Trabajo especialización:
Análisis de los años de estudio de individuos Americanos

Maestría en la Explotación de Datos y Descubrimiento del Conocimiento

Victoria Busto

(DNI.: 32.618.034)

Contents

Índice de Tablas

Índice de Ilustraciones

Introducción

Lograr entender los factores que determinan el nivel educativo de los individuos es de gran importancia para los gobernantes, las agencias no gubernamentales y en especial para la sociedad. De esto radica la importancia de predecir de forma acertada la cantidad de años de educación de un individuo a partir de datos y características socio-económicas y demográficas, dado que pueden ser utilizadas para realizar políticas públicas focalizadas.

Este trabajo utiliza 14 variables socio-económicos de 1980 de una base de 4.739 individuos de nacionalidad EE.UU. para determinar los años de estudio.

El objetivo principal es determinar el modelo que prediga con mayor certeza la variable dependiente (cantidad de años de estudio), es decir, aquel que otorgue una menor perdida de predicción. La función de pérdida utilizada es la del error cuadrático medio. Para ello, se construyen diversos modelos de regresión y se evalúa la performance de cada uno de ellos.

Material y métodos desarrollo

Descripción de la base

El dataset utilizado proviene de una encuesta de corte transversal a Colegios secundarios y Universidades realizada por el Departamento de Educación de Estados Unidos en 1980. El dataset contiene 14 variables socio-económicas y demográficas de 4.739 jóvenes adultos Americanos, las cuales pueden ser clasificadas en:

- 1 variable dependiente,
- 1 variable categórica,
- 8 variables dicotómicas o dummies, y
- 4 variables continuas.

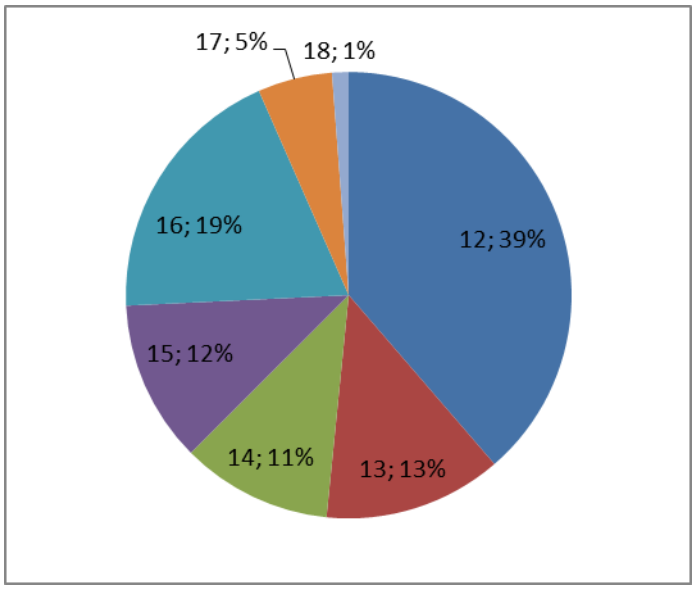
Las variables contenidas en el dataset son las que se describen a continuación:

- ***education:*** *Variable discreta. Variable dependiente. Indica la cantidad de años de educación del individuo.*
- ***gender:*** *Variable dicotómica. Indica el género de la persona.*
- ***ethnicity:*** *Variable categórica. Indica la raza del individuo. Esta variable puede tomar los siguientes valores: afro-americano (afam), hispano (hispanic) u otros (other).*
- ***score:*** *Variable continua. Indica la calificación de la prueba compuesta del año base. Estas son las pruebas de rendimiento otorgadas a los alumnos de secundaria de la muestra.*
- ***fcollege:*** *Variable dicotómica. Indica si el padre del individuo es graduado de la universidad o no.*
- ***mcollege:*** *Variable dicotómica. Indica si la madre del individuo es graduada de la universidad o no.*
- ***home:*** *Variable dicotómica. Indica si la familia del individuo es dueña del hogar en el que viven.*
- ***urban:*** *Variable dicotómica. Indica si la universidad se encuentra en un área urbana.*
- ***unemp:*** *Variable dicotómica. Indica la tasa de desempleo del condado en 1980.*

- ***wage:*** *Variable continua. Indica el salario promedio por hora para la industria para el estado del individuo en 1980.*
- ***distance:*** *Variable continua. Indica la distancia en 10 millas de la universidad más cercana con plan de 4 años.*
- ***tuition:*** *Variable continua. Indica el costo promedio del universitario (matricula y cuota) por mes por un plan de 4 años (miles de USD).*
- ***income:*** *Variable dicotómica. Indica si el ingreso familiar se encuentra por encima de 25.000 USD por año.*
- ***region:*** *Variable dicotómica. Indica si la región pertenece al Oeste (West) o no.*

La variable dependiente son los años de educación del individuo, variable con un rango de 12 a 18 años. Un 38,7% de los individuos solo tiene 12 años de estudio (n=1.832), mientras que solamente 1,2% estudio 18 años. La distribución de los individuos dentro de los años de estudio no es homogénea, como se muestra a continuación:

Ilustración 1: Distribución del dataset de acuerdo a los años de educación



Fuente: Elaboración propia en base al dataset

Métodos utilizados

Con el fin de obtener el mejor modelo que prediga los años de estudio en base a las variables socio-económicas, se comienza el análisis realizando una descripción de las variables en cuestión por medio de técnicas de estadística descriptiva. Además, se aplica el análisis de correspondencia para determinar relación entre las variables categóricas y la variable dependiente, buscando obtener los perfiles de los individuos.

