Problem Set 1: Predicting Income

Big Data

MEcA

Taller_1_BigData

Grupo 2

Alejandro García 201810340

Andrés Torres 202115992

Ignacio Serrano 201618789

https://github.com/AndresT2022/Taller_1_BigData

Introducción

Este trabajo toma la Gran Encuesta de Ingresos y Hogares - GEIH de 2018, realizada por el DANE. Con este taller se busca dar explicación a partir de la teoría económica y utilizando análisis básico y herramientas de machine learning para poder captar el concepto de no buscar el mejor ajuste sino la capacidad del modelo de replicarse y predecir en otros conjuntos muestrales con el menor error posible. El enfoque de este taller es aplicar la idea anterior, de permitir una mayor variabilidad que tener un modelo que se ajuste y tenga una alta significancia estadística. Para aplicar lo anteriormente descrito, se busca analizar varios modelos que expliquen la variable dependiente ingreso en función de las variables disponibles en la GEIH, dicho análisis se realizará por medio de los conceptos vistos en clase. Se decidió utilizar el lenguaje de programación R y el interfaz de ejecución R-Studio, dado que R tiene paquetes que permiten una mejor ejecución de lo que se quería probar con la muestra propuesta.

Desarrollo

1. Obtención de datos

Se utilizó la metodología de webscrapping con el paquete rvest que nos permitió navegar dentro de la página creada para este taller donde se encontraban anidados los datos. Dicha página no tenía ningún tipo de restricción para el uso del scrapping como pudimos comprobar al validar que no contaba con el documento robots.txt.

El proceso de lo anterior consistió en encontrar el vínculo html que contenía las tablas utilizando las herramientas de desarrollo de inspección de la página en el navegador Chrome, una vez identificado esto, se decidió crear un ciclo que permitiera hacer la descarga de cada uno de los 10 chunks, y se unificaron en un dataframe que se exportó a un archivo Excel para facilidad de acceso a los datos. La muestra unificada es de 32,177 registros y 178 columnas.

2. Limpieza de datos

Se delimito la base de datos a los individuos mayores de 18 años y menores o iguales a 80 años, ocupados que registraron su residencia en Bogotá.

Consideramos la variable de ocupados como la más efectiva para indicar los individuos que están trabajando y con ingreso, ya que esta variable considera que la persona sea de la PEA (Población Económicamente Activa) y tenga cualquier tipo de relación salarial, sea formal o informal, y que le permita tener un ingreso mensual. Se decidió acotar la muestra a 80 años, dado que hay adultos mayores sin pensión que deben seguir ocupados y la esperanza de vida en Colombia oscila entre 77 – 82 años. Adicionalmente, se consideró esta variable dado que era la que presentaba menos missing values, lo nos permite tener una mayor cantidad de observaciones.

Decidimos dejar como variables relevantes:

- Edad
- Género
- Oficio
- Máximo nivel de educación

Respecto a la variable de máximo nivel de educación, en base a ella se generó una nueva variable que contiene la "experiencia potencial" que tiene cada observación, lo anterior, dado que revisando la literatura de Mincer nos indica que la experiencia es una variable importante al momento de explicar el ingreso de un individuo. Dicha variable se creó asociando el máximo nivel educativo de los individuos con años de estudio acumulado de la siguiente manera:

maxEducLevel	Descripción categoría	Años de estudio acum
1	Ninguna	0
2	Prescolar	2
3	Primaria incompleta	6
4	Primaria completa	7
5	Secundaria incompleta	12
6	Secundaria completa	13
7	Terciaria	19
9	N/A	N/A

Posteriormente, para estimar los años de experiencia de cada individuo se utilizó la siguiente ecuación:

$$Exp = age - 5 - años\_estudio\_acum$$

La ecuación anterior resta 5 años dado que en los 5 primeros años un individuo en promedio no estudia. Dada la estimación realizada, algunos datos tenían experiencia negativa, lo cual no tiene sentido, por lo anterior, estos datos se reasignaron a cero años de experiencia.

Estadísticas descriptivas.

