

**O B L I G A T O R I O**

**Machine Learning No Supervisado**

Lic. En Gerencia y Administración

Docente: Ec. Damian Coltzau

Pioli Gastón – N° 166371

Machado Cecilia – N° 213640

Silveyra Andrés – N° 207336

Fecha de entrega: 18 de Noviembre de 2020

Índice

[1 – Introducción 3](#_Toc45559913)

[1.1 - Descripción de la base de datos 3](#_Toc45559914)

[2 – Parte 1 3](#_Toc45559915)

[2.1 - Análisis descriptivo 3](#_Toc45559916)

[2.1.1 Medidas de tendencia, dispersión y separación de las variables: 3](#_Toc45559917)

[2.1.2 - Gráficos 4](#_Toc45559918)

[2.1.3 Tabla de frecuencias de variable Ingreso 5](#_Toc45559919)

[2.1.4 Conclusiones finales del análisis descriptivo 6](#_Toc45559920)

[2.2 Análisis de correlación 7](#_Toc45559921)

[2.2.1 Tabla con coeficientes de correlación entre variables cuantitativas 7](#_Toc45559922)

[2.2.2 Gráficos de dispersión 8](#_Toc45559923)

[2.2.3 Conclusiones finales del análisis de correlación 8](#_Toc45559924)

[2.3 – Análisis de la variable categórica 9](#_Toc45559925)

[2.3.1 - Tabla de frecuencias absolutas relacionando Ingreso y NivelS.E 9](#_Toc45559926)

[2.3.2 - Tabla de frecuencias porcentuales relacionando Ingreso y NivelS.E 9](#_Toc45559927)

[2.3.3 - Tabla de frecuencias absolutas para Nivel S.E Alto y las que no son de Nivel S.E Alto 9](#_Toc45559928)

[2.3.4 Diagramas de cajas 10](#_Toc45559929)

[2.3.5 Histogramas 10](#_Toc45559930)

[2.3.6 - Conclusiones finales de los datos obtenidos 11](#_Toc45559931)

[2.4 – Regresión Lineal 12](#_Toc45559932)

[2.4.1 - Comentarios reg1: 12](#_Toc45559933)

[2.4.2 - Comentarios Reg2 14](#_Toc45559934)

[2.4.3 Interpretación y conclusiones del modelo. 15](#_Toc45559935)

[2.4.4 - Conclusiones finales de Regresión Lineal: 16](#_Toc45559936)

[3 - Parte 2 17](#_Toc45559937)

[3.1 Regresión Logística 17](#_Toc45559938)

[3.2 Modelo de Árbol de clasificación 23](#_Toc45559939)

[3.3 Conclusión sobre los resultados y selección del mejor modelo 27](#_Toc45559940)

[4 – Parte 3 28](#_Toc45559941)

[4.1 Modelo K-MEANS 28](#_Toc45559942)

[4.1.1 - Conclusión final: 31](#_Toc45559943)

[5 – Script 32](#_Toc45559944)

# 1 – Introducción

## - Descripción de la base de datos

La base de datos utilizada en el siguiente informe se llama “*base.cvs*”. La muestra tiene información sobre los integrantes de un hogar promedio en Montevideo y en la investigación realizada se obtuvo una muestra de 150 observaciones con 6 variables distintas. Estas son:

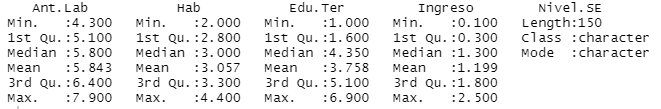
* *ID*: Identificador único de las observaciones.
* *Ant.Lab*: Antigüedad laboral promedio de los hogares.
* *Hab:* Cantidad de habitantes promedio de los hogares.
* *Edu.Ter:* Cantidad de años de Educación Terciaria en el hogar.
* *Ingreso:* Ingreso promedio de los hogares (valores cada 10.000 pesos)
* *Nivel.SE:* Nivel Socioeconómico, en tres categorías: Alto, Medio, Bajo.

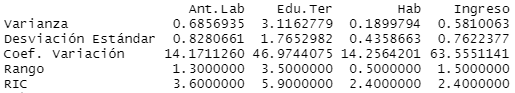
# 2 – Parte 1

El objetivo de la primer parte del presente informe es comprobar si es posible predecir valores de la variable Ingreso mediante el análisis de estimaciones puntuales y elaboración de modelos estadísticos*.*

## - Análisis descriptivo

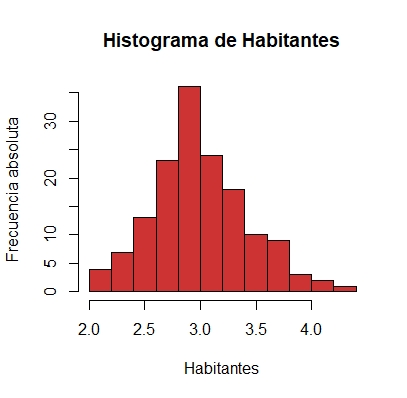
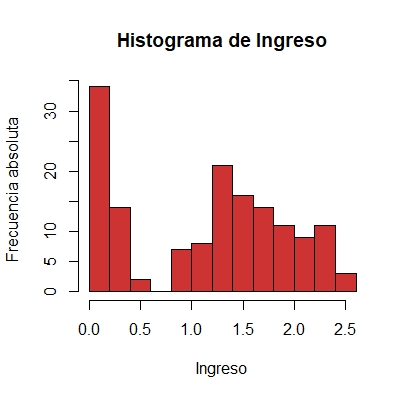
### 2.1.1 Medidas de tendencia, dispersión y separación de las variables:

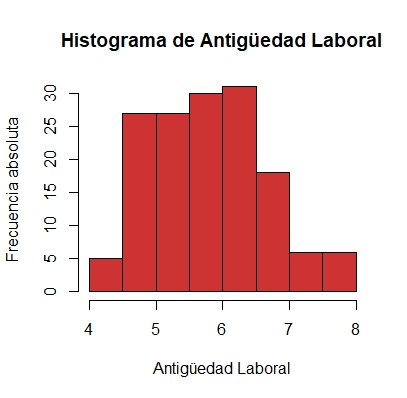




### 2.1.2 - Gráficos

#### 2.1.2.1 - Histogramas





#### 2.1.2.2 – Diagrama de caja

### 2.1.3 Tabla de frecuencias de variable Ingreso



### 2.1.4 Conclusiones finales del análisis descriptivo

A continuación, se detallan observaciones generales de cada variable con respecto a los datos y gráficos obtenidos en los puntos anteriores.