Luego, se aplican diversos modelos, los cuales se testean bajo determinados parámetros (las variables utilizadas en los modelos se encuentra estandarizados). Los métodos que se utilizaron son los que se listan a continuación:

- Mínimos Cuadrados (LS - Least Square),
- Ridge,

- Lasso,
- Regresión de componentes principales (PCR – Principal Component Regression),
- Mínimos Cuadrados Parciales (PLS – Partial Least Square),
- Máquina de Soporte de Vector (SVM – Support Vector Machine),
- Regresión con Arboles (Regression Trees),
- Bagging,
- Boosting, y
- Random Forest.

Para cada uno de estos modelos, se aplicaron dos técnicas de validación:

- Splitting (partición de 80/20 entre training y testing)
- Cross-validation
    - 5 folds,
    - 10 folds,
    - 15 folds, y
    - 20 folds.

Para evaluar cada modelo, se calculó el error de identificación de cada uno por medio de dos métodos diferentes: dividiendo el dataset entre test y training (80% y 20% respectivamente) y cross-validation. En el caso de splitting, se tomó el 80% del dataset y se utilizó para generar el modelo (training) y luego se testeo con el 20% restante (testing). En el caso de cross-validation se utilizaron distintas divisiones con el fin de observar la evolución del error al dividir el dataset en un mayor número de partes.

Se realizó un análisis posterior del modelo obtenido con menor error.

El análisis principal de la base se realizó utilizando el software R, no obstante, mucho de los soportes gráficos se realizaron con Tableau y con Excel.

Resultados

Como se mencionó con anticipación, se comienza el análisis con una descripción de las variables utilizadas. Luego se aplican los modelos de predicción y se evalúa la performance del modelo.

Análisis descriptivo

La Tabla 1 muestra la variable dependiente y las variables categóricas contenidas en el dataset con la cantidad de instancias para cada uno de las categorías que puede tomar la variable. Los datos se muestran en valores absolutos como en porcentajes.

Tabla 1: Características de la población estudiada. Variables discretas (n=4.739)

	Niveles	Cantidad	
		#	%
education	12	1.832	38,7%
	13	613	12,9%
	14	518	10,9%
	15	556	11,7%
	16	907	19,1%
	17	256	5,4%
	18	57	1,2%
etnicity	afam	786	16,6%
	hispanic	903	19,1%
	other	3.050	64,4%
gender	female	2.600	54,9%
	male	2.139	45,1%
fcollege	no	3.753	79,2%
	yes	986	20,8%
mcollege	no	4.088	86,3%
	yes	651	13,7%
home	no	852	18,0%
	yes	3.887	82,0%
urban	no	3.635	76,7%
	yes	1.104	23,3%
income	high	1.365	28,8%
	low	3.374	71,2%
region	other	3.796	80,1%
	west	943	19,9%

Fuente: Elaboración propia en base al dataset

Como se mencionó con anticipación, los años de educación se dividen en 7 valores ordinales: de 12 años a 18 años de estudio, con una diferencia de 1 año entre cada valor. En lo que respecta a los individuos del datase, 38,7% solo estudió 12 años, 12,9% de los individuos estudió 13 años, 10,9% estudió 14 años, 11,7% de los individuos estudió 15 años, mientras que el 19,1% estudió 16 años; por ultimo solamente el 5,4% y 1,2% estudiaron 17 y 18 años respectivamente. Es decir, que un 74,3% de los individuos estudió como máximo 15 años.

Al observar las instancias por etnicidad, se puede ver que el 64,4% de los individuos pertenecen a otro grupo étnico distinto a los afro-americanos e hispanos, estos 2 grupos tienen 16,6% y 19,1% de las observaciones del dataset respectivamente.

En lo que respecta al género, el dataset se encuentra levemente desbalanceado, con la mayoría de las instancias comprendidas por mujeres (54,9%). El 45,1% de las instancias corresponde a hombres.

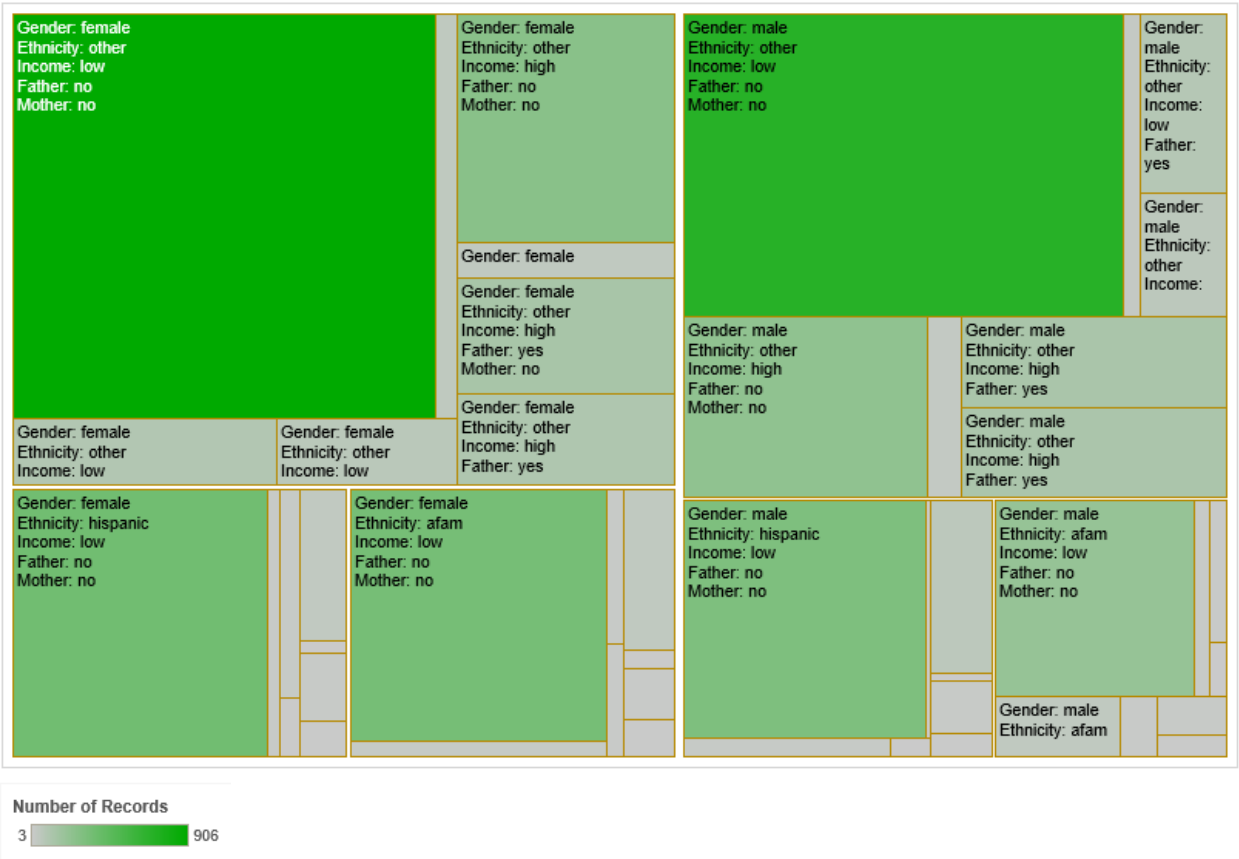
Al dividir la base entre los individuos relevados que tienen a su padre que concurrió a la universidad y los que no (fcollage), se observa que el 79,2% de los padres no fue a la universidad; dejando