La base de datos final cuenta con 13,504 observaciones y 31 variables. Entre las variables se encuentran datos sociodemográficos que nos han servido para segmentar la muestra de la población en personas mayores de 18 años, ocupadas y habitantes de la ciudad de Bogotá. Se incluyen variables que permiten caracterizar la muestra dentro de sus actividades laborales y formación educativa, así, tenemos una muestra dividida homogéneamente entre sexo femenino y masculino, donde el 70% reporta tener un trabajo formal y el 30% tener formación superior. Se reporta 1,684,365cop de salario promedio y trabajan 47.3 horas a la semana, en promedio. Para entender la muestra debemos tener en cuenta la definición de los encuestadores (DANE) sobre las personas ocupadas, pues se considera que una persona tiene empleo si reporta 1 hora de actividad remunerada en la semana. Se debe considerar igualmente que la mayoría de la muestra se encuentra en estratos socioeconómicos bajos. Se observa una dispersión muy alta entre los datos individuales de la variable de ingreso mensual, esto en consecuencia de salarios reportados considerablemente grandes comparados con los ingresos promedio.

Summary Statistics

Statistic	Mean	St. Dev.	Min	Max
directorio	4,657,702.0	81,876.8	4,514,331	4,804,455
secuencia_p	1.0	0.1	1	4
orden	1.9	1.2	1	12
clase	1.0	0.0	1	1
mes	6.5	3.4	1	12
estrato1	2.5	1.0	1	6
sex	0.5	0.5	0	1
age	39.2	13.1	18	80
p6210	5.0	1.1	1	6
maxEducLevel	6.0	1.2	1	7
regSalud	1.3	0.7	1	3
cotPension	1.4	0.5	1	3
sizeFirm	3.3	1.7	1	5
oficio	49.0	28.0	1	99
wap	1.0	0.0	1	1
ocu	1.0	0.0	1	1
dsi	0.0	0.0	0	0
pea	1.0	0.0	1	1
inac	0.0	0.0	0	0
totalHoursWorked	47.3	14.6	1	130
formal	0.7	0.5	0	1
informal	0.3	0.5	0	1
cuentaPropia	0.3	0.4	0	1
microEmpresa	0.4	0.5	0	1
college	0.3	0.5	0	1
Escol	14.5	4.4	0	19
y_total_m	1,684,365.0	2,494,130.0	84.0	70,000,000.0
y_total_m_ha	8,872.8	14,032.4	2.5	350,583.3
LnIng	13.9	0.9	4.4	18.1
agesqr	1,708.2	1,120.8	324	6,400
exp	19.8	14.9	0	74

Tabla 1. Estadística descriptiva base de datos muestra limpia.

3. Estimación de modelo y perfiles

Dado el enunciado, se propone inicialmente el siguiente modelo:

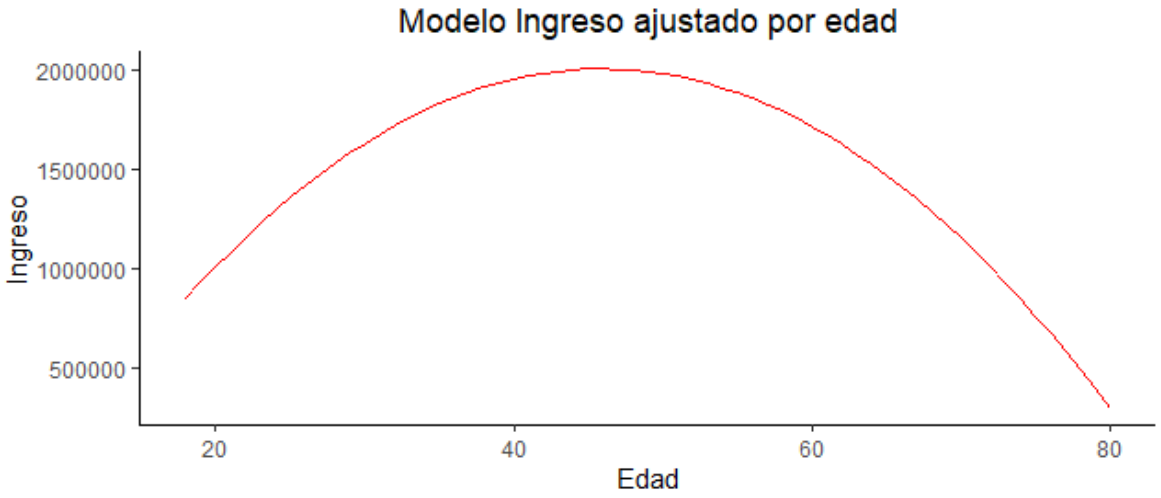
$$Income = \beta_1 + \beta_2 Age + \beta_3 Age^2 + \mu$$

Dicho modelo se estimó por OLS, en donde se encontró que el ajuste del modelo es muy bajo, a pesar de que es estadísticamente significativo como se muestra en la siguiente tabla:

	Dependent variable:	
	y_total_m	
	(1)	(2)
age	11,568.940***	135,725.900***
	(1,629.500)	(9,866.201)
agesqr		-1,476.482***
		(115.737)
Constant	1,231,061.000***	-1,111,709.000***
	(67,346.940)	(195,465.600)
Observations	13,504	13,504
R2	0.004	0.016
Adjusted R2	0.004	0.015
Residual Std. Error	2,489,580.000 (df = 13502)	2,474,801.000 (df = 13501)
F Statistic	50.406*** (df = 1; 13502)	106.877*** (df = 2; 13501)
Note:		*p<0.1; **p<0.05; ***p<0.01

Tabla 2. Regresión Ingreso Vs Edad.

En la siguiente gráfica se muestra que los valores predichos precios a partir de la ecuación propuesta, sugieren una edad de máximo ingreso de aproximadamente 46 años (45.96 años) con un intervalo de confianza del 95% entre 44.19 y 47.68 años.



Gráfica 1. Modelo ingreso individuos ajustado por edad.

4. Existencias de brecha de género

Usando el siguiente modelo sugerido:

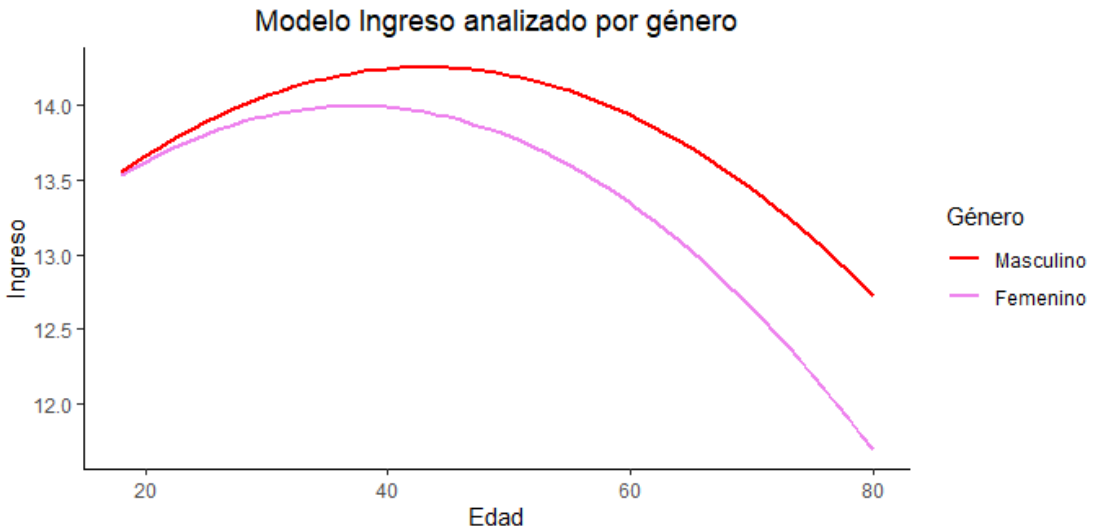
$$\log(Income) = \beta_1 + \beta_2 Female + \mu$$

Se encontró que las mujeres sin clasificar por oficio tienen una brecha respecto a los ingresos de los hombres, es decir, mientras mayor edad tengan las y más ingresos, hay una brecha que se mantiene, como se muestra a continuación:

	Dependent variable:		
	LnIng	y_total_m	
	(1)	(2)	(3)
age	0.094***	147,964.700***	136,279.800***
	(0.005)	(13,973.610)	(13,869.510)
agesqr	-0.001***	-1,767.104***	-1,366.418***
	(0.0001)	(165.411)	(161.508)
female			
male			
Constant	12.251***	-1,246,490.000***	-1,185,959.000***
	(0.108)	(275,422.900)	(275,769.400)
Observations	6,544	6,544	6,960
R2	0.075	0.017	0.022
Adjusted R2	0.075	0.017	0.021
Residual Std. Error	0.930 (df = 6541)	2,370,229.000 (df = 6541)	2,553,725.000 (df = 6957)
F Statistic	264.423*** (df = 2; 6541)	57.111*** (df = 2; 6541)	76.551*** (df = 2; 6957)
Note:			*p<0.1; **p<0.05; ***p<0.01

Tabla 3. Regresión Ingreso Vs Género.