La Antigüedad laboral promedio de la muestra es de 5,843 años. Posee una leve distribución asimétrica con sesgo hacia la derecha, esto confirma que el valor de la media se encuentre a la derecha de la mediana (5,8).

La mayor diferencia de antigüedad laboral de la muestra es de 1,3 años, asimismo, el 50% central de los datos tiene una diferencia máxima de 3,6 años.

En cuanto a la dispersión de los datos con respecto a la media se afirma que es baja debido al valor del coeficiente de variación (14,17%). Esta tendencia también se puede observar en el histograma al ver que la frecuencia absoluta de los intervalos está distribuida uniformemente. No obstante, al ver el diagrama de caja de la Antigüedad laboral, se constata que poseer su límite inferior y superior lejos de la media, ocasiona la dispersión ya mencionada, aunque no tenga valores atípicos.

La educación terciaria promedio de la muestra es de 3,758 años. Posee una distribución asimétrica con sesgo hacia la izquierda, lo que confirma que el valor de la media se encuentre a la izquierda de la mediana (4,35).

La mayor diferencia de educación terciaria de la muestra es de 3,5 años, asimismo, el 50% central de los datos tiene una diferencia máxima de 5,9 años.

En cuanto a la dispersión de los datos con respecto a la media se afirma que es media debido al valor del coeficiente de variación (46,97%). Al observar el histograma se constata dicha dispersión y sesgo debido a la presencia de una concentración alta de frecuencias absolutas en el primer intervalo del gráfico. El diagrama de caja también verifica una mayor concentración de los datos por debajo de la media, observando que hay más observaciones del lado del límite inferior. En otras palabras, hay más personas con bajo nivel educativo con respecto al promedio de la muestra.

La cantidad promedio de habitantes en los 150 hogares es de 3,057 habitantes. Posee una distribución casi normal con un muy leve sesgo hacia la derecha, esto lo confirma la baja diferencia que hay entre la media y la mediana (3,0).

La mayor diferencia de cantidad de habitantes por hogar es de menos de una persona (exactamente 0,5), asimismo, el 50% central de los datos tiene una diferencia máxima de 2,4 personas.

En cuanto a la dispersión de los datos con respecto a la media se afirma que es baja debido al valor del coeficiente de variación (14,26%). Al observar el histograma se constata dicha dispersión debido a la alta concentración de frecuencias absolutas en los intervalos centrales del gráfico, así como también la baja concentración de observaciones en los intervalos extremos. El diagrama de caja de Habitantes constata lo dicho anteriormente, observando que hay cantidades uniformes de observaciones en torno a la media de la muestra. Y también que existe un leve sesgo debido a la existencia de valores atípicos evidenciados fuera de ambos límites.

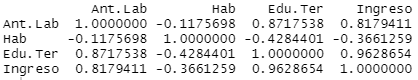
El Ingreso promedio de las observaciones del dataset es de $11.990. En cuanto a la mayor diferencia de Ingreso la misma es de $15.000, asimismo, el 50% central de los datos tiene una diferencia máxima de $24.000.

Posee una distribución muy similar a la variable Educación terciaria, es por esto que los histogramas y diagramas de caja también coinciden en su representación. De todas formas, la variable Ingreso tiene más personas representadas por encima de la media, es por esto que su coeficiente de variación es mayor (63,56%).

Si se observa la tabla de frecuencias de la variable Ingreso, también se ve que, si se divide a la variable en 3 partes iguales, el intervalo central (de $9.000 a $17.000) es el que tiene mayor representación con un 37,24% de los 150 ingresos observados. No obstante, las apariciones de algunos ingresos altos ratifican que haya una leve mayor concentración en el tercio superior con una representación de un 31,72%, mientras que el tercio inferior es representado por un 31,03%.

## Análisis de correlación

### 2.2.1 Tabla con coeficientes de correlación entre variables cuantitativas



### 2.2.2 Gráficos de dispersión

### 2.2.3 Conclusiones finales del análisis de correlación

A continuación, se detallan observaciones generales de la relación que hay entre la variable Ingreso y las restantes variables cuantitativas.

En la muestra de observaciones hay una correlación lineal positiva entre la variable Ingreso y Antigüedad Laboral, la misma está representada en el gráfico de dispersión con una recta de pendiente positiva que pasa por el centro de los datos. Esta relación también es comprobada con el valor obtenido en la tabla de correlación, el coeficiente tiene un valor de 0,818.

En otras palabras, es probable que una persona que tenga más años de Antigüedad laboral obtenga un Ingreso mayor con respecto a una persona que tenga menor Antigüedad laboral.

Las variables Ingreso y Educación Terciaria también tienen una correlación lineal positiva, pero es más intensa que la de Ingreso y Antigüedad Laboral, ya que los datos representados en el gráfico de dispersión se ubican en coordenadas muy similares y están más alineados con la recta central. El coeficiente de correlación obtenido de la tabla de correlación es de 0,963.

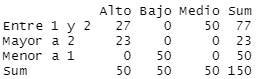
En resumen, cuantos más años de educación terciaria tenga una persona la probabilidad de que aumente su Ingreso es alta.

En lo que respecta a la relación entre Ingreso y Habitantes, se considera que hay puntos dispuestos alrededor de una recta, es decir, hay correlación lineal, pero la pendiente de dicha recta es negativa. Asimismo, se observa que el coeficiente de la pendiente de la recta es bajo debido a que hay una gran dispersión de las observaciones. Si se observa la tabla de coeficientes de correlación, el valor para la dicha relación es de -0,366, confirmando también lo comentado anteriormente.

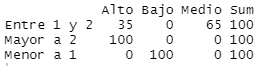
Por lo cual, se puede explicar que si hay más habitantes promedio en un hogar el Ingreso del mismo va a ser menor que el de un hogar que tiene menos habitantes.