solamente el 20,8% de los individuos con padre universitario. Lo mismo sucede en el caso de las madres universitarias, que solo representan el 13,7% de la base (mcollage).

El 82,0% de las familias de los individuos es dueña del hogar en el que viven, mientras que el 18,0% no lo es. Un 76,7% de los individuos concurre a una universidad que no se encuentra situada en una zona urbana, mientras que el 23,3% restante sí. Solamente un 28,8% de los individuos tiene ingresos por encima de 25.000 USD al año (high). Por último, el 80,1% de los individuos pertenece a una región distinta que la región Oeste de EE. UU.

En la Ilustración 2 se puede ver un Tree Map con las variables categóricas gender, ethnicity, income, fcollege y mcollege. El tamaño y el color están dados por la cantidad de registros.

Ilustración 2: Tree Map variables categóricas



Fuente: Elaboración propia en base al dataset

La Ilustración 2 muestra que los dos grupos más grandes de individuos pertenecen a hombres y mujeres de otra etnicidad distinta a afro-americanas o hispanas, de bajo ingreso económico y que ninguno de sus padres (ni madre ni padre) estudio en la universidad. En el caso del grupo anterior de mujeres, se observa un total de 906 registros (19,12%), mientras que el de los hombres 709 registros.

Existen otros 3 grupos, tanto en el caso de las mujeres como de los hombres, que presentan más o menos la misma cantidad de observaciones (teniendo en consideración que el número de mujeres es superior):

- etnicidad hispano, madre y padre sin estudios universitarios y bajos ingresos;
- etnicidad afro-americano, madre y padre sin estudios universitarios y bajos ingresos; y
- otra etnicidad, madre y padre sin estudios universitarios y altos ingresos.

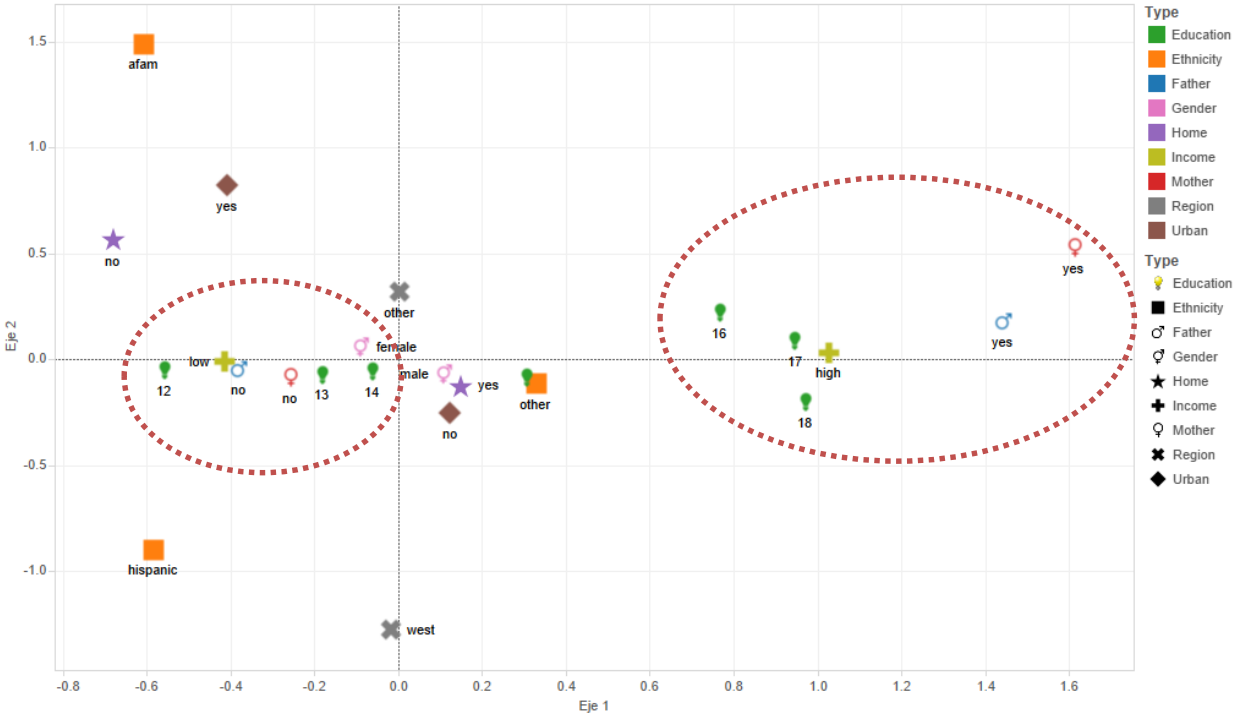
Para el primer caso, el número de observaciones es 366 y 310 para mujeres y hombres respectivamente; el segundo caso tiene 345 mujeres y 209 hombres; y el último caso 264 mujeres y 233 hombres.

