

Predicción del nivel de pobreza en los hogares de Costa Rica.

Proyecto Final Aprendizaje de Máquina

Dante Ruiz Martínez 183340
Laura López S. Jácome 144089

Problemática

La pobreza en Costa Rica
aumentó en 2018:
21% de los hogares se
encuentra en estas
condiciones:



→
Gobiernos e instituciones
como el Inter-American
Development Bank tiene
programas de ayuda
social.



→
Programas de ayuda social tienen
dificultades para asegurarse de
que las personas adecuadas
reciban la ayuda necesaria.
**Las personas más pobres no
pueden proporcionar los
registros de ingresos y gastos
necesarios para demostrar que
califican.**



Objetivo

Desarrollar un modelo de aprendizaje de máquina que pueda predecir el nivel de pobreza (nivel de necesidad) de los hogares utilizando las características tanto del hogar como de los individuos.

Datos

Se obtuvieron de la competencia de Kaggle “Costa Rican Household Poverty Level Prediction”. Las **variables explicativas** proporcionadas se dividen en **2 categorías**:

1. Características del **HOGAR**



2. Características del **INDIVIDUO**



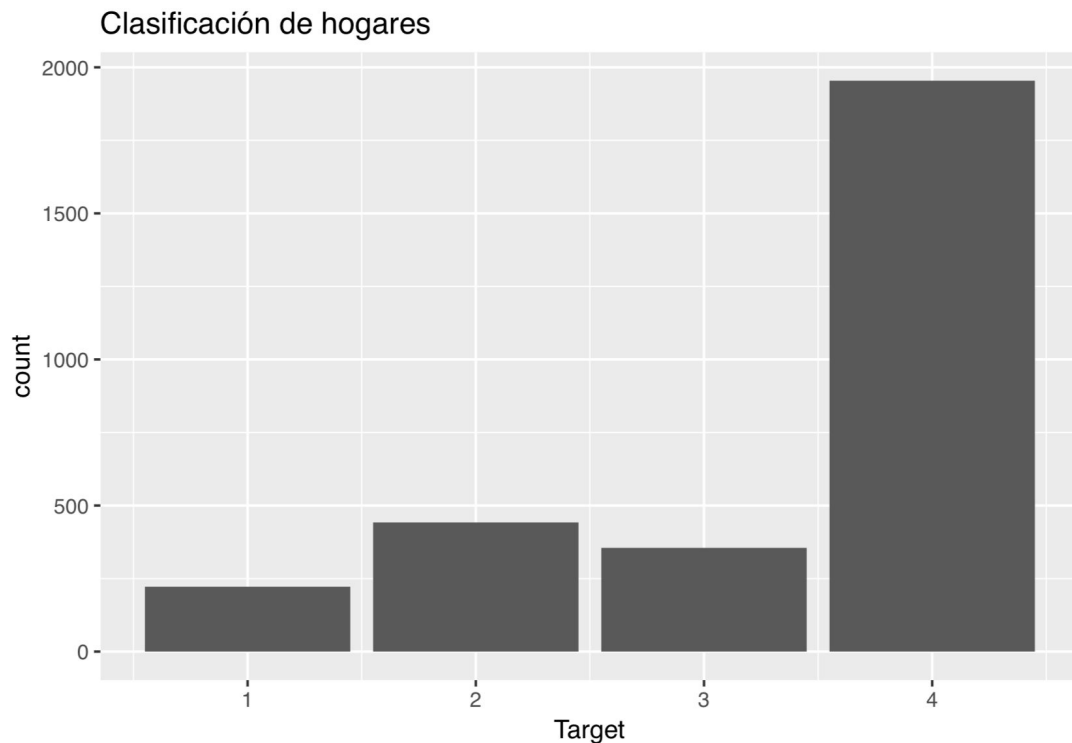
TRAIN SET

- **9557 filas (individuos) y 143 variables.**
- Variable Target es la variable de interés. Es una variable categórica con 1 siendo la pobreza extrema y 4 sin riesgo.
- **Se debe entrenar los modelos ÚNICAMENTE CON LOS JEFES DE FAMILIA.**

TEST SET

- **23,856 filas (individuos) y 142 columnas.**
- Se debe hacer 1 predicción por cada **INDIVIDUO** del conjunto de prueba.