## 2.3 – Análisis de la variable categórica

### 2.3.1 - Tabla de frecuencias absolutas relacionando Ingreso y NivelS.E



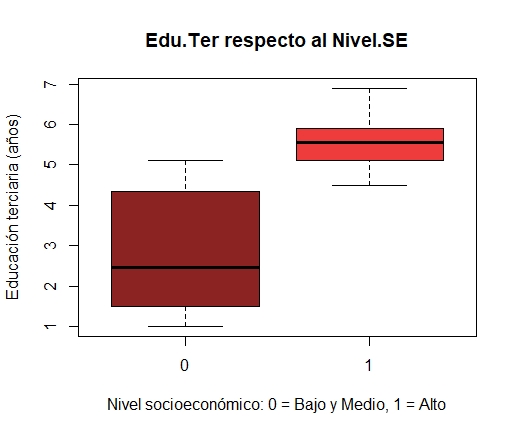
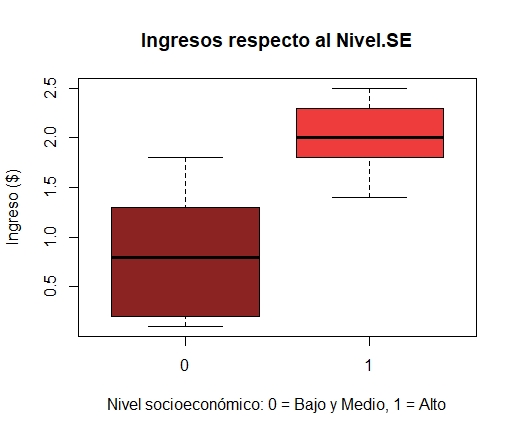
### 2.3.2 - Tabla de frecuencias porcentuales relacionando Ingreso y NivelS.E



## 2.3.3 - Tabla de frecuencias absolutas para Nivel S.E Alto y las que no son de Nivel S.E Alto



### 2.3.4 Diagramas de cajas



### 2.3.5 Histogramas

### 2.3.6 - Conclusiones finales de los datos obtenidos

A continuación, se realizan observaciones de los datos obtenidos en las tablas y los gráficos representados en los puntos anteriores.

Cuando se contrastan las variables Ingreso y Nivel socioeconómico, se observa que el 100% de las personas que tienen un Ingreso mayor a $20.000 pertenecen al Nivel socioeconómico Alto, y que el 100% de las personas que perciben un Ingreso menor a $10.000 pertenecen al Nivel socioeconómico Bajo. En cambio, cuando los ingresos son entre $10.000 y $20.000, el 35% son de Nivel socioeconómico Alto y el restante 65% son de Nivel socioeconómico Medio. Es decir, que existe más probabilidad de que las personas pertenezcan a un Nivel socioeconómico Medio o Bajo que al Alto.

En lo que respecta a la distribución de los datos, hay una mayor concentración de frecuencias dentro del grupo que es de Nivel socioeconómico Medio o Bajo ya que representan a 100 observaciones de las 150 del dataset. Asimismo, dentro dicho grupo también se constata la proporción mencionada anteriormente observando las áreas de los diagramas de caja de Ingreso respecto al Nivel.SE. Al ver los histogramas se visualiza la existencia de una asimetría hacia la derecha, lo cual confirma que haya más observaciones por encima de la media.

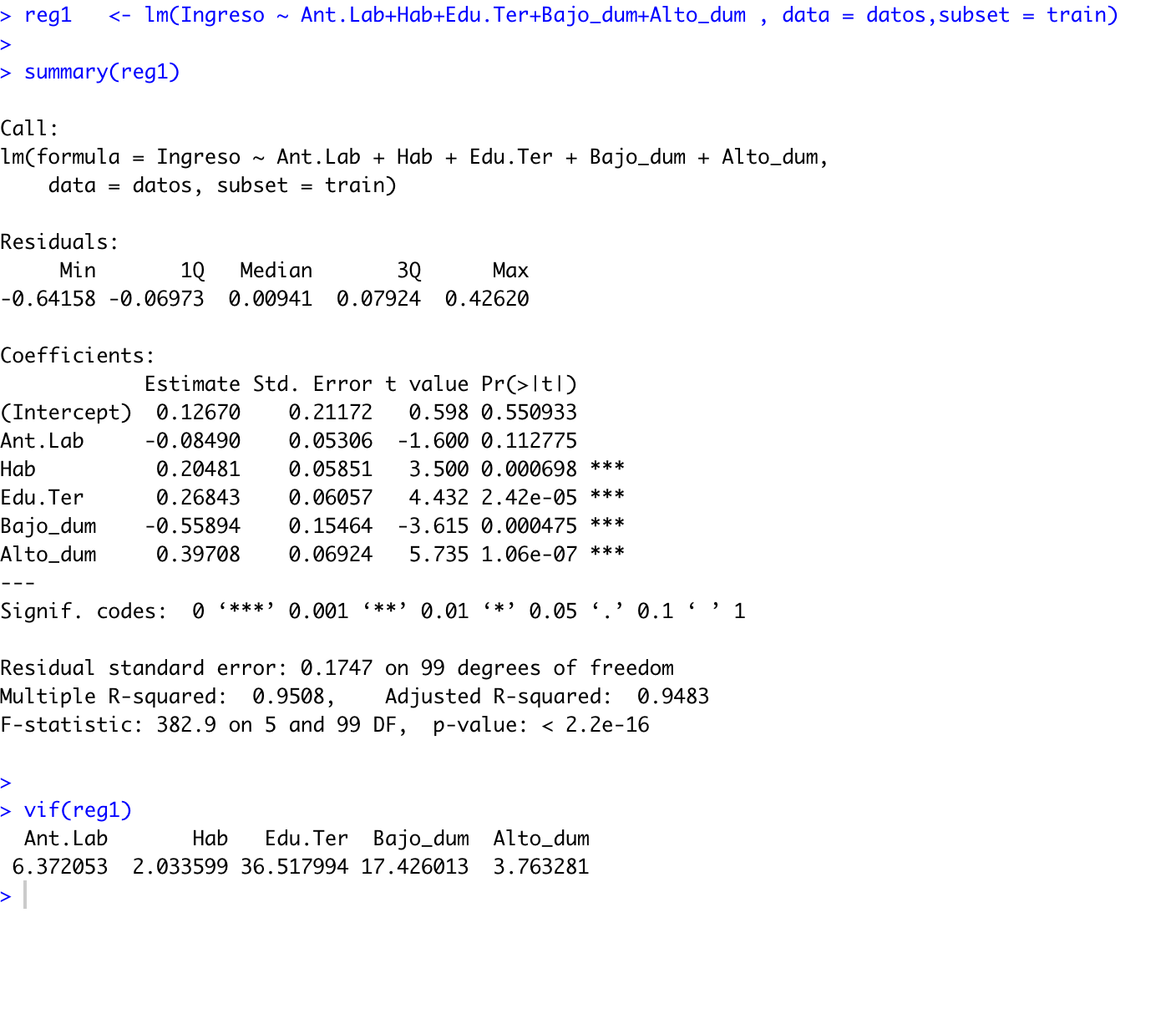
En cuanto al contraste de la variable Educación terciaria y Nivel socioeconómico se observa que hay un comportamiento similar al análisis anterior con respecto a la distribución de los datos. De todas formas, se constata mediante la observación del área del diagrama de caja que hay una mayor ausencia de frecuencias absolutas dentro del Nivel socioeconómico Alto con respecto al Ingreso. Es decir, que una persona que tiene un Ingreso y Nivel socioeconómico Alto, no siempre tiene un nivel de Educación terciaria alta.