Por último, se destaca que uno de los grupos más pequeños corresponde a hombres hispanos de alto ingreso económico, cuyo padre no fue a la universidad pero la madre si (este grupo tiene solamente 3 registros/individuos).

Con el fin de determinar la interacción de las variables categóricas, se aplica la técnica de análisis factorial de correspondencia. El objetivo es lograr explicar el comportamiento de los datos en un número reducido de dimensiones, intentando la menor perdida de información.

La Ilustración 3 muestra los resultados del biplot simétrico para el análisis de correspondencia. Se destaca que tanto el color como la forma de los puntos se encuentran dados por las variables.

Ilustración 3: Biplot simétrico. Codificación por tipo de figura y color



Fuente: Elaboración propia en base al dataset

Por medio de la Ilustración 3 podemos observar dos relaciones interesantes:

- como primera observación, se destaca que los individuos con mayores años de educación (16, 17 y 18 años) tienen altos ingresos y sus padres fueron universitarios;
- la segunda observación radica en que los individuos con menor cantidad de años de estudio (12, 13 y 14 años) tienen bajos ingresos y sus padres no fueron a la universidad.

Por otro lado, los valores próximos al origen son los más usuales. Esto sucede con la región "otra", que las universidades no se encuentren en zonas urbanas, que la familia de los individuos sea dueña de la casa y que la etnicidad no sea ni afro-americano ni hispano. En el caso de la variable género (femenino y masculino), ambos valores se encuentran en el origen dado que el dataset se encuentra balanceado.

La Tabla 2 muestra las medidas de tendencia central y de posición de las variables continuas. La tabla muestra el mínimo y máximo, el 1er y 3er cuartil, y la mediana y media.

Tabla 2: Características de la población estudiada. Variables continuas (n=4.739)

	score	unemp	wage	distance	tuition
Min.	29,0	1,4	6,6	0,0	0,3
1st Qu.	44,0	5,9	8,8	0,4	0,5
Median	51,0	7,1	9,7	1,0	0,8
Mean	51,0	7,6	9,5	1,8	0,8
3rd Qu.	58,0	8,9	10,2	2,5	1,1
Max.	73,0	24,9	13,0	20,0	1,4
Desvio Estandard	8,7	2,8	1,3	2,3	0,3

Fuente: Elaboración propia en base al dataset

La variable *score* tiene un mínimo de 29,0 puntos y un máximo de 73,0 puntos, con una media situada en 51,0 (mismo valor que la mediana). El desvío estándar de esta variable es 8,7 puntos. Un 50% de las observaciones se sitúan dentro del rango de puntos de 44,0 y 58,0.

La tasa de desempleo, variable *unemp,* tiene un mínimo de 1,4 puntos y un máximo de 24,9 puntos, con una media situada en 7,6 y una mediana de 7,1. El desvío estándar de esta variable es 2,8 puntos. Un 50% de las observaciones se sitúan dentro del rango de puntos de 5,9 y 8,9.

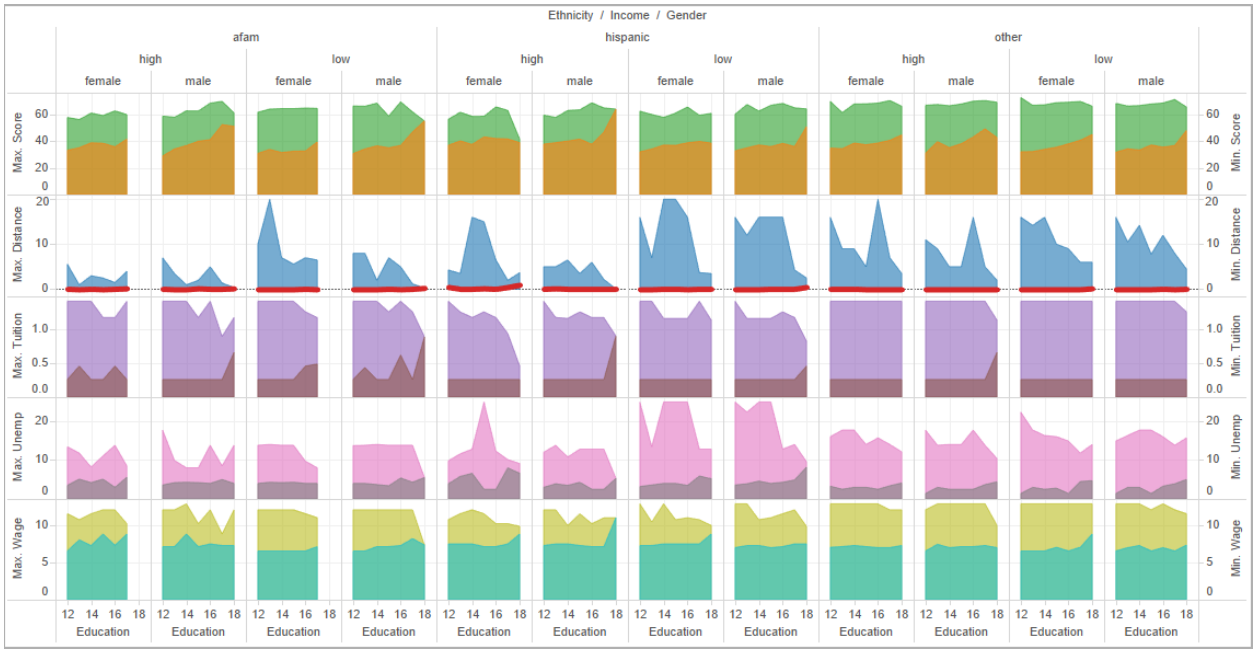
El salario promedio por hora, variable *wage,* tiene un mínimo de 6,6 USD/h y un máximo de 13,0 USD/h, con una media situada en 9,5 y una mediana de 9,7 USD/h. El desvío estándar de esta variable es 1,3 USD/h. Un 50% de las observaciones se sitúan dentro del rango de 8,8 y 10,2 USD/h.

