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

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

**Taller 1**

**2023-02**

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

Angie Ariza Quitian– Código

**Presentado por**

Andrés Felipe Diaz Barreto - Código

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 | Edad de la persona – Solo se tienen en cuenta mayores de 18 años | La teoría económica argumenta |
| clase | Sector de residencia | Indica si la persona vive en el sector rural o urbano | Según estudios de la CEPAL y el DANE, se evidencian diferencias entre los salarios que se perciben a nivel rural y urbano, siendo los primeros más bajos. Lo anterior asociado al nivel de experticia requerida en los trabajos relacionados con el campo, que principalmente están enfocados agrícola y agropecuaria. La brecha salarial ha disminuido durante los últimos años en Colombia, principalmente por la caída de ingresos de los sectores urbanos y la mejora en el acceso a educación de la población rural, sin embargo, hay evidencia sobre la brecha que se presenta entre los dos sectores. |
| college | Educación terciaria | Indica si la persona tiene o no estudios de educación terciaria (educación universitaria y/o formación profesional) | En el modelo de Mincer (1958) se indica que el logaritmo del ingreso depende linealmente de la escolaridad. Luego, en 1974, Mincer propone un modelo similar incluyendo la experiencia, que genera mucha más confianza dados los fuertes supuestos asociados únicamente a la educación de una persona. |
| cuentaPropia | Independiente | Indica si la persona trabaja de forma independiente o recibe salario |  |
| dsi | Desempleado | Indica si la persona se encuentra desempleada o no |  |
| estrato1 | Estrato | Estrato de energía para las 13 a.M., y sextil de icv para otras cabeceras |  |
| hoursWorkUsual | Horas trabajadas por semana | Horas usuales trabajadas por la persona a la semana |  |
| informal | Informal | Define si la persona es informal  (no cuenta con seguridad social) |  |
| ingtotob | Ingreso total observado | Ingreso total observado |  |
| maxEducLevel | Nivel educativo (incluye estado) | Indica el máximo nivel alcanzado de escolaridad y si la persona finalizó los estudios asociados al nivel |  |
| microEmpresa | Microempresa | Indica si la persona trabaja en una empresa de 5 o menos empleados o en una con más de 5 empleados |  |
| ocu | Ocupado | Indica si la persona está ocupada o activa laboralmente |  |
| oficio | Ocupación | Indica la ocupación o profesión de la persona |  |
| p6210 | Nivel educativo (no incluye estado) | Indica el máximo nivel alcanzado de escolaridad. No incluye estado de finalización |  |
| p6620s1 | Estimado de ingreso | Estimación de ingresos recibidos por la persona |  |
| relab | Tipo de ocupación | Tipo de ocupación y categoría de empleabilidad (cuenta propia, empleado, otros) |  |
| sex | Sexo | Sexo: Hombre o mujer |  |
| sizeFirm | Tamaño de la empresa | Tamaño de la empresa por número de empleados |  |
| y\_bonificaciones\_m | Ingreso monetario al mes | Ingreso monetario en el mes |  |
| y\_salarySec\_m | Salario nominal mensual (secundario) | Salario nominal mensual occ. secundario |  |
| y\_ingLab\_m\_ha | Ingreso por salario (Asalariados) | Ingresos laborales de asalariados – nominales por hora (Incluye propinas y comisiones) |  |
| y\_total\_m | Ingresos asalariados por mes | Ingresos asalariados + independientes total - nominal mensual |  |
| y\_total\_m\_ha | Ingresos asalariados por hora | Ingresos asalariados + independientes total - nominal por hora |  |

**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**

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[[1]](#footnote-1)[[2]](#footnote-2) |
| --- | --- |
| Modelo 1 |  |
| Modelo 2 |  |
| Modelo 3 |  |
| Modelo 4 |  |
| Modelo 5 |  |
| Modelo 6 |  |
| Modelo 7 |  |
| Modelo 8 |  |
| Modelo 9 |  |
| Modelo 10 |  |

**Desempeño predictivo**

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 gráfica, se muestran los resultados por modelo:

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figura 1. RMSE de prueba de las especificaciones sujetas a evaluación

Como se puede ver en la gráfica anterior, los RMSE de los modelos tienen una magnitud significativa (entre 6.8% y 8.8%) respecto a la media del logaritmo del salario en el conjunto de prueba, que es 8.565. Cuando se incluyen los controles especificados (*informal, maxEducLevel, sizeFirm, relab*), el RMSE se reduce en aproximadamente 20%, variando este del modelo, y se mantiene relativamente estable con diferentes grados del polinomio de edad.

**Interpretación del desempeño predictivo**

Respecto al modelo con menor error de predicción, en el que se incluye un polinomio de grado 3 para la edad y los diferentes controles (*female, informal, maxEducLevel, sizeFirm, relab*), se tiene la siguiente distribución del error:

Gráfico, Histograma

Descripción generada automáticamente

Figura 2. Distribución del error de predicción en el training set – Validation Set Approach.

De acuerdo con la figura anterior, hay varios datos atípicos, especialmente en la cola de la izquierda de la distribución. Con el fin de identificar si esta distribución obedece a datos atípicos del salario por hora registrados en la base de datos o al desempeño del modelo, se presenta la siguiente figura:

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Figura 3. Diagrama de dispersión entre el error de predicción y el logaritmo del salario por hora observado en el training set – Validation Set Approach.

Los puntos con color rojo son aquellos que corresponden a datos atípicos en la distribución del logaritmo del salario por hora observado, mientras que los verdes son aquellos que están entre las líneas punteadas A y B, que delimitan el rango comprendido por tres desviaciones estándar de la media. De 4286 observaciones en el *training set*, 28 se identifican como datos atípicos (0.65% de la muestra); en términos de la participación en la suma de los residuos al cuadrado (RSS, por sus siglas en inglés), estos datos atípicos representan el 10.3%, por lo que se puede ver la influencia que tienen en el error de la predicción. Sin embargo, casi el 90% del RSS corresponde a datos del logaritmo del salario que no se consideran atípicos.

Por lo anterior, este modelo requiere un mejor ajuste para identificar si hay casos en los que haya un riesgo de fraude, dado que gran parte del error se concentra en observaciones que no son atípicas.

**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**

1. En algunos modelos se utilizan un conjunto de controles identificado con la variable , que contiene las siguientes variables: *informal,* *maxEducLevel, sizeFirm,* *relab* [↑](#footnote-ref-1)
2. Con el conjunto de variables de control se hicieron varias combinaciones, quitando una a la vez, y la mejor configuración es aquella con todos los controles. [↑](#footnote-ref-2)