Solucion Datathon

Sahlre

1 de julio de 2017

Introducción

Lo que esta en negrilla es un link que llevará al respectivo documento, para que esto funcione debe descargar primero el pdf.

En este informe se presenta la solución del equipo Sahlre al Desafio Inversión Educativa. Desafio que consiste en responder: Si tuviera un fondo de 100 millones de dólares ¿Qué inversiones público - privadas priorizaría con el objetivo de garantizar la calidad, pertinencia, equidad e inclusión de la educación en América Latina?

Los datos abiertos relacionados al sector educativo contienen diversas mediciones (variables o características) asociadas a los hogares que podrían dar solución al desafió de inversión educativa en Colombia. Una de ellas es el nivel educativo de las personas, variable que se convierte en la variable resultado o dependiente que se intentará predecir con las otras características asociadas a cada una de los individuos, como por ejemplo, sólo por mencionar algunas, el género, el material predominante de las paredes exteriores y estado laboral.

Con dichas características se construyeron diversos algoritmos supervisados para predecir el nivel educativo de las personas: *Uno vs todos Regresión Logistica Multivariada*, *Árbol de clasificación*, *Bagging con árboles de clasificación*, *Boosting con árboles de clasificación* y *Bosques Aleatorios de clasificación*. La decisión de incluir varios algoritmos es poder propener en esta solución ambos aspectos: predicción e interpretabilidad. Es sabido el *trade off* que existe entre esos aspectos, a mayor capacidad de predicción de un algoritmo menor será su interpretabilidad. Sin interpretabilidad es díficil propener solución alguna al desafio de inversión educativa.

Metodología

Se elegieron las bases de datos del mes de abril de la **Gran Encuesta Integrada de Hogares (GEIH) - 2017**. Luego con el objetivo de tener una base de datos a nivel nacional con la mayor cantidad de observaciones se seleccionaron las bases de datos: *Características Generales*, *Fuerza de trabajo*, *Otras actividades y ayudas en la semana*, *otros ingresos* y *Vivienda y hogares*, tanto para Areas metropolitanas como para Cabeceras y zonas Rurales. Con cada una de esas bases de datos se procedió a seleccionar aquellas variables con un porcentaje de valores pérdidos inferior al 10%. Tal decisión se toma con la intención de crear un modelo predictivo del nivel educativo que aproveche la mayor cantidad de observaciones de las bases de datos de la GEIH. De esta manera, la propuesta de este análisis de datos no se convierte en una respuesta única a la variación del nivel educativo de los colombianos, sino en una aproximación al mejor modelo predictivo del nivel educativo de los colombianos a partir de la GEIH. El código para cargar y unir las bases de datos está en **Carga de datos**, este produce una base de datos que se ha denominado *bd_Nac.csv*. El nombre de las variables, su significado y los valores asumidos, con su respectiva codificación se puede explorar en **Definicion de variables**

En la base de datos *bd_Nac.csv* aún hay presencia de valores pérdidos ya que estos no fueron representados a través de un valor vació sino como un valor númerico. Se procede a eliminarlos, además se crea la variables resultado o dependiente sobre la que se intentará realizar la predicción. Se decide transformar la variable P6210 (Nivel educativo mas alto alcanzado) en una variable con tres niveles: Bajo, Medio, Alto.

- El nivel bajo esta conformado por los valores: 1. Ninguno, 2. Preescolar, 3. B?sica primaria, 4. B?sica secundaria.
- El nivel medio por: 5. Media.
- El nivel alto por: 6. Superior o universitaria.

Este último procedimiento se encuentra soportado en el archivo **Preproceso de datos 1**. Allí el código produce la base de datos: *bd_Nac_final.csv*

Finalmente en el proceso de construcción de la base de datos sobre la que se realizará el entrenamiento de los algoritmos así como su evaluación de pronóstico, en el archivo **Datos de entrenamiento y de prueba** se encuentra el código que produce las bases de datos de entrenamiento, validación y prueba. La base de datos de entrenamiento *bd_Train.csv* tiene 54397 observaciones y 61 variables, y la base de datos de prueba *bd_Test.csv* tiene 13598 observaciones y 61 variables.