## 2.4 – Regresión Lineal

Dada la ciencia en la que estamos abordando el análisis decidimos trabajar con un nivel de α=0,10.

Con la finalidad de observar el modelo con todas las variables del dataset decidimos correr la siguiente regresión la cual denominamos *reg1*.

### 2.4.1 - Comentarios reg1:



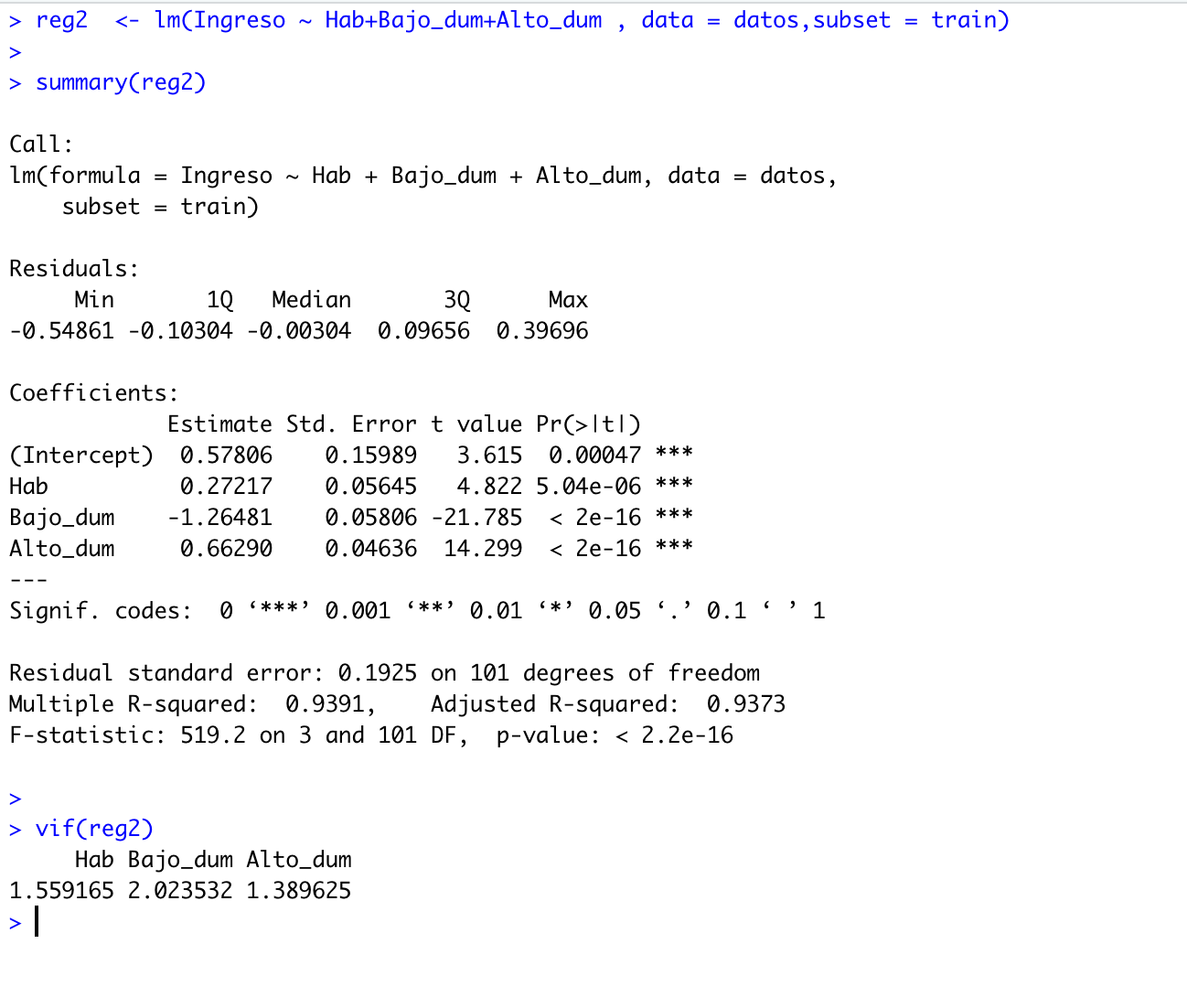
* Prueba F-statistic: Esta prueba de hipótesis nos plantea si los coeficientes de la regresión son 0, la cual rechaza y nos indica que al menos un coeficiente es distinto de 0. Como vemos en la salida de RStudio, todos los coeficientes de las variables explicativas son ≠0
* Variables no significativas a un nivel de α=0,10.
* Problemas de multicolinealidad[[1]](#footnote-1).

Variables predictoras: Ant.Lab, Edu.Ter y Bajo\_dum con VIF > 5.