Naturaleza del problema de clasificación



Nivel de pobreza	Conteo	Proporción
1	755	0.07
2	1597	0.16
3	1209	0.12
4	5996	0.62

Procedimiento para atacar el problema



TOP 20 de variables predictivas

Educación

Edad

Calidad del hogar

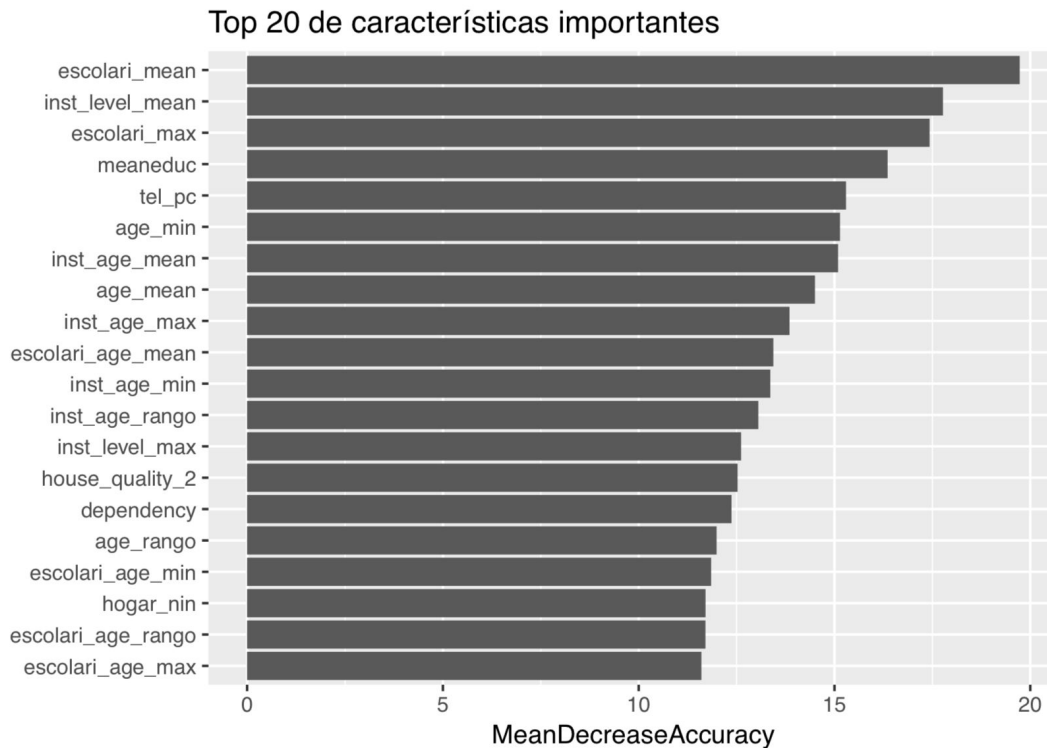
Dependencia
económica

Teléfono

Hijos



variable



Macro F1 Score

$$F_1 = \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

$$\text{Macro F1} = \frac{\text{F1 Class 1} + \text{F1 Class 2} + \text{F1 Class 3} + \text{F1 Class 4}}{4}$$

Resultados

Macro
F-Score

0.4

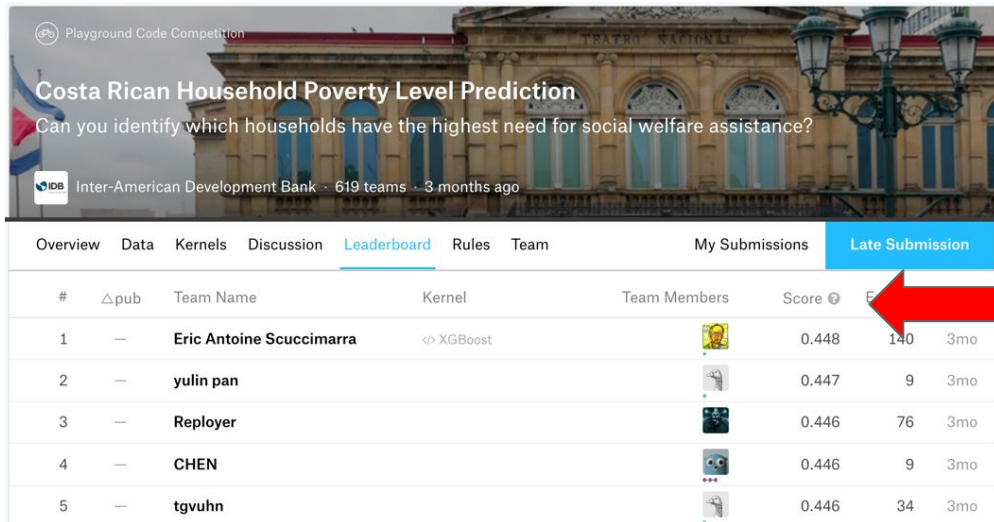
XGBOOST

0.39

Logit

0.38

Random Forest



The screenshot shows the Kaggle Playground Code Competition interface. The title is 'Costa Rican Household Poverty Level Prediction' with the subtitle 'Can you identify which households have the highest need for social welfare assistance?'. It is organized by the Inter-American Development Bank and has 619 teams. The 'Leaderboard' tab is selected, showing a table of top teams. A red arrow points to the 'F' column, which represents the Macro F-Score.

