Costa Rican Household Poverty Level Prediction

Entrega 2 Introducción a la inteligencia artificial para ciencias e ingenierías

Por: Alejandro Mesa Mesa, Esteban Alzate Pérez y Maria Paula Sedano Sarmiento

Resumen

En este avance se realizaron 3 notebooks de Python en Google Colab, que contienen un primer acercamiento a los datos, así como una manipulación, una limpieza y un análisis basados en los conocimientos adquiridos durante el curso y técnicas y metodologías aprendidas en el desarrollo de los laboratorios.

Progreso alcanzado

En primer lugar, se logró un mejor entendimiento y una mejora en la capacidad del manejo de GitHub, ya que ninguno de los integrantes sabía cómo usarlo previamente. Gracias a esto, se simplificó la comunicación entre los diferentes notebooks, así como el manejo de los data sets sin necesidad de descargarlos.

El trabajo realizado hasta el momento está dividido en 3 notebooks. El primero es 01_Exploración_de_datos.ipynb, que contiene un primer acercamiento a los datos mediante el uso de la función df.hist(), que da una mirada global a los datos y la frecuencia con la que aparecen dentro del dataset. Conociendo los datos, se comprobaron y ajustaron las condiciones iniciales del proyecto dadas por el docente, que decían que el dataset debía tener al menos un 5% de datos nulos en 3 columnas y por lo menos un 10% de columnas categóricas; esta última condición tuvo que ser simulada ya que no había ninguna columna categórica previamente. Para esto, se escogieron columnas que tuvieran por valor 0 y 1, y se cambiaron por “si” y “no”; por otro lado, se vio que en el dataset existían varias columnas que daban una clasificación para un solo tipo de dato, por ejemplo, el estado del piso de la vivienda, por lo cual, se crearon nuevas columnas que redujeran esta clasificación a diferentes categorías, para que luego pudieran ser nuevamente desglosadas. Finalmente, en el primer notebook, se hizo un análisis de los valores nulos, y cuál era la mejor forma de llenarlos al relacionarlos con valores de otras columnas.

Terminado el primer notebook, se guardó el dataset editado en GitHub para utilizarlo en el segundo, 02_Limpieza_de_datos.ipynb, donde se buscaron columnas que fueran redundantes entre sí, es decir, que dieran la misma información para luego eliminarlas. Se hizo el mismo proceso para algunas columnas que se consideraron irrelevantes. Se eliminaron columnas que correspondían a los datos de otras columnas elevados al cuadrado y se deshicieron las columnas categóricas creadas en el primer notebook.

En el tercer notebook, 03_Análisis_primario.ipynb, se utilizó el dataset obtenido de la limpieza de datos para hacer un análisis de cómo se relacionaban los diferentes datos con el objetivo, para tener una visión de cómo se puede relacionar cada variable con el nivel de pobreza de un hogar, así como tener un panorama de cuáles tienen más peso a la hora de implementar el modelo. También se juntaron varias de estas columnas que estaban muy

correlacionadas entre sí para continuar con la limpieza de datos y tener una visión más simplificada.

Dificultades y pasos a seguir

Las únicas dificultades que se presentaron durante el desarrollo de esta entrega, fueron el manejo y la visualización de una cantidad tan grande de datos y de columnas, que a simple vista puede resultar abrumador, además del manejo de funciones avanzadas de Pandas y de librerías de visualización y graficación de datos.

En cuanto a los modelos, aún no se tiene mucho conocimiento sobre la implementación de modelos predictivos, por lo que no se tocó este asunto para esta entrega. Se espera seguir adquiriendo conocimiento en este ámbito para poder implementar una predicción satisfactoria del problema que se plantea e implementarla a más problemas.

Al adquirir conocimiento de herramientas de inteligencia artificial, se espera también realizar una mejor limpieza y análisis primario de los datos, para ver si hay o no aspectos faltantes a tener en cuenta en los notebooks realizados.