Dado los comentarios de Reg.1 y lo ineficiente[[2]](#footnote-2) que sería correr manualmente otras regresiones hasta llegar a una versión óptima, decidimos llegar a la regresión final aplicando un método automático de selección de atributos: Backward selection.

Aplicar el método automático en RSTUDIO nos otorga exactamente el mismo modelo que reg1.

Dada esta situación generamos la regresión número 2, llamada reg2 en la cual descartamos aquellas variables no significativas a un α=0,10, las mismas son: Ant.Lab y Edu.Ter.



### 2.4.2 - Comentarios Reg2

Análisis Summary Reg2

* Sintaxis

Y= β0 + β1.X1+ β2.X2 + β3.X3 + β4. X4

* Función

Ingreso = 0,57806 + 0,27217 ∗ Hab -1,26481 ∗ Bajo\_dum + 0,66290 ∗Alto\_dum

* Prueba F-statistic: El estadístico F tomó un valor cercano a 0, esto nos indica que existe relación entre la respuesta y los predictores.
* t value s. Nos muestra los valores del estadístico (t-student) para cada variable explicativa la cual nos sirve como dato para la prueba de hipótesis.
* Pr(>| t |)

Los pvalues individuales son menores a 0,10 la cual nos indica que todas las variables predictoras están relacionados a nuestra variable Y=Ingreso.

Esto indica que todas las variables son significativas a un nivel de α=0,10.

* R2 reg1 0,9508 Vs. R2 reg2= 0,9391.

Habiendo quitado 2 variables explicativas no se advierte una diferencia significativa, es decir la bondad de ajuste de un modelo con 5 variables difiere muy poco respecto al modelo con 3 variables.

Concluimos que es aceptable perder bondad de ajuste para que en nuestro modelo todas nuestras variables sean significativas.

* R2 ajustado reg1 =0,9483 Vs. R2 ajustado reg2= 0,9373

Este dato sirve para comparar dos modelos y observamos que tampoco presenta una diferencia significativa.

* Sin problemas de multicolinealidad.

En todas las variables predictoras VIF < 5

* ECM en Train = 0.03564802

ECM en Test = 0.02409314

El error cuadrático medio disminuye en test, es decir el modelo tiene menos error en la base de test.

* r2  Es la proporción de la variabilidad de la variable explicada que el modelo logra explicar.

r2 en Train = 0.9391004

r2 en Test = 0.9594645

r2 en Test mejora respecto a la base de Train, con este dato podríamos concluir que nuestro modelo no presenta problemas de overfitting.

### 2.4.3 Interpretación y conclusiones del modelo.

Ejemplo de predicción

* ¿Cuál es el ingreso de una persona con una cantidad de 3 Habitantes, y un NSE Alto?

Ingreso = 0,57806 + 0,27217 ∗ Hab -1,26481 ∗ Bajo\_dum + 0,66290 ∗Alto\_dum

Ingreso= 0.57806 + 0.27217\*(3) -1.26481\*(0) + 0.66290\*(1)

Ingreso= 2.05747 x (10.000) = $20.574,7

Interpretaciones

* Alto\_dum=1

Ingreso = 0,57806 + 0,27217∗ Hab -1,26481∗ Bajo\_dum + 0,66290 ∗(1)

Una persona de NSE alto tiene + 0.66290 de ingreso respecto a una persona de NSE medio dejando todo lo demás constante.

* Variable Hab = 0,27217\*Hab

El aumento en una unidad en la variable Hab, dejando las demás variables constantes provoca un aumento en el ingreso.

### 2.4.4 - Conclusiones finales de Regresión Lineal:

El modelo de regresión lineal generado apela al principio de parsimonia, el cual hace relación a que un modelo sencillo como este, puede explicar la realidad relativamente bien.

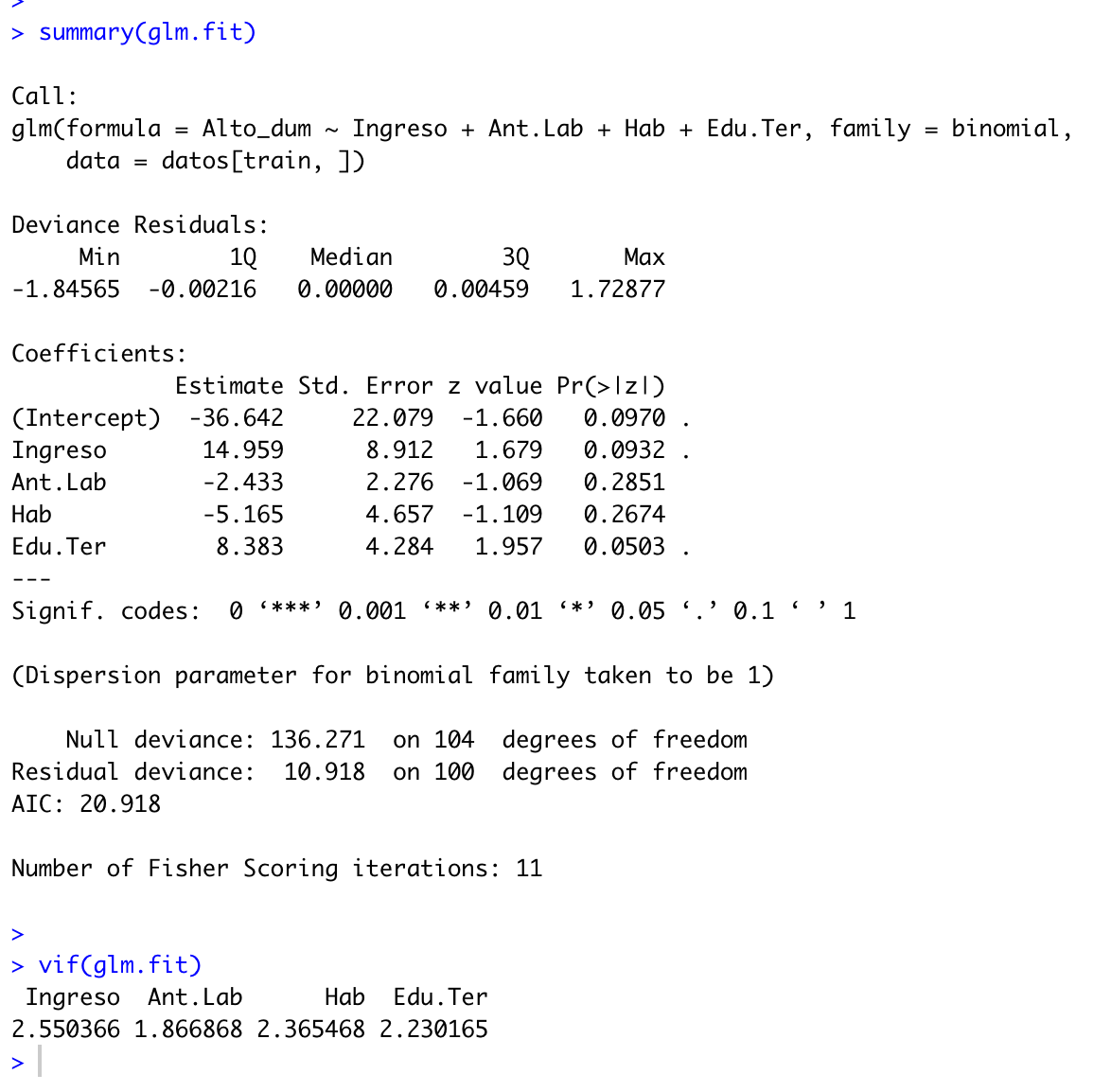
Esto es consecuencia de:

* La bondad de ajuste que presenta nuestro modelo con un r2 en nuestra base de test de 0,96 aproximadamente.
* El uso de 3 variables explicativas.

# 3 - Parte 2

## 3.1 Regresión Logística

Con la finalidad de observar el modelo con todas las variables del dataset decidimos correr la siguiente regresión la cual denominamos, glm.fit.

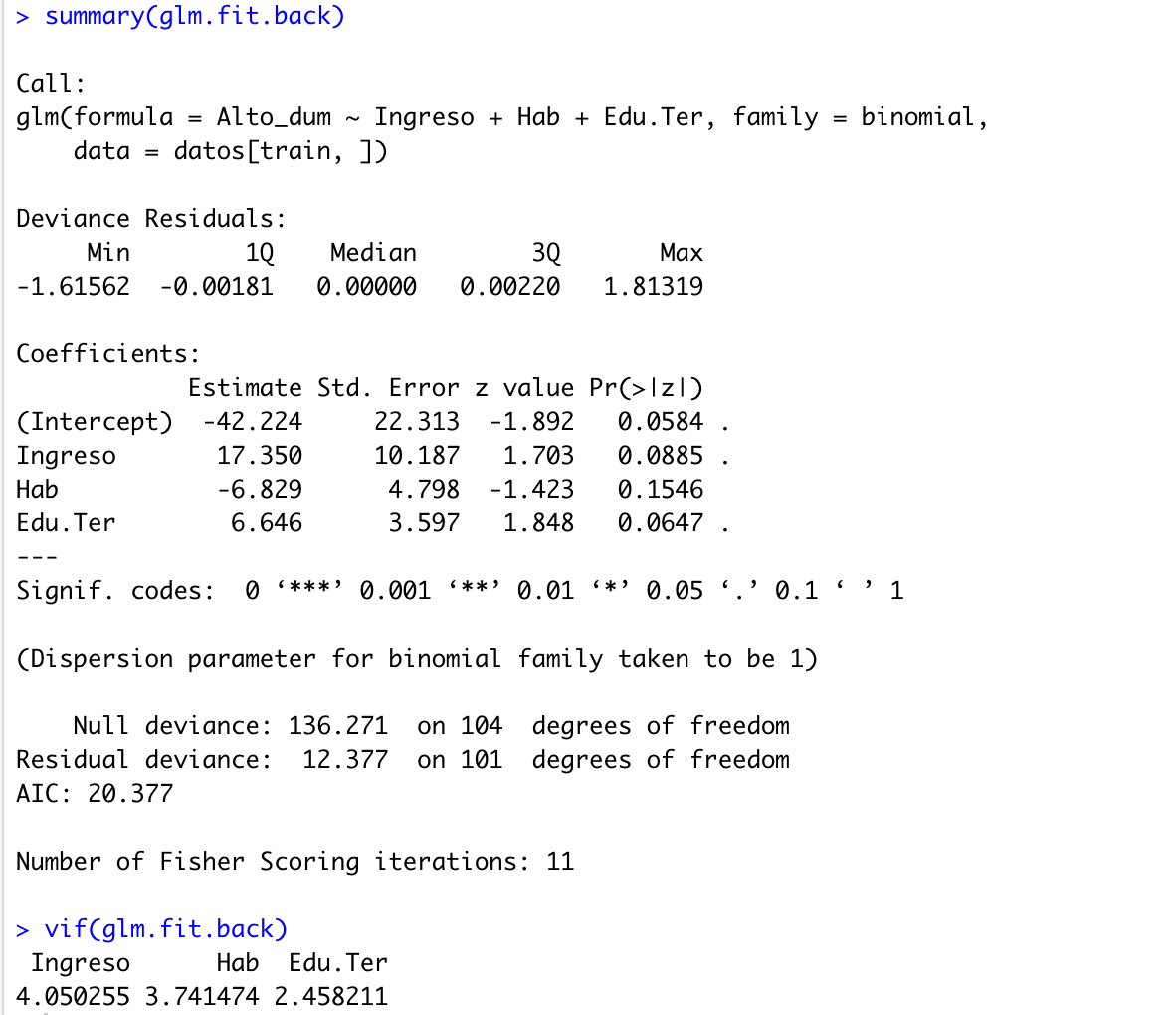


Comentarios Glm.fit:

* Variables no significativas a un nivel de α=0,10.
* Sin problemas de multicolinealidad.