Predicción del nivel educativo

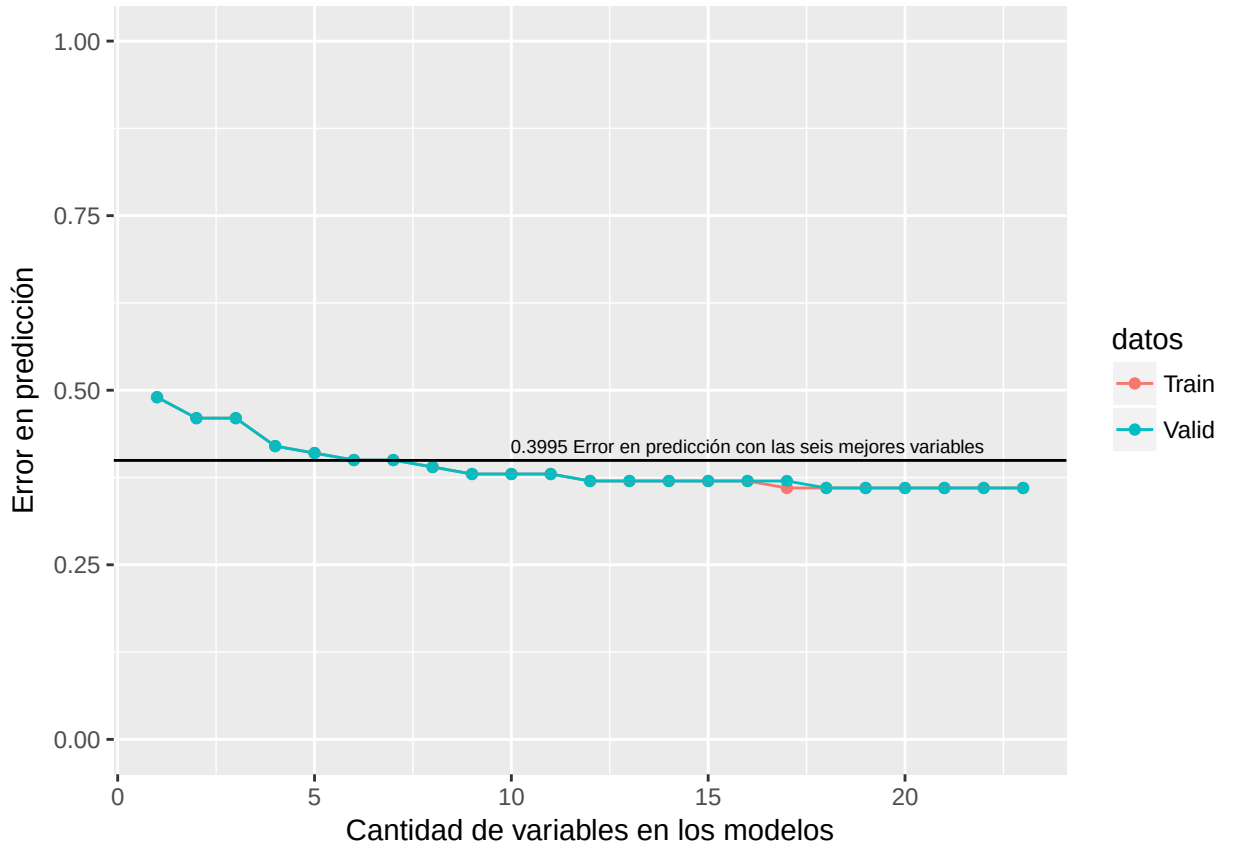
Uno vs todos Regresión Logistica Multivariada

El código para estimar el algorítmo se encuentra en el archivo **FSS Uno vs Todos**