De los resultados anteriores se puede concluir que, si el individuo se asume como femenino, la brecha de ingresos se incrementa conforme llega la edad cenit laboral, y luego aumenta. Gráficamente se puede observar este resultado.



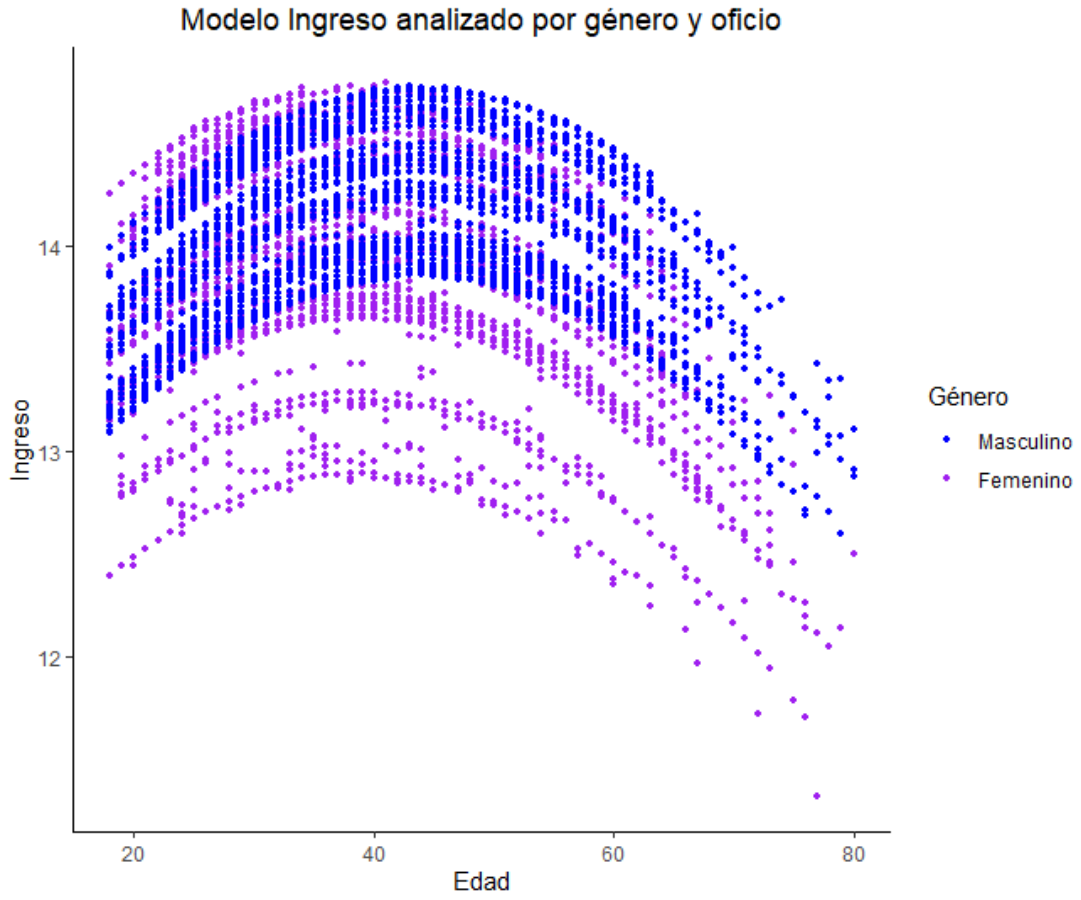
Gráfica 2. Modelo ingreso individuos ajustado por género.

De los resultados mostrados anteriormente, puede observarse que las mujeres en promedio tienen un pico de ingresos a los aprox. 42 años (41.87 años) con un intervalo de confianza del 95% entre 40.84 y 42.96 años; Y en los hombres, se observa que las mujeres en promedio tienen un pico de ingresos a los aprox. 50 años (49.87 años) con un intervalo de confianza del 95% entre 44.74 y 54.29 años.

De lo anterior se puede concluir que, a nivel general según la muestra, las mujeres llegan a su pico de ingreso más rápido que los hombres y esto genera la brecha, sin embargo, puede ser que esto se relacione con la escolaridad y el oficio por el que se perciben los ingresos.

Género y oficios

1. Condicionando por oficios, lo que se ve es que la brecha, aunque sigue existiendo, se reduce especialmente para oficios similares, ya que la variable oficio ajusta las diferencias en el crecimiento de los salarios. Lo que muestra es que en general tanto hombres como mujeres se sitúan en niveles cercanos durante los primeros años de vida laboral, pero la brecha aumenta conforme se avanza la edad y el cenit de ingresos.



