Poverty Level Prediction for Costa Rican Households

Efraín García Valencia
CC. 1001370984

1. Descripción del problema predictivo a resolver:

El problema predictivo a resolver es la mejora de un algoritmo utilizado para determinar la calificación de ingresos de las familias más pobres en América Latina, específicamente en Costa Rica. Actualmente, se utiliza un método llamado "Proxy Means Test" (PMT), que utiliza atributos observables del hogar, como el material de las paredes y el techo o los activos presentes en el hogar, para clasificar a las familias y predecir su nivel de necesidad económica. Sin embargo, este método no es lo suficientemente preciso, especialmente a medida que la población crece y la pobreza disminuye. El objetivo es mejorar el rendimiento de este algoritmo utilizando métodos más avanzados y basados en un conjunto de datos de características de hogares costarricenses.

2. Dataset a usar:

Para la realización de este proyecto se usará el dataset de la competicion de Kaggle (https://kaggle.com/competitions/costa-rican-household-poverty-prediction) que tiene más de 9000 instancias y 143 columnas, cada una de estas instancias representa una persona, sin embargo es importante aclarar que solo se evalua el valor de clasificación asignado a la persona cabeza de familia.

Algunas de las columnas más importantes que encontramos en el Dataset son:

- **Id:** identificación única para cada instancia (persona)
- **Target:** Clasificación del nivel de pobreza de 1 a 4, siendo 1 pobreza extrema y 4 hogares no vulnerables
- **Idhogar:** Identificación propia para cada hogar, es decir, todas las personas que pertenezcan a un mismo hogar tendrán el mismo Idhogar; especialmente útil para hacer análisis de las características conjuntas del hogar.
- **Parentesco1:** Nos permite identificar si esta persona es la cabeza de hogar

Además tenemos muchas otras variables que a grandes rasgos nos comunican:

- **Condiciones de Vivienda:** Las variables como "Monthly rent payment" (pago mensual de alquiler), "rooms" (número de habitaciones) y las relacionadas con las características de las paredes y el piso ofrecen información sobre las condiciones de vivienda de los hogares.
- **Composición del Hogar:** Las variables que desglosan la composición del hogar en términos de género y edad, como "Total males in the household" (total de hombres en el hogar) y "Total females in the household" (total de mujeres en el hogar), son cruciales para comprender la dinámica familiar.

- **Nivel de Educación:** Las variables cómo "years of schooling" (años de escolaridad) y "Years behind in school" (años de retraso en la educación) proporcionan información sobre el nivel educativo de los residentes, lo que puede estar relacionado con la pobreza y la necesidad de asistencia.
- **Características de la Comunidad:** Variables como la ubicación geográfica ("region") y la presencia de tecnología en el hogar (como "computer" y "television") pueden indicar el acceso a servicios y recursos en la comunidad.

3. Métricas de desempeño requeridas:

- **Métrica de Machine Learning:** La métrica de desempeño requerida es el "macro F1 score". El F1 score es una medida que combina la precisión y la exhaustividad del modelo y es particularmente útil cuando hay un desequilibrio en las clases objetivo. El "macro F1 score" considera el promedio de los valores F1 para todas las clases, lo que significa que se valora igualmente el rendimiento en todas las categorías. Esto es importante para garantizar que el modelo sea equitativo en su capacidad para clasificar a las familias en diferentes niveles de necesidad económica.

- **Métrica de Negocio:** Una métrica de negocio relevante podría ser la tasa de error en la asignación de beneficios de asistencia social. El objetivo principal de este problema es mejorar la precisión en la calificación de ingresos de las familias más pobres. Por lo tanto, reducir la tasa de error en la asignación de beneficios es esencial para garantizar que las ayudas lleguen a quienes más las necesitan de manera efectiva.

4. Desempeño deseable en producción:

El desempeño deseable en producción sería un alto Macro F1 score que indique que el modelo es capaz de clasificar con precisión a las familias en diferentes niveles de necesidad económica en Costa Rica y, potencialmente, en otros países. Un Macro F1 score superior o igual a 0.7 nos indicaría que el algoritmo es efectivo en identificar a las familias más pobres y proporcionarles la asistencia necesaria de manera equitativa y justa.

Un rendimiento alto es fundamental porque impacta directamente en la eficacia de los programas de asistencia social. Si el algoritmo tiene un alto F1 score, se minimizan los errores en la asignación de beneficios, lo que significa que las familias necesitadas recibirán la ayuda adecuada. Esto no solo beneficiará a las familias en situación de pobreza, sino que también garantizará un uso más eficiente de los recursos de desarrollo financiero en América Latina y el Caribe, lo que es fundamental para el éxito de los programas de desarrollo. Además, si el modelo demuestra ser exitoso en Costa Rica, podría ser implementado en otros países que enfrentan problemas similares, lo que ampliaría su impacto y utilidad.