Cómo el nivel educativo, variable dependiente, asume tres niveles: Bajo-Medio-Alto, la predicción de esta variable no se puede ejecutar a través de la Regresión Logistica Multivariada adecuada para una variable resultado binaria. Este algoritmo consiste en aplicar una regresión logística multivariada por cada nivel asumido en la variable dependiente. De esta manera se estiman tres modelos: el primero con la variable dependiente asumiendo sólo dos valores: "Bajo" y "El resto", el segundo con la variable dependiente: "Medio" y "El resto", y el tercero con: "Alto" y "El resto".

Con el propósito de crear un modelo con mayor nivel de interpretación se utilizó la técnica de selección de variables: Forward Stepwise Selection (FSS).

Los resultados de las primeros 23 modelos se presentan en la siguiente gráfica.



Se observa que a partir de las seis mejores variables el error de predicción tanto para datos de entrenamiento (Train) como para datos de validación (Valid) disminuye lentamente. De esta manera se eligen las primeras seis variables seleccionadas a partir del método FSS que son: P5210s16, P6100, P6040, P6240, P6170, P4020.

Al estimar el modelo con las seis variables se producen los siguientes indicadores:

- Error de clasificación de entrenamiento = 0.3995
- Error de clasificación de prueba = 0.4002

Este modelo es aceptable si se tiene en cuenta que al estimar el modelo con todas (60) las variables los indicadores son:

- Error de clasificación de entrenamiento = 0.3590
- Error de clasificación de prueba = 0.3602

También es importante mencionar que en la gráfica anterior se observa que la línea de error de predicción para datos de validación y datos de entrenamiento se solapan. Esto sugiere que hay alto sesgo. Es decir la hipótesis del algoritmo planteado es muy simple para modelar el nivel educativo. Otra señal de alto sesgo es cuando el error de clasificación es alto y el error de clasificación en datos de entrenamiento y en datos de validación son muy similares.

Resultados que indican varios caminos a seguir si se desea mejorar la predicción:

- Permitir al modelo capturar relaciones no lineales. Esto se lograría a través de interacción entre variables y aumentar el grado de polinomio de aquellas variables que sean continuas.
- Usar algorítmos como las Redes Neuronales, las Máquinas de Soporte Vectorial o Bosques Aleatorios, reconocidos por la capacidad de predicción en escenarios complejos.
- Adicionar nuevas variables.

Árbol de Clasificación.

El código para estimar el algorítmo se encuentra en el archivo **Tree**

Los métodos basados en árboles consisten en segmentar el espacio de variables predictoras (independientes) dentro de un número más simple de regiones. Luego de estimar el árbol de clasificación para las 60 variables predictoras o independientes se obtiene:

- Error de clasificación de entrenamiento = 0.3945
- Error de clasificación de prueba = 0.4029

Con el algortimo Árbol de clasifiacación sucede algo similar al resultado del anterior algoritmo. Los errores de clasificación en datos de entrenamiento y de prueba son altos y similares, lo que indica alto sesgo. De esta manera, para aumentar el nivel de predicción sería necesario utilizar algorítmos que produzcan hipótesis no lineales complejas como lo son las Redes Neuronales, las Maquinas de Soporte Vectorial o los Bosques Aleatorios.

Bagging en Árboles de clasificación

El código para estimar el algorítmo se encuentra en el archivo **Bagging**

Uno de los problemas más recurrentes de los árboles de decisión es la débil precisión en el pronóstico. El método Bagging o Bootstrap aggregation es un procedimiento para reducir la varianza de los algorítmos supervisados. Luego de estimar el algorítmo con 25 árboles para todas la variables predictoras se obtiene:

- Error de clasificación de entrenamiento = 0.007
- Error de clasificación de prueba = 0.1925

