# Conjunto de Problemas 1: Predicción de Ingresos

Fecha de entrega: Domingo, 24 de noviembre

## 1. Introducción

En el sector público, el reporte preciso de los ingresos individuales es fundamental para calcular impuestos. Sin embargo, el fraude fiscal de todo tipo siempre ha sido un problema significativo. Según el Servicio de Impuestos Internos (IRS), aproximadamente el 83,6% de los impuestos se pagan voluntariamente y a tiempo en los EE. UU. Una de las causas de esta brecha es la subdeclaración de ingresos por parte de los individuos. Un modelo de predicción de ingresos podría potencialmente ayudar a identificar casos de fraude, reduciendo esta brecha. Además, un modelo de predicción de ingresos puede ayudar a identificar a personas y familias vulnerables que puedan necesitar asistencia adicional.

El objetivo de este conjunto de problemas es aplicar los conceptos aprendidos utilizando datos del 'mundo real'. Para ello, vamos a extraer información del siguiente sitio web: https://ignaciomsarmiento.github.io/GEIH2018\_sample/. Este sitio contiene datos de Bogotá provenientes del informe de 2018 'Medición de Pobreza Monetaria y Desigualdad', basado en información de la GEIH.

Por favor, envíe un documento en formato PDF a ignaciomsarmiento@gmail.com.

### 1.1 Instrucciones Generales

El objetivo principal es construir un modelo de ingresos por hora individual:

w = f(X) + u

donde w es el ingreso por hora y X es una matriz que incluye variables explicativas/predictoras potenciales. En este conjunto de problemas, nos centraremos en f(X) = Xβ.

El documento final en formato PDF debe contener las siguientes secciones:

1. Introducción: Expone brevemente el problema y si existen antecedentes. Describe brevemente los datos y su idoneidad para abordar la pregunta del conjunto de problemas. Contiene una vista previa de los resultados y conclusiones principales.

2. Datos: Usaremos datos de Bogotá del informe de 2018 'Medición de Pobreza Monetaria y Desigualdad', basado en información de la GEIH. El conjunto de datos contiene todas las personas muestreadas en Bogotá y está disponible en este enlace: https://ignaciomsarmiento.github.io/GEIH2018\_sample/. Para obtener los datos, debe realizar un proceso de scraping.

#### Restricciones y análisis descriptivo de los datos:

• Centrarse solo en personas empleadas mayores de 18 años.  
• Limitar el análisis a estas personas y realizar un análisis descriptivo de las variables utilizadas.  
• Muchos registros tienen datos faltantes o ingresos iguales a 0. Usted deberá encontrar una forma de manejar estos datos.

Pautas para escribir esta sección:

(a) Describir brevemente los datos, incluyendo su propósito y cualquier información relevante.  
(b) Explicar el proceso de adquisición de los datos y si existen restricciones para acceder o extraerlos.  
(c) Describir el proceso de limpieza de los datos.  
(d) Realizar un análisis descriptivo, incluyendo: Una tabla de estadísticas descriptivas con interpretación. Un análisis profundo que ayude al lector a entender los datos, su variación y la justificación de las elecciones realizadas.

### 3. Predicción de ingresos:

En esta sección, evaluaremos el poder predictivo de diferentes especificaciones:

(a) Dividir la muestra en dos:  
 - Muestra de entrenamiento (70%).  
 - Muestra de prueba (30%).  
 - Utilizar el valor 123 como semilla aleatoria para reproducibilidad.  
(b) Reportar y comparar el desempeño predictivo (RMSE) de al menos 10 especificaciones, explorando diferentes niveles de complejidad y no linealidades.  
(c) Discutir los resultados, incluyendo:  
 - Desempeño general de los modelos.  
 - Especificación con el menor error de predicción.  
 - Explorar observaciones con errores de predicción significativos.  
(d) Realizar validación cruzada (LOOCV):  
 - Para los dos modelos con menor error, calcular el error predictivo utilizando LOOCV.  
 - Comparar estos resultados con los obtenidos previamente.

### 2. Pautas adicionales

• Enviar un documento en PDF a ignaciomsarmiento@gmail.com.  
• Incluir un enlace al repositorio de GitHub.

Requisitos del repositorio:  
• Seguir la plantilla.  
• Un README detallado que explique el proyecto y cómo reproducirlo.  
• Instrucciones para replicar el trabajo.  
• Al menos 5 contribuciones sustanciales por miembro del equipo.

Criterios de código:  
• Reproducible y bien documentado.  
• Tablas y figuras presentadas de manera ordenada y con etiquetas.