La distancia de la universidad, variable *distance,* tiene un mínimo de 0,0 millas y un máximo de 20,0 millas, con una media situada en 1,8 y una mediana de 1,0 millas. El desvío estándar de esta variable es 2,3 millas. Un 50% de las observaciones se sitúan dentro del rango de 0,4 y 2,5 millas.

El promedio por mes del costo universitario, variable *tuition,* tiene un mínimo de 0,3 miles USD y un máximo de 1,4 miles USD, con una media situada en 0,8 (valor igual al de la mediana). El desvío estándar de esta variable es 0,3 miles USD. Un 50% de las observaciones se sitúan dentro del rango de 0,5 y 1,1 miles USD.

La Ilustración 4 muestra el valor máximo (eje izquierda) y el mínimo (eje derecha) de las variables numéricas por años de educación dividido para las diferentes categorías de etnicidad, ingreso y por género.

Ilustración 4: Máximo y mínimo de las variables numéricas por etnicidad, ingreso y genero para cada uno de los años de educación



Fuente: Elaboración propia en base al dataset

La Ilustración 4 permite observar de forma gráfica como varían los valores extremos de las variables numéricas para cada uno de los años de educación. Al agregar la división con las variables categóricas se busca determinar diferencias importantes entre los perfiles estipulados. De esta forma, se analizan las variables numéricas por etnicidad, ingreso y género.

La variable *score* muestra valores similares entre los perfiles, se observa un leve incremento en los individuos con etnicidad "other". Al aumentar los años de estudio el score mínimo aumenta, siendo más pronunciado en los hombres que en las mujeres. No se observa un aumento significativo en el score máximo al aumentar los años de estudio.

En lo que respecta a la *distancia*, en promedio se observa que las mujeres viajan más que los hombres, teniendo la misma distancia mínima (distancia igual a 0). Se observa que existe una relación entre las personas de mismo perfil que viajan menos y alto nivel económico.

Los individuos con etnicidad distinta a afro-americanos o hispanos tiene los gastos universitarios (*tuition*) más altos. Para este grupo, el valor mínimo es igual salvo para los hombres de altos ingresos que estudiaron 18 años, donde el valor mínimo aumenta. Los afro-americanos e hispanos pagan un valor menor. En el caso de los hombres hispanos, al aumentar los años de estudio, aumenta el mínimo abonado y cae el máximo. En las mujeres hispanas, sucede algo similar, solo que el mínimo se mantiene. Y para los afro-americanos, a medida que aumentan los años de estudio, en muchos casos cae el máximo.

La tasa de desempleo (*unemp*) del estado muestra tasas inferiores para la etnicidad afro-americanos, y tasas muy dispares entre los hispanos al analizarla por los años de estudio. Los individuos con otrasetnicidad muestran tasas de desempleo sin tantas variaciones. Con excepción de las mujeres hispanas de altos ingresos, se observa un comportamiento similar entre los individuos de

sexo femenino y masculino. El perfil de individuos hispanos con bajos recursos presentan las tasas más altas de desempleo, bajando con el aumento de los años de estudio.

Por último, en lo que respecta al salario promedio por hora (*wage*), se observa una leve diferencia entre los hombres y las mujeres, siendo el salario de las mujeres levente menor. Los afro-americanos tienen brechas más chicas para los individuos de altos ingresos, no obstante los valores son dispares según los años de estudio. Finalmente, los individuos con otra etnicidad distinta a afro-americano o hispano presentan la mayor brecha entre el salario máximo y mínimo.

La Tabla 3 muestra la matriz de correlación entre las variables. En este caso se omiten las variables *education* y *ethnicity* dado que son variables categóricas. Se destaca que las variables dicotómicas se transforman en valores dummies, tomando valor 1 y 0.

