*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

**Big Data y Machine Learning para Economía Aplicada**

**Taller 1**

**2023-02**

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

Angie Ariza Quitian– Código

**Presentado por**

Andrés Felipe Diaz Barreto - 200610686

Duvan Javier Farfán López – 201317299

Daniel Eduardo Sandoval - 200712968

**Introducción - Dani**

El salario de una persona depende de diferentes variables y contextos particulares. En Colombia, el salario (…)

**Descripción de las fuentes de datos – Dani**

**Gran Encuesta Integrada de Hogares**

La Gran Encuesta Integrada de Hogares, se viene desarrollando desde el año 2005 y es realizada por el Departamento Administrativo Nacional de Estadística (DANE). Esta encuesta, analiza los hogares como unidad de análisis para obtener información sobre cómo están formados estos, además de los aspectos socioeconómicos de cada uno; de esta forma la información recolectada permite describir estas características que evidencia la realidad de los hogares colombianos.

De esta manera, la GEIH ayuda a generalizar la situación sobre el empleo, salarios, formalidad e informalidad, así como actividades económicas adicionales que puedan realizar las personas.

**Adquisición de datos**

El sitio web indicado en el enunciado (https://ignaciomsarmiento.github.io/GEIH2018 sample/) contiene la siguiente información:

1. Acceso a las diez partes (*data chunks*) en las que se dividió la muestra del GEIH 2018

2. Enlace al GEIH del 2018

3. Diccionario de las variables de la muestra

4. Descripción del DANE de la metodología y de las variables utilizadas en el GEIH 2018.

Durante la exploración de las páginas en las que estaban los *data chunks*, se evidenció que las tablas con las observaciones se demoraban en cargar. Al intentar raspar su información en R, mediante las funciones del paquete *rvest*, se obtuvo que no había ninguna tabla; esto sucedió porque al momento de crear el objeto *hmtl* para el respectivo *data chunk* la tabla todavía no había sido cargada, por tanto, el objeto carecía de la información que se quería obtener.

Luego de identificar que la razón por la que había una demora en el cargue de las tablas era porque estas páginas estaban trayendo la información de otros sitios web, se cambiaron las URL que estaban utilizándose para raspar las observaciones, logrando conseguir las 32,177 observaciones de la muestra mencionada.

**Descripción del proceso de limpieza de datos – Javier**

Una vez se realizó a la adquisición de la data y se exportó como archivo .Rda, se creó un nuevo script para realizar la limpieza de la misma. Una vez cargada, se identificó que la base de datos estaba conformada por 24.054 observaciones y 178 variables.

En la primera revisión de la estructura de lavase de datos, se evidenció que algunas de las variables poseen campos con NAN, lo cual distorsionaría el resultado de las estimaciones a realizar, sin embargo, previo a verificar el contenido de cada variable, se procedió a verificar en la literatura cuales son los criterios identificados en diferentes estudios para explicar el salario de una persona.

Incluir teoría asociada: Mincer, otros estudios, estudios colombianos

Teniendo en cuanta la literatura, y con el animo de identificar las características sociodemográficas de la muestra, se realizó la **primera selección de variables** que conformarían la base de datos para el desarrollo de estimación de parámetros y predicción de salario y brechas salariales por género. Las variables seleccionadas, su descripción y justificación de uso basada en la teoría económica y consideraciones propias se puede observar en la tabla 1

Tabla 1 Primera selección de variables

| Variable | Nombre de la variable | Descripción de la variable | Justificación de uso |
| --- | --- | --- | --- |
| age | Edad |  |  |
| clase |  |  |  |
| college |  |  |  |
| cuentapropia |  |  |  |
| dsi |  |  |  |
| estrato1 |  |  |  |
| hoursWorkUsual |  |  |  |
| informal |  |  |  |
| ingtotob |  |  |  |
| maxEducLevel |  |  |  |
| microEmpresa |  |  |  |
| ocu |  |  |  |
| oficio |  |  |  |
| p6210 |  |  |  |
| p6620s1 |  |  |  |
| relab |  |  |  |
| sex |  |  |  |
| sizeFirm |  |  |  |
| y\_bonificaciones\_m |  |  |  |
| y\_salarySec\_m |  |  |  |
| y\_ingLab\_m\_ha |  |  |  |
| y\_total\_m |  |  |  |
| y\_total\_m\_ha |  |  |  |

