio. Específicamente, se buscó ajustar el número de estimadores, la profundidad máxima, el máximo número de características consideradas en cada división, el criterio de división (Gini o entropía), el mínimo número de hojas y el mínimo número de muestras requeridas para dividir un nodo. Para realizar esta optimización, se implementó una búsqueda aleatoria de hiperparámetros que opera sobre una matriz que contiene posibles valores para cada uno de estos hiperparámetros a optimizar.

Los valores considerados fueron los siguientes:

Número de estimadores: de 10 a 100 en incrementos de 10.
Profundidad máxima: desde 1 hasta 20, con la opción de no establecer un límite.

Criterio de división: Gini o entropía, como mencionado anteriormente.
Mínimo de muestras: 2, 5 o 10.
Mínimo de hojas: 1, 2 o 4.

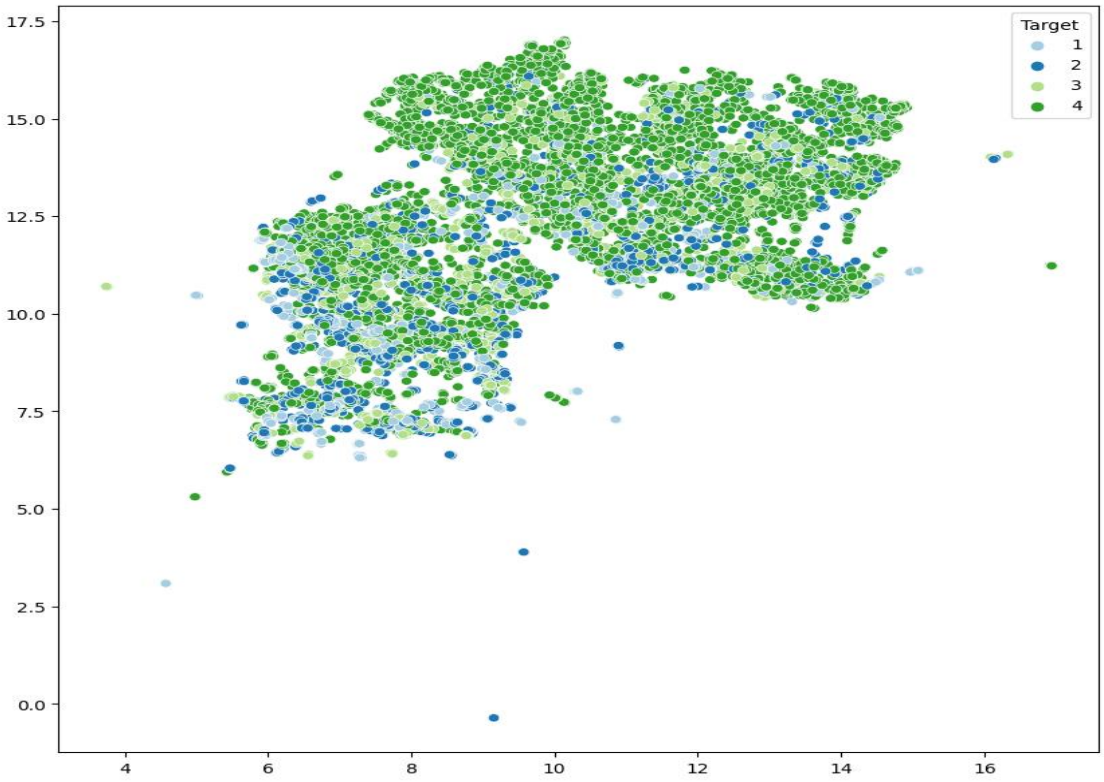
Es importante mencionar que este enfoque de búsqueda aleatoria no realiza una exploración exhaustiva de todas las combinaciones posibles. Sin embargo, ofrece mayor rapidez y requiere menos recursos computacionales en comparación con otros métodos de búsqueda más intensivos en recursos. Tras implementar esta técnica utilizando una configuración de 3 "pliegues" para explorar 100 combinaciones candidatas (300 en total), se obtuvo como resultado un bosque con 10 estimadores (árboles), sin límite de profundidad, utilizando el criterio de entropía, con un mínimo de 5 muestras para la división de un nodo y un mínimo de 1 hoja.