En esta oportunidad se obtienen menor error de predicción tanto en datos de entrenamiento como de prueba. También se presenta una amplia brecha de entre el error de entrenamiento y el error de prueba. Esto indica que el modelo estimado sufre de varianza. Aunque el modelo sufre de alta varianza posee mejor poder de predicción que los anteriores algorítmos estimados. Tal resultado sugiere posibles caminos para mejorar el algorítmo:

- Usar diferente cantidad de árboles al momento de estimar el algorítmo. Si se explora el archivo **Bagging** la cantidad de árboles utilizado fue 25.
- Utilizar una técnica de selección de variables como Forward Stepwise Selectión, Backward Stepwise Selection o Principal Components Analysis.

Bosques Aleatorios

El código para estimar el algorítmo se encuentra en el archivo **RandomForest**

Los Bosques Aleatorios proporcionan una mejora sobre el algorítmo Bagging en Árboles de Clasificación, a través de un ajuste aleatorio que evita la asociación de los árboles creados. Luego de estimar el algorítmo con 25 árboles para todas la variables predictoras se obtiene:

- Error de clasificación de entrenamiento = 0.008
- Error de clasificación de prueba = 0.1958

Los resultados son similares al los obtenidos con el algorítmo Bagging de Árboles de Clasificación. El algorítmo Bosques Aleatorios estimado sufre de varianza aunque tienen mejores resultados en predicción que los tres primeros algoritmos estimados en este estudio. Este resultado sugiere los siguientes caminos para mejorar la capacidad de predicción:

- Usar diferente cantidad de árboles al momento de estimar el algorítmo. Si se explora el archivo **RandomForest** la cantidad de árboles utilizado fue 25.
- Utilizar una técnica de selección de variables como Forward Stepwise Selectión, Backward Stepwise Selection o Principal Components Analysis.

Boosting en Árboles de Clasificación.

La idea del algorítmo Boosting en Árboles de Clasificación es mejorar la poder de predicción de los árboles de clasificación. Es parecido al algorítmo Bagging en Árboles de Clasificación excepto por el método de remuestreo utilizado para estimar los árboles de decisión. Luego de estimar el algorítmo con 25 árboles para todas la variables predictoras se obtiene:

- Error de clasificación de entrenamiento = 0.3602
- Error de clasificación de prueba = 0.3649

Con el algortimo Boosting en Árbols de clasifiacación sucede algo similar al resultado de los dos primeros algorítmos. Los errores de clasificación en datos de entrenamiento y de prueba son altos y similares, lo que indica alto sesgo. De esta manera, para aumentar el nivel de predicción sería necesario utilizar algorítmos que produzcan hipótesis no lineales complejas como lo son las Redes Neuronales, las Maquinas de Soporte Vectorial o los Bosques Aleatorios, reducir el número de varibles a través de métodos de selección de variables o usar otras variables.

Interpretación de modelos

De los algorítmos estimados *Uno vs todos Regresión Logistica Multivariada* y *Árbol de clasificación* son los de interpretación más sencilla, aunque son los de menor poder de predicción. En cambio aquellos de mayor poder de predicción son de interpretación más díficil debido a la hipótesis complejas sobre las que se fundamentan.

Recordar que hay un trade off entre interpretación y nivel decomplejidad de las hipótesis producidas por los algorítmos. Además con la intención de responder al desafió de inversión es necesario identificar la relación entre las variables dependientes con el nivel de educación.

