Primer entrega de proyecto

Cristian Camilo Serna Betancur[1] Diego Alonso Herrera Ramírez[2] Sharid Samantha Madrid Ospina[3]

I. Problema predictivo

DADAS las características socioeconómicas de un hogar (tales como nivel educativo, situación laboral, acceso a servicios de salud, entre otros), se busca predecir si dicho hogar se encuentra en una situación de pobreza o vulnerabilidad económica.

II. Dataset

El conjunto de datos proviene de la Gran Encuesta Integrada de Hogares proporcionada por el Departamento Administrativo Nacional de Estadística (DANE) en colaboración con el Municipio de Medellín.

A. Descripción

Esta encuesta proporciona información socioeconómica detallada con énfasis en la fuerza laboral. El conjunto de datos está desagregado por el Municipio de Medellín y sus 16 comunas. Contiene información demográfica y laboral recopilada de los hogares encuestados en Medellín.

B. Enlace del dataset

Gran Encuesta Integrada de Hogares

C. Número de filas

5138

D. Número de columnas

397

E. Porcentaje de columnas categóricas

7.05%

F. Columnas relevantes seleccionadas

1) DIRECTORIO: Llave de vivienda.
2) ORDEN: Orden.
3) HOGAR: Hogar.
4) SECUENCIA_P: Llave de hogar.
5) SEGMENTO: Segmento.
6) AREA: Área.
7) MES: Mes.
8) DPTO: Departamento.
9) MPIO: Municipio.
10) PERIODO: Periodo.
11) TRIMESTRE: Trimestre.
12) ANIO: Año.
13) P6016: Número de orden de la persona que proporciona la información.
14) P6020: Sexo.
15) P6030S1: Cuál es la fecha de nacimiento de...? Mes.
16) P6030S2: P6030S2.
17) P6030S3: Cuál es la fecha de nacimiento de...? Año.
18) P6040: ¿Cuántos años cumplidos tiene...? Si es menor de 1 año, escriba 00.
19) P6050: Cuál es el parentesco de... con el jefe o jefa del hogar?
20) P6090: ¿está afiliado, es cotizante o es beneficiario de alguna entidad de seguridad social en salud?
21) P6100: A cuál de los siguientes regímenes de seguridad social en salud está afiliado.
22) P6125: En los últimos doce meses dejó de asistir al médico o no se hospitaliza, por no tener con qué pagar estos servicios en la EPS o ARS?
23) P6160: ¿Sabe leer y escribir?
24) P6170: Actualmente... ¿asiste al preescolar, escuela, colegio o universidad?
25) P6210: Cuál es el nivel educativo más alto alcanzado por... y el último año o grado aprobado en este nivel?
26) p6210s1: Cuál es el nivel educativo más alto alcanzado por... y el último año o grado aprobado en este nivel? Grado o Años de escolaridad.
27) comuna: Comuna.
28) DSCY: Desempleo coyuntural.
29) P9460: Desocupados.
30) P7422: Recibió o ganó el mes pasado ingresos por concepto de trabajo?
31) DSI: DSI.
32) P744: Si le hubiera resultado algún trabajo a... ¿estaba disponible la semana pasada para empezar a trabajar?
33) P7440: ¿Cuánto hace que trabajó por última vez?
34) P7450: Por qué motivo o razón principal... dejó ese trabajo?
35) P7450S1: Por qué motivo o razón principal... dejó ese trabajo? Cuál?
36) P7452: Después de su último trabajo, ¿ha hecho alguna diligencia para conseguir otro trabajo o instalar un negocio?
37) P7454: Ha buscado trabajo alguna vez?
38) P7456: ¿Cuánto hace que buscó trabajo por última vez?
39) P7458: Por qué razón principal... dejó de buscar trabajo?
40) P7458S1: Por qué razón principal... dejó de buscar trabajo? Cuál?
41) P7472: Recibió o ganó el mes pasado ingresos por concepto de trabajo?
42) P7472S1: Recibió o ganó el mes pasado ingresos por concepto de trabajo? ¿Cuánto? $.

III. Métricas de desempeño

Como métrica de machine learning, utilizaremos principalmente la Función de Costo Logística o **Log Loss** para evaluar el rendimiento de nuestro modelo. Además, también consideraremos el Error Cuadrático Medio (**ECM**) como métrica complementaria.

A. Función de Costo Logística o Log Loss

La métrica principal para evaluar el modelo es la Log Loss. Esta métrica mide la eficiencia del modelo en la clasificación de hogares en situación de pobreza o vulnerabilidad económica. Cuanto más bajo sea el valor de Log Loss, mejor será el rendimiento del modelo. Por ejemplo, si obtenemos un valor de Log Loss de 0.2, significa que el modelo tiene una buena capacidad para predecir correctamente la situación de los hogares.

B. Error Cuadrático Medio (ECM)

Aunque la Log Loss es la métrica principal, también consideramos el ECM como métrica complementaria. El ECM mide la precisión de las predicciones del modelo en términos de las variables numéricas. En nuestro caso, el ECM podría proporcionar información adicional sobre la precisión de las predicciones en variables relevantes para la pobreza y la vulnerabilidad económica.

C. Métricas de negocio

El proyecto se considerará exitoso en términos de métricas de negocio si el modelo proporciona información relevante y útil para los investigadores y tomadores de decisiones en las comunidades vulnerables. Si el modelo es capaz de destacar patrones de desempleo en estos grupos vulnerables y facilita la toma de decisiones informadas, se considerará que cumple con su propósito principal. La efectividad del modelo se medirá por su capacidad para proporcionar conocimientos útiles y guiar futuras investigaciones y políticas de reducción del desempleo en estas comunidades.