Gráfica 3. Modelo ingreso individuos ajustado por género y oficio.

2. Si estimamos por medio del teorema FWL, lo que nos muestra es que efectivamente el estimador de la variable oficios es equivalente, como se muestra a continuación

	LnIng
Intercept	13.094
age	0.088
agesqr	-0.001
oficio	-0.020

Tabla 4. Estimación por OLS.

Coeffcients	**Estimate**	**STd. Error**	**PR(>|t|)**
(Intercept)	1.309e+0.1	9.914e-02	<2e-16***
age	8.825e-02	4.916e-0.3	<2e-16***
agesqr	-1.131e-03	5.826e-05	<2e-16***
oficio	-1.983e-02	4.961e-04	<2e-16***

signif. codes: 0 ***

Tabla 5. Estimación Teorema FWL.

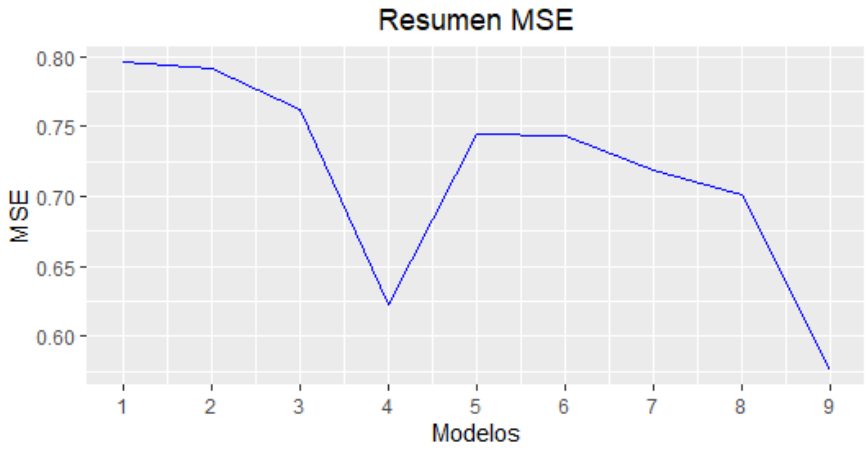
Ajustando con oficio la conclusión es que efectivamente, analizando la variable "Female", el estimador es equivalente, y por tanto ayuda a explicar la dinámica de la brecha.

3. Lo que puede concluirse es que a nuestra muestra el modelo se ajusta, mostrando la existencia de una brecha en la población ocupada con ingresos mensuales, pero que esta misma brecha depende más del tipo de trabajo para que sea mayor o menor, en especial en el inicio de la vida laboral. Por tanto, la brecha obedece más a un problema de selección que a una discriminación directa, y a condicionantes como la edad y el oficio.

Predicciones

1. 70% muestra entrenamiento – 30% muestra prueba

Se realizaron 9 modelos combinando las variables que según la literatura y nuestro conocimiento son relevantes al momento de explicar el ingreso de los individuos. El primer modelo tiene únicamente una constante.



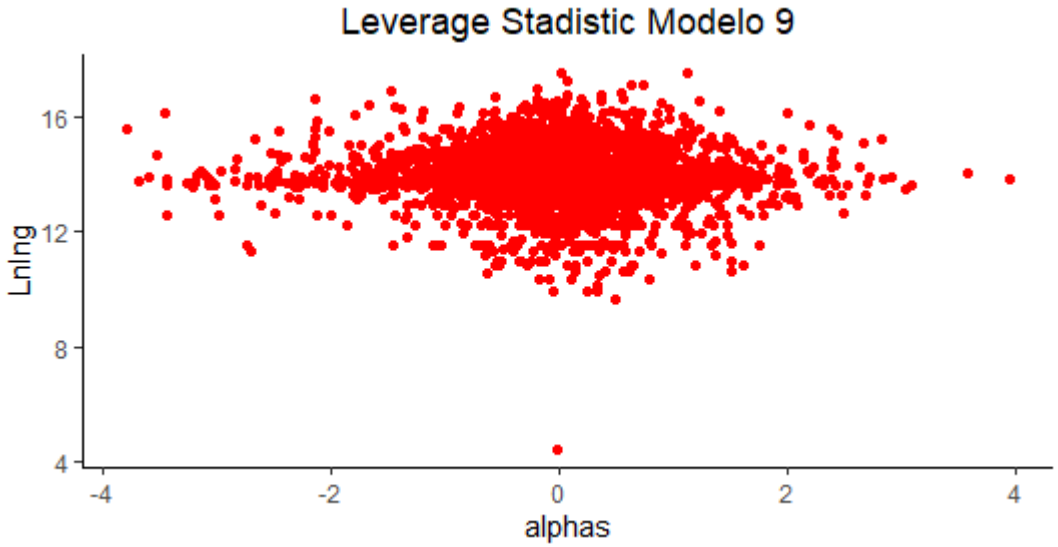
Gráfica 4. Resumen de los MSE predicción con 70 – 30.