Tabla 3: Matriz de correlación entre las variables

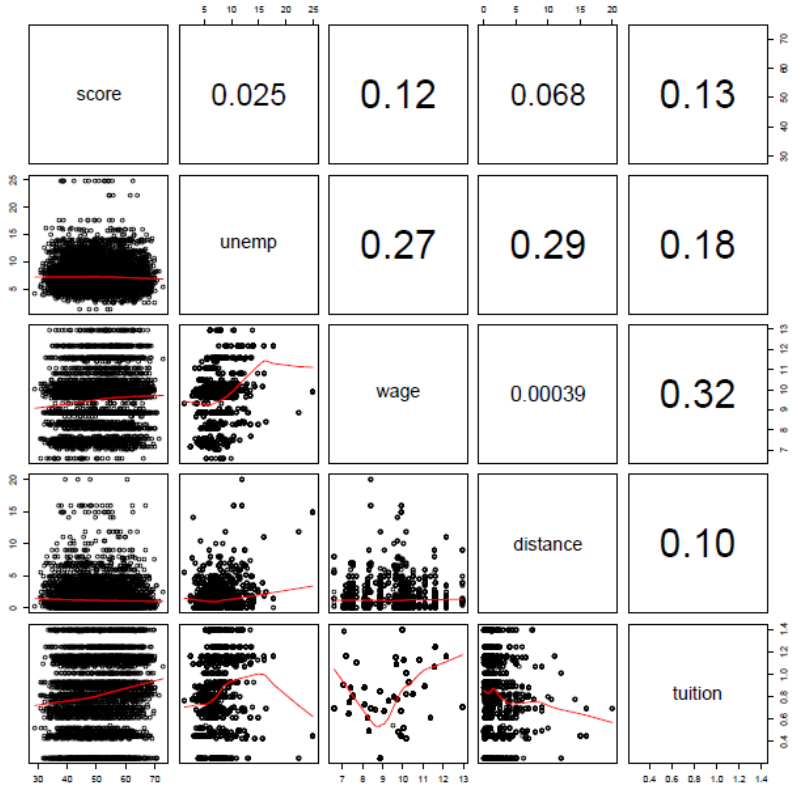
	gender	score	fcollege	mcollege	home	urban	unemp	wage	distance	tuition	income	region
gender	1,00	0,08	0,04	0,02	0,04	0,01	-0,03	0,03	0,00	0,01	-0,06	-0,01
score	0,08	1,00	0,25	0,19	0,13	-0,09	-0,03	0,12	-0,07	0,13	-0,18	-0,03
fcollege	0,04	0,25	1,00	0,43	0,08	-0,05	-0,10	0,03	-0,11	0,03	-0,35	0,03
mcollege	0,02	0,19	0,43	1,00	0,06	-0,03	-0,09	0,02	-0,08	0,04	-0,25	-0,01
home	0,04	0,13	0,08	0,06	1,00	-0,10	0,01	0,07	0,02	0,00	-0,14	0,00
urban	0,01	-0,09	-0,05	-0,03	-0,10	1,00	-0,05	-0,03	-0,29	-0,02	0,07	-0,05
unemp	-0,03	-0,03	-0,10	-0,09	0,01	-0,05	1,00	0,27	0,29	0,18	0,08	-0,04
wage	0,03	0,12	0,03	0,02	0,07	-0,03	0,27	1,00	0,00	0,32	-0,07	-0,08
distance	0,00	-0,07	-0,11	-0,08	0,02	-0,29	0,29	0,00	1,00	-0,10	0,08	0,07
tuition	0,01	0,13	0,03	0,04	0,00	-0,02	0,18	0,32	-0,10	1,00	-0,05	-0,58
income	-0,06	-0,18	-0,35	-0,25	-0,14	0,07	0,08	-0,07	0,08	-0,05	1,00	-0,01
region	-0,01	-0,03	0,03	-0,01	0,00	-0,05	-0,04	-0,08	0,07	-0,58	-0,01	1,00

Fuente: Elaboración propia en base al dataset

Podemos observar que las variables no tienen una correlación fuerte entre sí. Notemos que existe una leve correlación negativa entre la región y el costo universitario igual a -0,58 (región y tuition).

La Ilustración 5 muestra un scatter plor matrix de la correlación entre las variables numéricas del dataset. En la diagonal se pueden ver los nombres de las variables en cuestión. Por debajo de la diagonal se ven los scatter plots, mientras que en la parte superior se ven los valores absolutos de la correlación.

Ilustración 5: Matriz de correlación entre las variables numéricas



Fuente: Elaboración propia en base al dataset

Se puede observar que no existe una relación entre las variables en cuestión. Notemos que los scatter plots no muestran una tendencia ni un patrón entre los datos. Tanto la Tabla 3 como la Ilustración 5 muestran que no existe correlación entre las variables, por lo tanto no se considera la existencia de multi-colinealidad que puede traer problemas en la estimación de los modelos.

Resultado de los modelos aplicados

La Tabla 4 muestra el error de identificación de los modelos aplicados. Además, se muestra en forma de mapa de calor los modelos con menor error de identificación utilizando splitting (azul) y para los de cross-validation (rojo).

Tabla 4: Error de identificación de los modelos

Error de identificacion		Splitting	Cross-validation			
			5 folds	10 folds	15 folds	20 folds
LS		2,2400	4,0459	4,0200	4,0888	4,0879
Ridge		2,4613	4,0005	3,9665	3,9363	3,9875
Lasso		2,2919	3,9230	3,9608	3,9230	3,8809
PCR		2,4012	3,9464	3,8840	3,8899	3,9398
PLS		2,2690	4,0866	4,0235	4,0143	4,0937
SVM	**Default**	2,3272	4,7179	4,6626	4,6465	4,7292
	Laplace	2,3478	4,6877	4,6286	4,6262	4,7073
Regression Trees		2,6727	3,7758	3,7460	3,8451	3,9007
Bagging		2,4200	3,8971	3,8674	3,8343	3,8879
Boosting		2,6538	3,3696	3,3535	3,3440	3,3754
Random Forest		2,3576	4,0807	4,0750	4,0235	4,1260

Fuente: Elaboración propia en base al dataset

La Tabla 4 muestra que el modelo que más se ajusta a los datos es el modelo de regresión lineal (LS). No obstante, al analizar el dataset con cross-validation, Boosting presenta las mejores estimaciones en todas las particiones realizadas. Es importante destacar que este modelo tiene uno de los mayores valores obtenidos del error cuadrático medio bajo splitting. Esto puede suceder por el algoritmo utilizado por Boosting, donde se utiliza un residuo de los árboles para generar nuevos árboles.

Los parámetros utilizados para generar el modelo de Boosting son:

- Distribución Gaussiana: esto permite obtener el error cuadrático medio,
- Cantidad de hojas: 10
- Fracción del dataset utilizada para el Bag fracción: 1%
- Número total de árboles: 1.000

Estos parámetros se repiten para todas las particiones realizadas en cross-validation.