Interpretación del Árbol de Clasificación

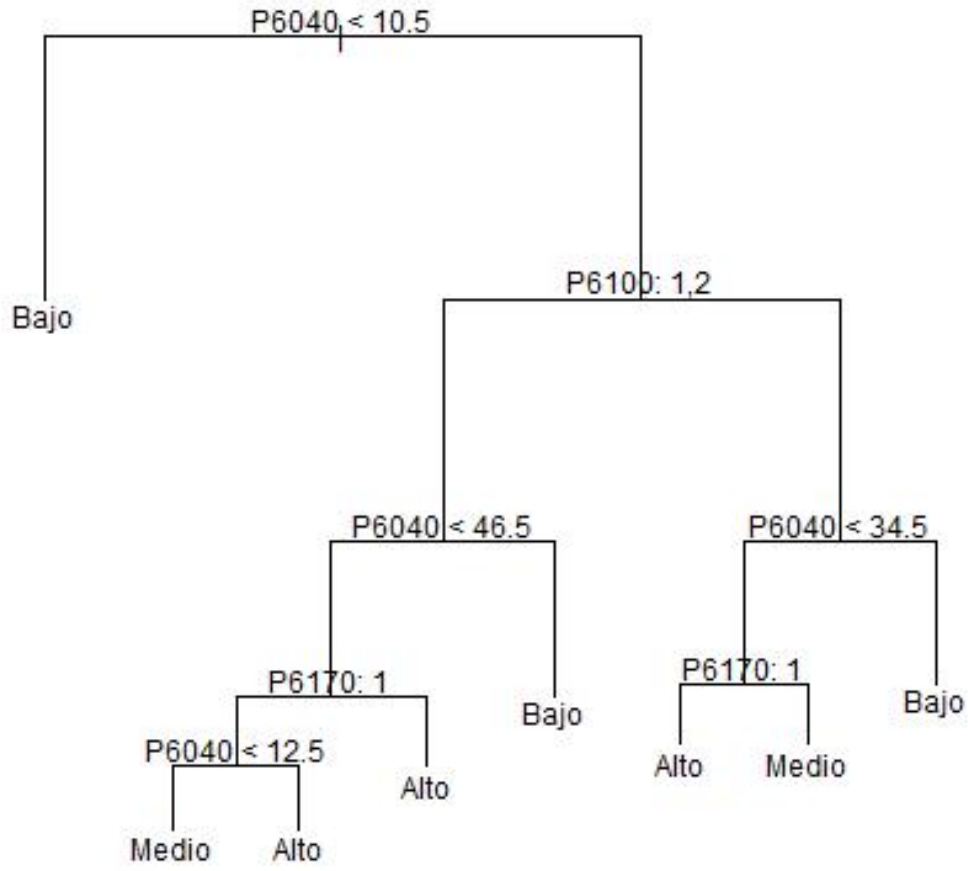
En la figura 1 se observa el Árbol de clasificacion estimado. Tiene 7 nodos internos, 8 nodos términales, y las variables usadas para su construcción son P6040 (Edad), P6100 (¿A cuál de los siguientes regimenes de seguridad social en salud esta afiliado? 1. Contributivo (EPS), 2. Especial (Fuerzas Armadas, Ecopetrol, Universidades Públicas), 3. Subsidiado (EPS-S)) y P6170 (¿Actualmente, asiste al preescolar, escuela, colegio o universidad? 1. Si, 2. No.). La interpretación es de este tipo de algorítmos es muy sencilla:

Nivel bajo

- Personas menores a los 11 años tienen nivel educativo bajo.
- Personas con Régimen de Seguridad Social en Salud (RSSS) subsidiado y mayores a 35 años tienen un nivel educativo bajo.
- Personas con RSSS contributivo o especial y mayores a 47 años tienen nivel educativo bajo.

Nivel medio y alto

- Personas mayores entre los 11 y 35 años con RSSS subsidiado que se encuentran estudiando tienen un nivel educativo alto, aquellos que no están asistiendo a centros educativos tienen un nivel educativo medio.
- Personas con RSSS contributivo o especial entre los 11 y 47 años que no se encuentran estudiando tienen un nivel educativo alto.
- Personas con RRSS contributivo o especial entre los 13 y 47 años que se encuentran estudiando o no estudiando tienen un nivel educativo alto.
- Personas con RSSS contributivo o especial entre los 11 y 13 años que se encuentran estudiando tienen nivel



educativo medio.