A pesar de este proceso, se observó una mejora mínima en el rendimiento del modelo. Esta escasa mejora puede atribuirse a la naturaleza aleatoria de los modelos. Por consiguiente, se decidió utilizar los valores por defecto de los hiperparámetros del modelo, ya que no se logró obtener una mejora significativa en el desempeño del bosque aleatorio a través de esta técnica de optimización.

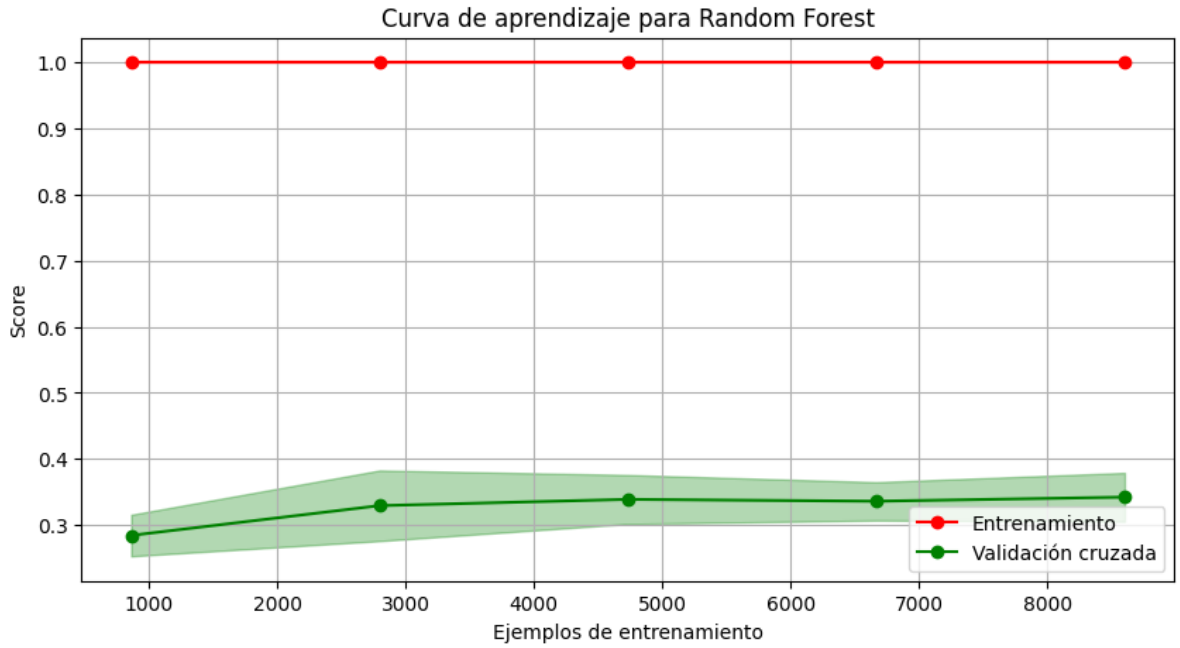
3.5. Reducción de dimensionalidad

Con el propósito de incrementar aún más el rendimiento del modelo, se optó por explorar la reducción de dimensionalidad mediante "UMAP". Este enfoque nos permitió transformar una matriz original de dimensiones 9557x137 a una matriz de 9557x2. Aunque estas dos últimas dimensiones carecen de un significado directo, en muchos casos, esta reducción puede conducir a una mejora en el rendimiento del modelo. No obstante, al examinar la gráfica resultante, se evidenció que los datos estaban altamente entremezclados, lo cual no sugería un posible aumento en el rendimiento del modelo.

De hecho, tal como se pudo apreciar en la representación gráfica, al reajustar todos los modelos tras esta transformación, se observó que todos mostraron un rendimiento inferior en comparación con las métricas anteriores. Esta disminución en el rendimiento llevó a la conclusión de descartar esta opción de reducción de dimensionalidad mediante "UMAP", ya que no mostraba indicios de mejorar el desempeño de los modelos, sino todo lo contrario.