En el caso del modelo LS el error cuadrático medio calculado sobre el set de testing (20% del dataset) es de 2,24. Dado que este error depende de la partición aleatoria realizada (por lo tanto distintas particiones pueden generar resultados diferentes), y dado que no utiliza toda la información para construir el modelo o calcular el error, se utiliza la estimación por cross-validation igual a 4,02 para 10 particiones. Notemos que a pesar que este error aumenta en comparación al error de Boosting, no deja de ser un error bajo.

Dado que el modelo LS presenta buenos resultados, se analizan los valores obtenidos de la regresión. La Tabla 5 muestra los coeficientes obtenidos de la regresión por mínimos cuadrados ante splitting. La tabla muestra los coeficientes de la regresión y el nivel de significatividad de cada una de las variables.

Tabla 5: Coeficientes de la regresión de mínimos cuadrados ordinarios (LS)

Coeficientes	Estimado	Error Estandar t	valor	Pr(>|t|)	
Intercepto	13,31	0,26	50,95	<2,00E-16	***
Gender	-0,11	0,05	-2,09	0,04	*
Score	0,79	0,03	2,83	<2,00E-16	***
fcollege	0,55	0,07	7,57	0,00	***
mcollege	0,38	0,08	4,60	0,00	***
home	0,12	0,07	1,76	0,08	.
urban	0,04	0,06	0,64	0,53	
unemp	0,08	0,03	2,97	0,00	**
wage	-0,06	0,03	-2,12	0,03	*
distance	-0,07	0,03	-2,60	0,01	**
tuition	-0,06	0,03	-1,70	0,09	.
income	-0,35	0,06	-5,85	0,00	***
region	-0,18	0,08	-2,28	0,02	*
ethnicity.hisp	0,35	0,07	4,93	0,00	***
ethnicity.afam	0,31	0,08	4,03	0,00	***

Fuente: Elaboración propia en base al dataset
*Nota: códigos de significatividad: 0-0,001 "***", 0,001-0,01 "**", 0,01-0,05 "*", y 0,05-0,1 ".".*

En la Tabla 5 se puede observar que algunas de las variables (como home, urban y tuition) no son significativas. Con el fin de mejorar los resultados, se realiza un proceso de selección de variables (stepwise). En este proceso, se elimina la variable *urban*, logrando mejorar el R^2 a 27,29% (con el anterior modelo el R^2 era 26,93%). La Tabla 6 muestra los resultados de la nueva regresión, para este modelo el error cuadrático medio es 2,24.

Tabla 6: Coeficientes de la regresión de mínimos cuadrados ordinarios (LS) con variables elegidas por stepwise

Coeficientes	Estimado	Error Estandar t	valor	Pr(>|t|)	
intercepto	13,40	0,22	61,11	<2,00E-16	***
gender	- 0,13	0,04	- 2,86	0,00	**
score	0,78	0,02	31,55	<2,00E-16	***
fcollege	0,55	0,06	8,60	<2,00E-16	***
mcollege	0,38	0,07	5,29	0,00	***
home	0,14	0,06	2,42	0,02	*
unemp	0,08	0,02	3,36	0,00	***
wage	- 0,05	0,02	- 2,01	0,04	*
distance	- 0,09	0,02	- 3,73	0,00	***
tuition	- 0,07	0,03	- 2,38	0,02	*
income	- 0,38	0,05	- 7,08	0,00	***
region	- 0,19	0,07	- 2,75	0,01	**
ethnicity.hisp	0,32	0,06	5,08	0,00	***
ethnicity.afam	0,32	0,07	4,72	0,00	***

Fuente: Elaboración propia en base al dataset
*Nota: códigos de significatividad: 0-0,001 "***", 0,001-0,01 "**", 0,01-0,05 "*", y 0,05-0,1 ".".*

El modelo obtenido bajo el proceso de eliminación de las variables stepwise, es el que se muestra en la siguiente ecuación:

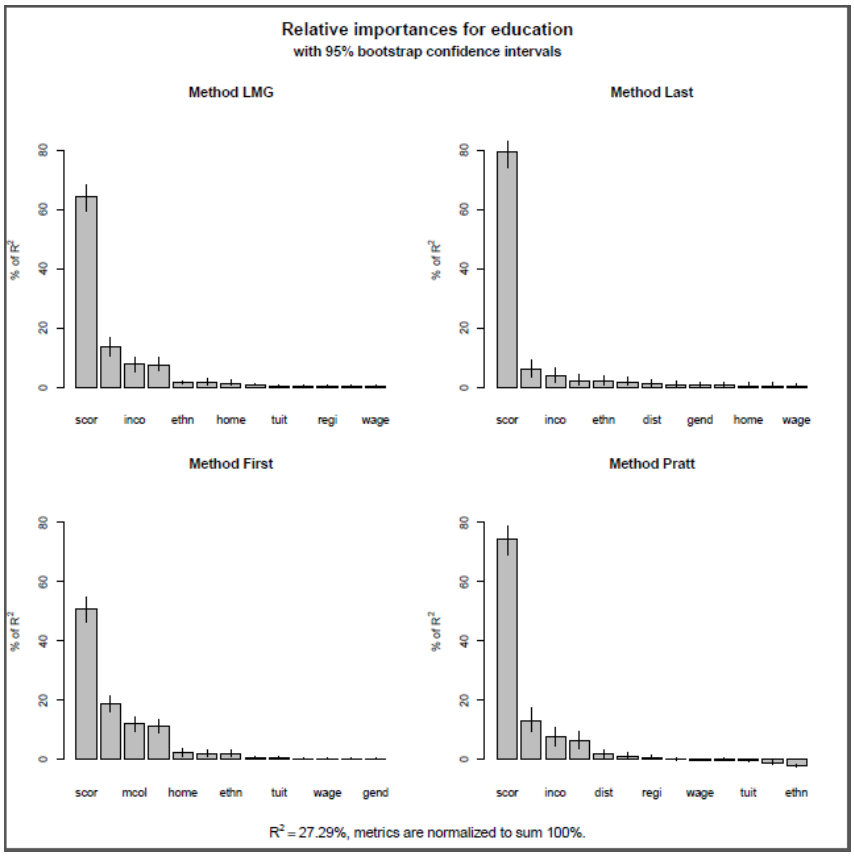
$$\widehat{education} = 13{,}40 - 0{,}13 * gender + 0{,}78 * score + 0{,}55 * fcollege + 0{,}38 * mcollege + 0{,}14 * home + 0{,}08 * unemp - 0{,}05 * wage - 0{,}09 * distance - 0{,}07 * tuition - 0{,}38 * income - 0{,}19 * region + 0{,}32 * ethnicity.hisp + 0{,}32 * ethnicity.afam$$

