**BIG DATA MACHINE LEARNING FOR APPLIED ECONOMICS**

**TALLER 1 – PREDICTING INCOME**

INTEGRANTES:

Luisa Cuellar - 201613942

Isabella Riveros - 201923015

Miguel Contreras - 202116189

CONTENIDO

Introducción

Datos

Perfil edad-ingresos

Brecha salarial de género

Predicción de ganancias

1. Introducción

En el presente documento se indaga acerca de las características que determinan el ingreso de un trabajador en Colombia. Para ello, se realizan una serie de análisis econométricos utilizando datos de la Gran Encuesta Integrada de Hogares (GEIH) del 2018. Se efectuaron regresiones con variables que determinan las características de las personas como, por ejemplo, la edad, el sexo, el estrato socio económico y el nivel educativo, entre otros.

Algunos de los resultados destacables de este estudio son que, la edad es un factor relevante en los salarios, teniendo un signo positivo en el estimador, lo cual indica que al momento de ingresar al mercado laboral (para el caso de este estudio, 18 años) la remuneración es más baja, y tiende a incrementarse con el paso de los años. Por su parte, el coeficiente de la edad al cuadrado cuenta con un signo negativo, lo que indica que el salario de las personas de la muestra aumenta, pero a un ritmo cada vez más bajo, hasta el punto en que tiende a estabilizarse o incluso, a declinar (a excepción de algunos datos atípicos de individuos de estratos altos). Otro de los hallazgos es que, las mujeres presentan en general, menores salarios que los hombres (xx % según estima el modelo), aunque la brecha se reduce al momento de comparar individuos con el mismo oficio. También se puede observar que, tal como lo predice la teoría, la inclusión de más variables en las regresiones tiende a aumentar la varianza de los modelos, mientras que el R2 no es un instrumento muy confiable para decidir si incluir o no un nuevo regresor en la regresión, dado que normalmente aumenta a medida que se incorporar más variables explicativas.

1. Datos (tabla de estadísticas)

Tras el proceso de depuración de la base de datos se obtiene que, la muestra seleccionada para el estudio corresponde a hombres y mujeres mayores de 18 años que se encuentran en el mercado laboral colombiano. La muestra se restringió a xxxxxx personas. En cuanto a las variables de las características personales inicialmente se plantearon las de la edad, el sexo, el nivel educativo (con 1 missing value), el estrato socioeconómico, el régimen de salud al cual se encuentra afiliado (con 1.420) y si es cotizante en pensiones.

Por otro lado, en cuanto a las características laborales se plantearon las variables tamaño de la firma, si es una microempresa, el oficio, las horas trabajadas usualmente, las horas trabajadas en el segundo trabajo actual, si es trabajador informal y la relación laboral. Como medida de ingreso las variables que se contemplaron fueron ingreso total mensual y la de cuándo ganó el mes anterior por la actividad principal (con 4.535 missing values). Tras un análisis conjunto se decidió utilizar las variables que se describen a continuación:

Ojo. Decir cuáles son las categóricas, hablar de la distribución de las variables.

1. Perfil edad-ingresos

Modelo 1: ing-vs-age-age2

1. Brecha salarial de género

Modelo 2: loging-vs-sex.

Modelo 3: loging-vs-sex-age-age2)

Modelo 3: loging-vs-sex-age-age2-ctrlvar

Modelo 4: loging-vs-sex-age-age2-ctrlvar FWL

1. Predicción de ganancias

Con el fin de evaluar el poder de predicción de la especificación, se dividió la muestra en dos: de entrenamiento (70%) y de prueba (30%)

……………………………..

Construcción de conjunto de datos y proceso de adquisición

Descripción de los datos, variables utilizadas, tabla y por qué usamos esas

Regresión ingreso vs edad. Interpretar el coeficiente, el ajuste

Gráfica perfil ingresos por edad

Discusión de la "edad máxima" con sus respectivos intervalos de confianza. (use bootstrap para construir los intervalos de confianza).

Estimación brecha salarial condicional con otras variables (características laborales )

Ver distribuciones, si hay outliers

Es cóncava o convexa

Calcular el error estándar con el Bootstrap (resampleo, cómo varía con cada)