Universidad de los Andes

Maestría Economía Aplicada: Big Data and Machine Learning for Applied Economics

Grupo: Laura Natalia Capacho (202121025), Sebastián David Beltrán (202121021) y

Yurani Gonzalez (201212100)

***Problem Set 1: Predicting Income***

El objetivo es construir un modelo predictivo para el ingreso individual, a partir de la Gran Encuesta Integrada de Hogares del DANE para 2018 en la ciudad de Bogotá.

**1. Adquisición de datos**

(a) El proceso de adquisición de datos está descrito en el *script* de R adjunto al documento.

(b) No hay restricciones para acceder a la información puesto que no existe ningún archivo (robots.txt) asociado a la raíz del sitio web que indique restricciones para rastrear la información de la página. Sin embargo, al momento de realizar el *web scraping* para obtener las bases de datos, nos enfrentamos a un problema debido a que la página web es dinámica y no estática. Por lo anterior, no es posible extraer las tablas de forma directa, en la medida que la información no carga inmediatamente. Las tablas realmente provienen de una promesa *json*, asociada a un link externo que es de donde pudimos extraer los datos.

(c) Para adquirir los datos seguimos el siguiente proceso:

1. Crear una lista vacía “Links” en la que se van a guardar los links de cada una de las páginas que contienen las tablas con la información
2. Crear una lista vacía “base” en la que se van a almacenar las tablas que se extraen de cada página web
3. Crear un *loop* que para i = 1, 2, …, 10, ejecuta los siguientes pasos:
   1. Guardar en la lista “Links” cada uno de los links de las 10 páginas web, definiéndolo como un documento de .*html*
   2. Descargar y preprocesar las páginas web descargadas de cada link (*read\_html)*
   3. Convertir los nodos de tipo tabla de la página web en tablas de R (*html\_table)*
   4. Almacenar cada una de las tablas en la lista “base”
4. Renombrar la primera columna de todas las bases, que está vacía, para evitar errores
5. Convertir cada una de las tablas en la lista “base” a un formato *tibble*
6. Hacer un *append* de las 10 bases almacenadas en la lista “base”, para obtener una base de datos completa con toda la información