Al analizar los MSE de los modelos propuestos, se encuentra que el mejor modelo es el 9, como se muestra en la anterior gráfica. El modelo 9 empleado es el siguiente:

$$LnIng = \beta_1 + \beta_2 exp + \beta_3 \exp^2 + \beta_4 sex + \beta_5 age + \mu$$

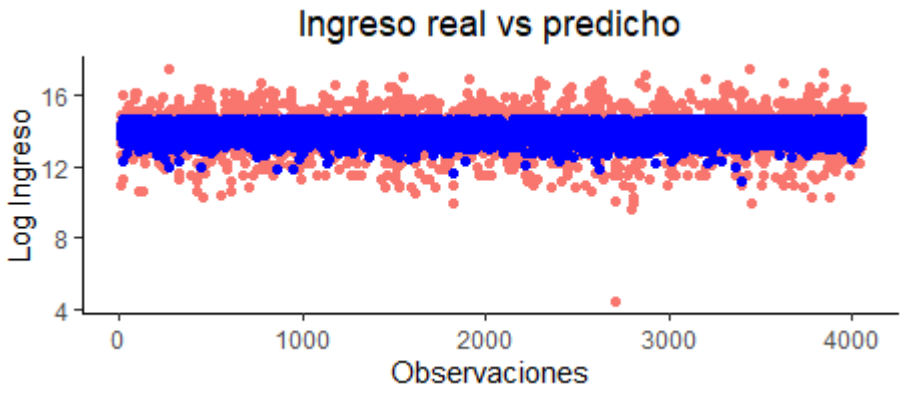
De acuerdo con la teoría económica y el modelo de ingresos propuesto por Mincer, encontramos que el modelo con mayor ajuste contempla las variables de edad, años de experiencia y sexo.

Usando leverage statistic en el modelo 9, se encuentra que existen datos atípicos, como se muestra en la siguiente gráfica.



Gráfica 5. Leverage stadistic modelo 9. Ingreso vs Alpha.

Al realizar una comparación entre los valores estimados o predichos por el modelo y los valores reales de los ingresos en la muestra de prueba se encuentra que se tiene evidencia para sospechar que hay individuos que están mintiendo en el reporte de sus ingresos, como se muestra a continuación:



Gráfica 6. Ingresos reales vs predichos. En rojo ingresos reales, en azul valores predichos por el modelo 9.

En este orden ideas, nuestro modelo 9 puede llegar a ser un buen elemento para ponerle atención a esas observaciones que al parecer reportan mal sus ingresos. Aunque en general la mayoría de los individuos, acorde a estos datos, reportan bien sus ingresos.

2. K-fold cross-validation.

Realizando el mismo ejercicio del punto anterior bajo la metodología de K-fold cross validation, se tiene el siguiente cuadro resumen con los MSE:

Modelo	MSE
1	0.795
2	0.890
3	0.876
4	0.789
5	0.772
6	0.866
7	0.853
8	0.844
9	**0.762**

Tabla 5. Mapa calor de MSE predicción CV.

Como se puede observar de la tabla anterior, el modelo 9 continúa siendo el mejor modelo.

Es importante mencionar que la metodología CV de esta sección es en esencia misma la metodología de la sección anterior, solo que acá se cuentan con 5 grupos de entrenamiento y prueba, con lo cual, se puede decir que es mejor CV que la metodología anterior.

3. LOOCV.

Realizando el mismo ejercicio del punto anterior bajo la metodología de LOOCV, se encuentra que para el modelo propuesto 9, el promedio del MSE es de 0.581. Dicho valor es inferior al de las dos metodologías anteriores, lo cual es deseable al momento de predecir, tener un menor MSE.

Es importante mencionar que esta metodología, aunque disminuye en alguna medida las desventajas de las metodologías anteriores, es computacionalmente intensivo.