Interpretación de Uno vs todos Regresión Logistica

Para facilitar la interpretación del algorítmo se estima con la varible respuesta como una variable binaria donde el valor 1 corresponde a nivel educativo alto y el valor 0 a nivel educativo bajo o medio. Se estima con las seis mejores variables obtenidas a través del proceso Forward Stepwise Selection. Ellas son:

- P5210s16 (¿Hay computador (para uso del hogar)? 1. Si, 2. No.)
- p6100 (¿A cuál de los siguientes regimenes de seguridad social en salud esta afiliado? 1. Contributivo (EPS), 2. Especial (Fuerzas Armadas, Ecopetrol, Universidades Públicas), 3. Subsidiado (EPS-S))
- P6040 (¿Cuántos años cumplidos tiene?)
- P6240 (¿En qué actividad ocupó la mayor parte del tiempo la semana pasada? 1. Trabajando, 2. Buscando trabajo, 3. Estudiando, 4. Oficios del hogar, 5. Incapacitado permanente para trabajar, 6. Otra actividad.)
- P6170 (¿Actualmente, asiste al preescolar, escuela, colegio o universidad? 1. Si, 2. No.)
- P4020 (¿Cuál es el material predominante de los pisos de la vivienda? 1. Tierra, arena, 2. Cemento, gravilla, 3. Madera burda, tabla, tablón, otro vegetal, 4. Baldosín, ladrillo, vinisol, otros materiales

sintéticos, 5. Mármol, 6. Madera pulida, 7. Alfombra o tapete de pared a pared.)

```
##
## Call:
## glm(formula = as.factor(ifelse(bd_Train1$P6210nuevo == "Alto",
##     "1", "0")) ~ P5210s16 + P6100 + P6040 + P6240 + P6170 + P4020,
##     family = "binomial", data = bd_Train1[, -c(61)])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4773  -0.8016  -0.4293   0.9831   3.4555
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.5929813  0.1758298   9.060  < 2e-16 ***
## P5210s162   -0.9746833  0.0236284 -41.251  < 2e-16 ***
## P61002       0.4221868  0.0431402   9.786  < 2e-16 ***
## P61003      -1.1101437  0.0267103 -41.562  < 2e-16 ***
## P6040       -0.0162665  0.0008129 -20.011  < 2e-16 ***
## P62402       0.5200007  0.0662628   7.848 4.24e-15 ***
## P62403      -2.7512209  0.0648725 -42.410  < 2e-16 ***
## P62404      -0.7941672  0.0313215 -25.355  < 2e-16 ***
## P62405      -2.0850455  0.2043389 -10.204  < 2e-16 ***
## P62406      -0.5852476  0.0443482 -13.197  < 2e-16 ***
## P61702      -2.3504711  0.0615326 -38.199  < 2e-16 ***
## P40202       0.8296848  0.1657374   5.006 5.56e-07 ***
## P40203       1.3439359  0.1784037   7.533 4.95e-14 ***
## P40204       1.4936954  0.1646385   9.073  < 2e-16 ***
## P40205       2.0698666  0.1986250  10.421  < 2e-16 ***
## P40206       2.3692712  0.1982843  11.949  < 2e-16 ***
## P40207       1.9019740  0.3763568   5.054 4.33e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 66700  on 54396  degrees of freedom
## Residual deviance: 52878  on 54380  degrees of freedom
## AIC: 52912
##
## Number of Fisher Scoring iterations: 5
```

Todos las variables tienen relación estadísticamente significativa con el nivel educativo. Para interpretar los coeficientes se transforman a posibilidades (odds).

```
## (Intercept)   P5210s162      P61002      P61003       P6040      P62402
##        4.92        0.38        1.53        0.33        0.98        1.68
##      P62403      P62404      P62405      P62406      P61702      P40202
##        0.06        0.45        0.12        0.56        0.10        2.29
##      P40203      P40204      P40205      P40206      P40207
##        3.83        4.45        7.92       10.69        6.70
```

- Las personas sin computadora para uso en el hogar (P5210s16) tienen 62% menos de posibilidades de tener nivel educativo alto que aquellos que si la tienen.
- Las personas con Régimen de Seguridad Social en Salud (RSSS) especial (P6100) tienen 1.53 veces más

de posibilidades de tener nivel educativo alto que aquellos que tiene RSSS contributivo. Y aquellas personas con RSSS subsidiado tienen 67% menos de posibilidades de tener nivel educativo alto que los que tienen RSSS contributivo.