En todas las variables predictoras VIF < 5

Por las mimas razones que en regresión lineal, corremos método automático de selección de atributos: Backward selection la cual nos otorga la siguiente regresión logística, denominada: glm.fit.back

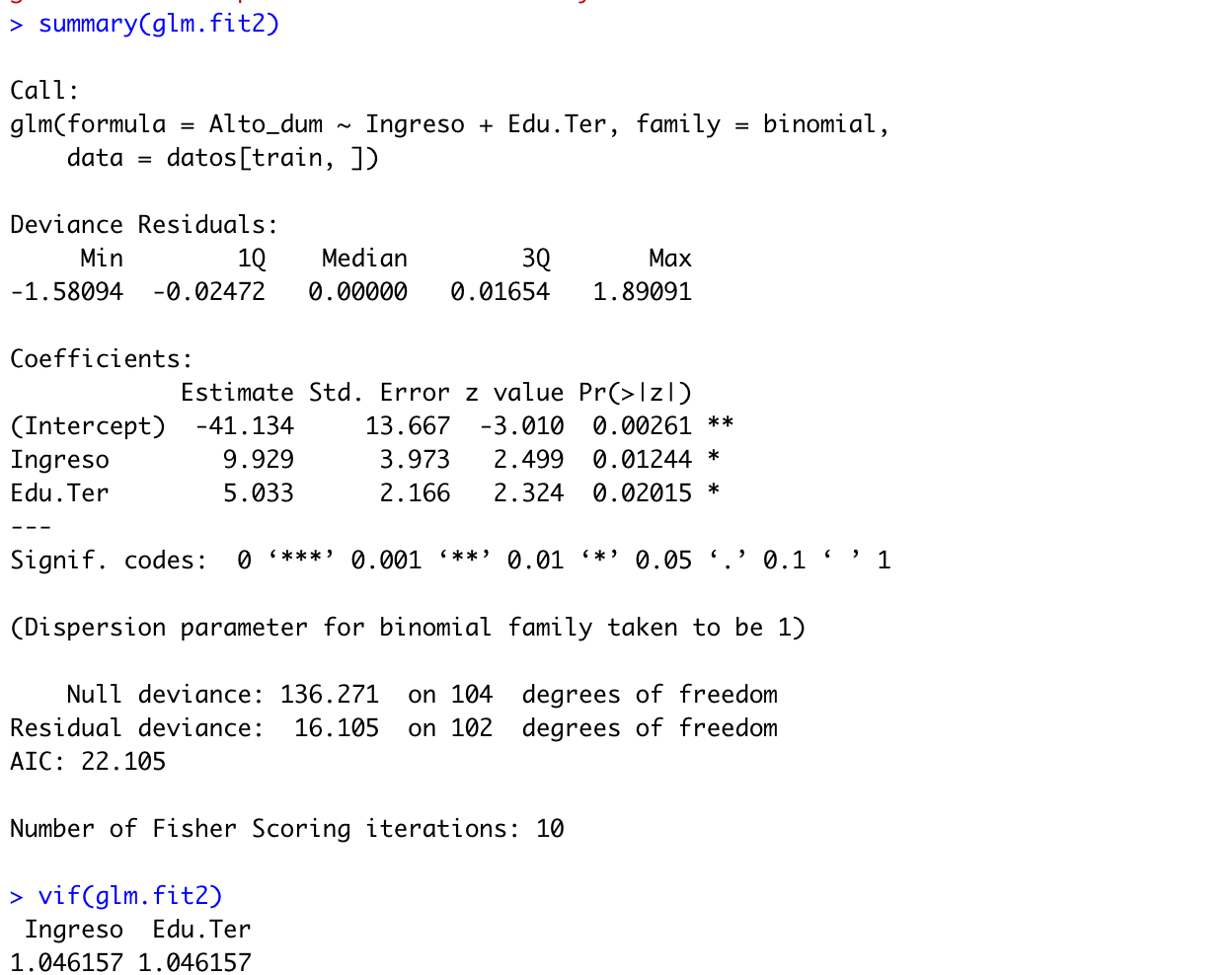


Comentarios Glm.fit.back:

* Variables Hab no significativas a un nivel de α=0,10.
* Sin problemas de multicolinealidad.

En todas las variables predictoras VIF < 5

Dado que tenemos una variable predictora, Hab no significativa a un nivel de α=0,10 decidimos quitarla y realizar una nueva regresión logística la cual denominamos: glm.fit2



Comentarios Glm.fit2:

* Todas las variables predictoras son significativas a un nivel de α=0,10.
* Sin problemas de multicolinealidad.

En todas las variables predictoras VIF < 5

* Sintaxis
* Función
* Null deviance: Nos indica cuanto nuestro modelo está explicando.
* Residual deviance: Nos indica cuanto nuestro modelo no está explicando.

Observamos que es relativamente poco lo que no puede explicar nuestro modelo.

Si comparamos glm.fit y glm.fit2 la variación de Residual Deviance es relativamente pequeña, pasó de 10,918 a 16,105.

Dicho en otras palabras, sacrificamos capacidad explicativa para que nuestro modelo trabaje con menos variables, es decir, le quitamos complejidad.