3.6. Curva de aprendizaje del bosque aleatorio



En el gráfico previo, que representa la curva de aprendizaje de nuestro mejor modelo hasta la fecha, se realizaron ajustes con modelos de boosting que no resultaron exitosos. Se puede observar claramente que el modelo está experimentando sobreajuste, indicado por el puntaje de entrenamiento significativamente alto en contraste con el puntaje de validación considerablemente bajo. Este desequilibrio sugiere que el modelo se ha adaptado demasiado a los datos de entrenamiento y tiene dificultades para generalizar a nuevos datos.

Las posibles soluciones para abordar este problema de sobreajuste incluyen la adquisición de más datos, la reducción de la complejidad del modelo o el aumento de la regularización. Sin embargo, considerando la información obtenida durante la competencia, donde el mejor modelo utilizado logra métricas aproximadas al 0.45, se vuelve más plausible que la principal causa del problema sea la insuficiencia de datos disponibles para el entrenamiento del modelo.

Aunque técnicas como la reducción de complejidad o la regularización podrían ser beneficiosas en circunstancias diferentes, en este caso particular, el bajo rendimiento del modelo comparado con el resultado óptimo esperado sugiere que la limitación principal probablemente sea la cantidad limitada de datos para entrenar el modelo. En este contexto, la adquisición de más datos se perfila como la solución más prometedora para mejorar la capacidad predictiva y reducir el sobreajuste del modelo.Se procede a usar el modelo de random forest, que fue el que mejor se comportó.

4. Retos y consideraciones de despliegue

- Escasez de Datos:

La limitación más significativa identificada durante el desarrollo del modelo fue la disponibilidad limitada de datos. Este desafío puede impactar negativamente la capacidad del modelo para generalizar patrones y comportamientos en datos no vistos. Obtener más datos relevantes y de calidad puede mejorar significativamente el rendimiento y la precisión del modelo.

- Sobreajuste y Selección de Modelo:

Se evidenció sobreajuste en los modelos, lo que indica la necesidad de una selección más adecuada del modelo o la implementación de técnicas de regularización para mejorar la generalización a nuevos datos. La identificación y evaluación cuidadosa de diferentes algoritmos de aprendizaje automático puede ser crucial para mitigar este problema.

- Optimización de Hiperparámetros:

A pesar de los intentos de optimización de hiperparámetros, no se logró mejorar significativamente el rendimiento del modelo. Es crucial realizar una búsqueda más exhaustiva o considerar técnicas avanzadas de optimización para encontrar la combinación óptima de hiperparámetros que maximicen el rendimiento sin incurrir en sobreajuste.

- Reducción de Dimensionalidad:

Se exploró la reducción de dimensionalidad con "UMAP", sin embargo, esta técnica no ofreció mejoras significativas y, de hecho, empeoró el rendimiento de los modelos. Se deben considerar cuidadosamente las técnicas de reducción de dimensionalidad y evaluar su impacto antes de implementarlas en un entorno de producción.

- Interpretación de Resultados:

Es esencial comprender la interpretación de los resultados del modelo para explicar de manera clara y coherente las predicciones a las partes interesadas y usuarios finales. Esto implica la generación de informes y visualizaciones que ayuden a comunicar las decisiones basadas en el modelo de manera comprensible y transparente.

5. Conclusiones

- Limitación de Datos: La escasez de datos adecuados afecta la capacidad del modelo para generalizar patrones, siendo la adquisición de más datos una solución fundamental.
- Exploración de Técnicas: La exploración de reducción de dimensionalidad y optimización de modelos no proporcionó mejoras significativas, lo que enfatiza la importancia de evaluar críticamente el impacto de las técnicas antes de implementarlas.
- Utilización de modelos boosting: Son una excelente alternativa, pero se debe profundizar en el conocimiento de estos para hallar los hiperparametros adecuados.
- Aprendizaje: Las competencias de Kagel permiten nutrirse no solo del reto que uno debe resolver, sino del conocimiento de los cientos de personas que también lo están intentando o ya lo hicieron, pudiendo aprender y conocer de códigos que quedan almacenados en cada competencia.
- Rendimiento general: El puesto alcanzado en la competencia es entre 300-400, aún hay espacio para mejora, pero fue un buen primer intento, teniendo en cuenta que habían mas de 700 participantes.

6. 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

1. Anexo

Anexo A

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