- A medida que las personas envejecen (P6040) tienen 2% menos de posibilidades de tener un nivel educativo alto.
- Las personas que estaban buscando trabajo la semana pasada (P6240) tienen 68% más de posibilidades de tener nivel educativo alto que aquellos que estuvieron trabajando. Y las personas que ocuparon la mayor parte de la semana pasada en oficios del hogar, incapacitados para trabajar, estudiando u otras actividades tienen menos posibilidades de tener nivel educativo alto que aquellos que estuvierón trabajando.
- Las personas que no están asistiendo actualmente al colegio o universidad (P6170) tienen 90% menos de posibilidades de tener nivel educativo alto que aquellos que están asistiendo.
- Las personas que habitan en viviendas cuyo material predominante de los pisos (P4020) es cemento o madera o baldosín o mármol o madera púlida o alfombra tienen myaor posibilidades de tener nivel educativo alto que aquellos que habitan en viviendas con pisos de tierra o arena.

Conclusión

Se construyo una base de datos con la mayor cantidad de observaciones obtenidas a partir de las bases de datos del mes de abril de la Gran Encuesta Integrada de Hogares (GEIH) 2017. Esta base de datos contiene una gran cantidad de características generales de los individuos entrevistados, desde el régimen de seguridad social en salud, características de la vivienda donde habita, posesión de bienes como automóvil, computadora.

Es posible predecir a partir de dichas características el nivel educativo de las personas. Utilizando el algorítmo Bagging en Árboles de Clasificación se obtuvo el menor error de clasificación de prueba (0.1925) esto significa que el algorítmo tiene un nivel de precisión en el pronóstico igual a 80,75%. El algorítmo sufre de varianza (sobreajuste) el cuál se puede mejorar utilizando diferentes parámetros para estimar el Bagging (número de árboles) o utilizando técnicas para reducir el número de variables. Con el propósito de mejor la capacidad de predicción también se recomienda utilizar algorítmos como las redes neuronales o máquinas de soporte vectorial que permiten estimar hipótesis complejas.

Sin embargo un algorítmo con gran poder de predicción es dificil de interpretar ya que es difícil determinar cómo es la relación de las variables predictoras con la variable resultado. Por ello se decidio estimar algorítmos que a pesar de su bajo poder de predicción son sencillos de interpretar, ellos fueron: *Uno vs todos Regresión Logística Multivariada* y *Árboles de Clasificación*.

En ambos modelos se presenta un patrón: el nivel educativo es alto cuando las personas jóvenes están estudiando. Un rubro importante de esos 100 millones de dólares se sugieré invertirlos en fortalecer las políticas de permanencia educativa a nivel básico, medio y alto. Esto no significa que es sólo invertir en cobertura, edificios, y capital humano sino también en estrategias pedagógicas y currículos de asignaturas que hagan del estudio no una tediosa tarea sino una maravillosa experiencia donde se construyen y materializan sueños.

Tecnología empleada

El análisis de datos se realizó utilizando el software R. Y las librerias:

- *ggplot2*. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- *tree*. Brian Ripley (2016). tree: Classification and Regression Trees. R package version 1.0-37. https://CRAN.R-project.org/package=tree

- *randomForest*. A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18–22.
- *gbm*. Greg Ridgeway with contributions from others (2017). gbm: Generalized Boosted Regression Models. R package version 2.1.3. https://CRAN.R-project.org/package=gbm