#	pub	Team Name	Kernel	Team Members	Score	F	Time
1	—	Eric Antoine Scuccimarra	XGBoost		0.448	140	3mo
2	—	yulin pan			0.447	9	3mo
3	—	Reployer			0.446	76	3mo
4	—	CHEN			0.446	9	3mo
5	—	tgvuhn			0.446	34	3mo

El ganador de la competencia en Kaggle alcanzó un 0.44

Conclusiones

- El XGBOOST es el algoritmo que puede alcanzar el mejor desempeño en este problema..
- El truco para subir más es score es en la ingeniería de características y construcción de variables.
- Para este tipo de problema los árboles y el xgboost pueden lidiar con el problema de datos faltantes y con el desbalance en las clases

¡Gracias!

Anexos

XGBoost

Table 16: Resultados del XGBoost

Entrena	Prueba	iteraciones	Eta	Lambda	Subsamples	Max_depth	Gamma	feature_subsample
0.9499116	0.400	2000	0.030	0.20	1.00	3	0.0	1.00
0.9900000	0.398	3000	0.030	0.20	0.20	3	0.0	1.00
0.8900000	0.398	3000	0.030	0.20	0.20	4	0.5	0.25
0.9500000	0.396	2000	0.030	0.20	0.20	3	0.0	1.00
1.0000000	0.396	3000	0.030	0.20	0.20	4	0.0	1.00
0.7900000	0.396	2500	0.030	0.20	0.20	4	0.5	0.05
0.7600000	0.396	2500	0.030	0.20	0.20	4	0.5	0.04
0.9900000	0.395	3000	0.030	0.20	0.20	4	0.5	1.00
0.9790000	0.394	2000	0.030	0.19	0.54	3	0.0	1.00
0.9500000	0.393	2000	0.030	0.19	1.00	3	0.0	1.00
0.9490000	0.392	2000	0.030	0.21	1.00	3	0.0	1.00
0.9383542	0.391	2000	0.030	0.00	1.00	3	0.0	1.00
0.7883000	0.389	1000	0.030	0.00	1.00	3	0.0	1.00
0.7800000	0.389	500	0.030	0.20	0.20	4	0.5	0.05
0.6699000	0.388	1000	0.030	0.20	0.20	4	0.5	0.05
0.9497000	0.386	2000	0.030	0.15	1.00	3	0.0	1.00
0.5700000	0.371	500	0.030	0.20	0.20	4	0.5	0.05
0.4849563	0.344	1000	0.003	0.00	1.00	3	0.0	1.00
0.4466259	0.337	200	0.003	0.00	1.00	3	0.0	1.00
0.4405086	0.331	100	0.003	0.00	1.00	3	0.0	1.00

20 modelos

Resultados Logit Multinomial

Table 14: Resultados de los 5 mejores modelos con Regresión Logística

n_variables	F1_Score_Macro_CV	F1_Score_Macro_Prueba
10	0.337	0.298
40	0.375	0.310
222	0.389	0.385

Random Forest

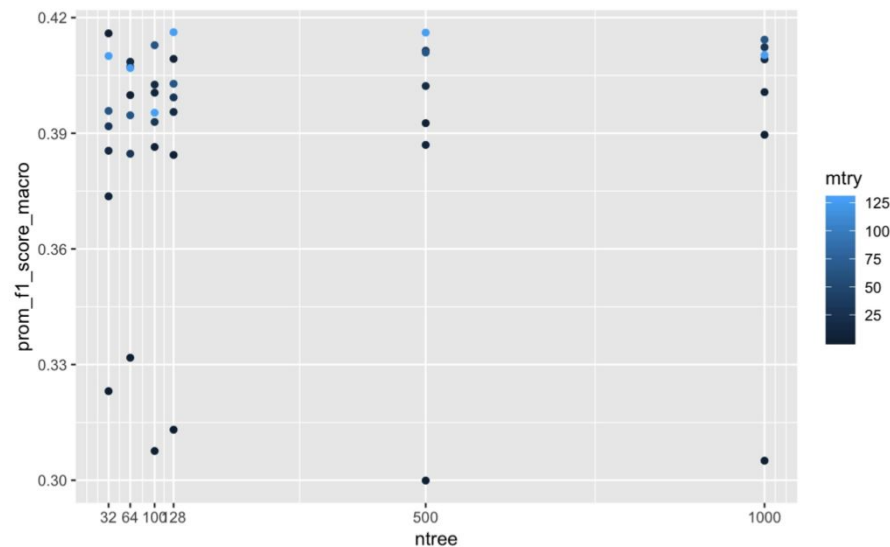


Table 15: Resultados de los 5 mejores modelos con random forest

ntree	mtry	cv	kaggle
128	128	0.416	0.376
500	128	0.416	0.373
32	4	0.416	0.347
1000	64	0.414	0.374
100	64	0.413	0.376

Desempeño de los 42 modelos.