En la ecuación se puede ver el impacto de las variables del modelo. Algunas variables tienen un impacto negativo en la cantidad de años de estudio, y otras tienen un impacto positivo. Haber tenidos padres universitarios aumentan los años de estudio de los individuos, y este aumento es mayor si el padre fue universitario (0,55 versus 0,38 años de estudio). El tener mejores resultados en las pruebas de los años base aumentan los años de estudio en 0,78 años. Otra variable que incrementa los años es poseer un hogar (0,14). Tener una etnicidad hispana aumenta 0,32 años de estudio en los individuos en relación a las otras etnicidades (mismo valor obtenido para los afro-americanos).

Notemos que la ecuación nos indica que existe una relación negativa con el género, es decir que cuando los individuos son hombres caen los años de educación en 0,13. El salario promedio por hora de la región disminuye los años de estudio en 0,05 años. Otro desincentivo a continuar estudiando es la distancia, disminuyendo los años de estudio en 0,09. En lo que respecta a los costos de estudio, la variable tuition disminuye los años de estudio en 0,07 años. Si el ingreso familiar se encuentra por encima de 25.000 USD los años de estudio disminuyen en 0,38. Los individuos que pertenecen al Oeste del país, disminuyen el estudio en 0,19 años.

La Ilustración 6 muestra la importancia relativa de cada variable en el modelo. Se utilizaron cuatro métricas distintas.

Ilustración 6: Importancia relativa de cada predictor del modelo seleccionado



Fuente: Elaboración propia en base al dataset

Como podemos observar, las variables más relevantes y que mayor impacto tienen en la variable dependiente para el modelo LS estipulado son:

- score,
- fcollege,
- mcollege, y
- income.

Las cuatro métricas calculadas otorgaron resultados similares.

Al omitir las variables que no tienen demasiado impacto, la ecuación nos indica que los mayores años de estudio se encuentran relacionados con ser mujer, haber tenido mejores resultados en los exámenes (variable score), tener padres con estudios universitarios, tener un hogar propio, vivir próximo a universidades con planes de 4 años y venir de un hogar con ingresos altos.

Conclusiones

El presente trabajo tiene por objetivo generar un modelo que pueda predecir los años de estudio de los individuos utilizando un dataset de 4.739 individuos con 14 variables económicas y demográficas.

El análisis realizado incluye analizar los datos con técnicas de estadística descriptiva y por medio de modelos de regresión. Se construyeron diferentes modelos y se aplicaron dos métodos para calcular

el error cuadrático medio: dividir la base en training- testing y la aplicación de cross-validation (con particiones de 5, 10, 15 y 20 folds).

Los dos perfiles con mayor cantidad de observaciones pertenece a hombres y mujeres de otra etnicidad, de bajo nivel económico y que ninguno de sus padres concurrió a la universidad. El grupo de mujeres corresponde al 19,12% del dataset.

Al realizar el biplot simétrico proveniente del análisis de correspondencias con las variables categóricas, se puede observar que los individuos con más años de educación tienen altos ingresos y sus padres tienen un título universitario. En contraposición, individuos con bajos ingresos y cuyos padres no son universitarios se relacionan con los años más bajos de estudio.

Al analizar las variables numéricas desagregadas por algunas de las variables categóricas, se pudo identificar las siguientes conclusiones:

- al aumentar los años de estudio el score mínimo aumenta, siendo más pronunciado el aumento en los hombres que en las mujeres, es decir, que los individuos con más años de estudio obtuvieron score más altos;
- se observa una relación entre las personas que viajan menos y el alto nivel económico;
- existe una disminución en los gastos de estudio al aumentar los años de estudio para los individuos de raza hispana o afro-americana.

Los resultados obtenidos por el modelo de regresión, expresados en la educación en el apartado anterior, indican que mayores años de estudio se encuentran relacionados con:

- ser mujer,
- haber tenido mejores resultados en los exámenes del secundario,
- tener padres con estudios universitarios,
- que el hogar sea propio,
- vivir próximo a una universidad con programa de 4 años, y
- venir de una familia con ingresos altos.

Es importante destacar, que lo analizado con técnicas de estadísticas descriptivas concuerda con los resultados obtenidos en el modelo LS.

Con el fin de mejorar el análisis, se propone para un trabajo futuro, realizar cambios en los parámetros de Boosting y Regression Trees. Estos dos modelos obtuvieron un error cuadrático medio bajo para la técnica de cross-validation. Además, se podría tratar de actualizar la información y tratar de obtener datos para otros años, para agrandar la muestra y mejorar el modelo de regresión, y también para realizar comparación entre los años en cuestión.

Bibliografía

- Peter Dalgaard (2008), “Introductory Statistics with R”, Springer.
- Trevor Hastie, Robert Tibshirani, Jerome Friedman, “The Elements of Statistical Learning Data Mining, Inference, and Prediction”, Springer.
- Stock & Watson http://wps.aw.com/aw_stock_ie_2/