**Análisis descriptivo de los datos - Javier**

**Estimación de perfil edad – salarios - Dani**

**Presentación perfiles edad-salarios y “edad-pico” - Dani**

**Estimación de la brecha salarial de género - Angie**

**Perfiles edad-salarios y “edad-pico” por género - Angie**

**Interpretación de las estimaciones de la brecha salarial de género - Angie**

**Construcción de Muestra para predicción - Andy**

**Desempeño predictivo**

La evaluación del desempeño predictivo de los modelos anteriormente especificados se realizó mediante la utilización de dos métodos de validación cruzada: el enfoque de conjunto de validación (*validation set approach*) y el LOOCV (por sus siglas en inglés).

Para el primer método, se dividió aleatoriamente la muestra en dos partes: conjunto de entrenamiento (*training set*) y conjunto de evaluación o de validación (*validation set*). La primera contiene el 70% de las observaciones y la última, el porcentaje restante. Con estos conjuntos, se entrenaron y evaluaron los siguientes diez (10) modelos:

Tabla 2. Especificaciones evaluadas en el método de validación cruzada: conjunto de validación

|  |  |
| --- | --- |
| Modelo | Especificación |
| Modelo 1 |  |
| Modelo 2 |  |
| Modelo 3 |  |
| Modelo 4 |  |
| Modelo 5 |  |
|  |  |
|  |  |
| Modelo 8 |  |
| Modelo 9 |  |
| Modelo 10 |  |

Como parámetro de evaluación, se utiliza el error cuadrático medio (RMSE, por sus siglas en ingles) de la prueba, que es estimado a partir de los valores predichos por el modelo ajustado en el conjunto de entrenamiento para las observaciones del conjunto de validación. En la siguiente tabla, se muestran los resultados, organizados de menor a mayor RMSE para identificar fácilmente a los que tienen mejor desempeño:

Tabla 2. RMSE de prueba de las especificaciones sujetas a evaluación

|  |  |
| --- | --- |
| Modelo | RMSE |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

*El desempeño general de los modelos…*

*El desempeño con el menor error de predicción*

*Revisar la distribución de los errores en el modelo con mejor desempeño*

**Interpretación del desempeño predictivo - Andy**

**LOOCV - Javier**

**Variables a utilizar:**

* Age
* Clase
* College
* cuentaPropia
* dsi
* estrato1
* hoursWorkUsual
* ie
* ina
* inac
* informal
* impa
* ingtot
* maxEducLevel
* microEmpresa
* ocu
* oficio
* p6210
* p6210s1
* p6426
* relab
* sex
* sizeFirm
* totalHoursWor~d
* y\_bonificacio~m
* y\_ingLab\_m\_ha
* y\_salary\_m\_hu
* y\_total\_m\_ha

**Justificación de variables:**

* **Estudio Michael page: remuneración en Colombia 2022:** Se evidencian diferencias en salario por tipo de industria, cargo profesional y tamaño de la empresa
* **Estudio Banco de la Republica de Colombia:** <https://publicaciones.banrepcultural.org/index.php/emisor/article/view/7839/8219> se muestra que los salarios varían por sector, tamaño de la empresa. Tambien que depende de la experiencia y el cargo

**Páginas Usadas:**

**Justificación de variables elegidas:**

* **Estudio\_Remuneraci\_\_n\_Michael\_Page\_2022**
* **https://publicaciones.banrepcultural.org/index.php/emisor/article/view/7839/8219**