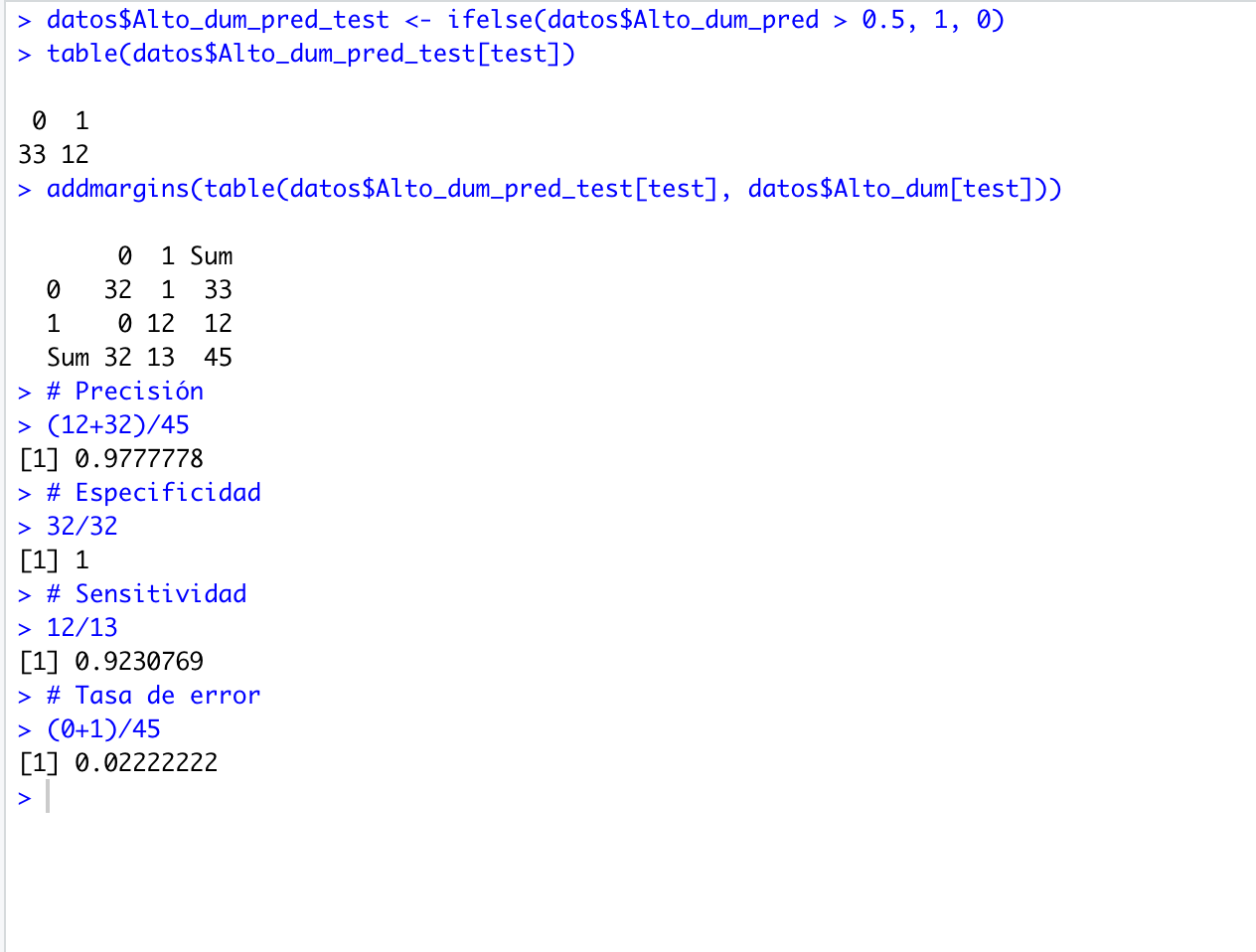
Una vez llegado al nuestro modelo optimo realizamos la predicción en nuestra base de test.

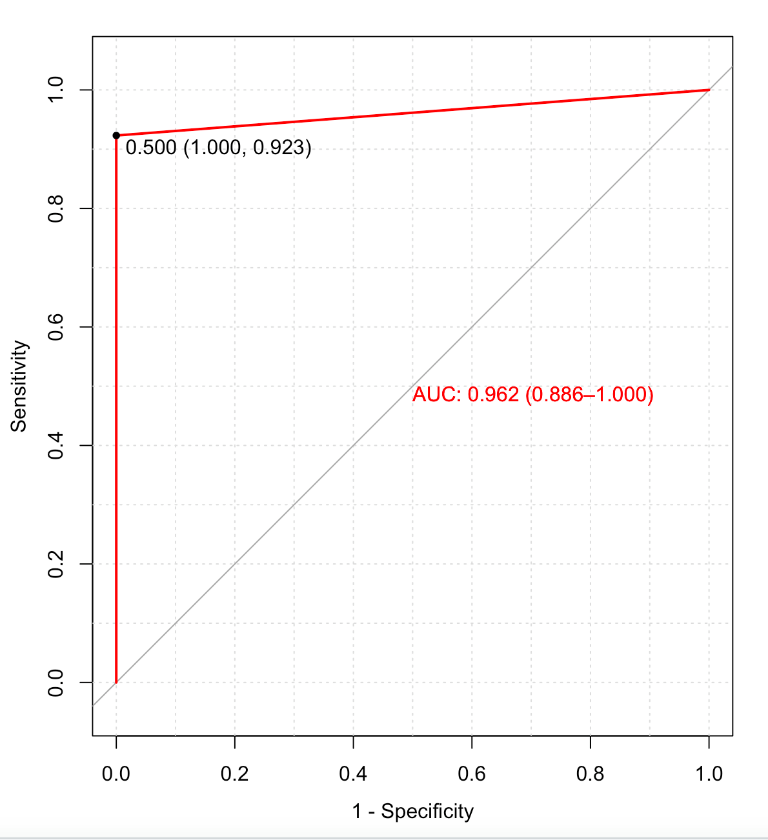
Interpretación:

En el modelo de regresión logística, los efectos de las variables explicativas sobre la variable dependiente no son lineales, pero podemos concluir lo siguiente:

Como el signo es positivo el aumento de nuestra X1, Ingreso, se asociará a un aumento de que la probabilidad suba.

Pasa lo mismo con nuestra X2, Edu.Ter.

Evaluación del modelo en Test.

Curva ROC – Punto de corte 0,5

Comentarios Finales:

Bondad de ajuste:

Especificidad= 1

Es la probabilidad de predecir un fracaso entre los fracasos.

Sensitividad = 0,9230769

Es la probabilidad de predecir un éxito entre los éxitos.

Observamos que contamos con un modelo relativamente bueno ya que la Especificidad es igual a uno y la sensitividad es muy cercana a 1.

Otro dato importante es el AUC= 0,962 que presenta nuestra curva ROC, considerando que un modelo perfecto tendría que tener un área bajo la curva de 1.

## 3.2 Modelo de Árbol de clasificación

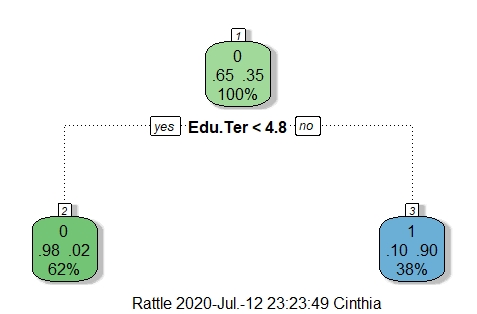
El objetivo de la elaboración del modelo de árbol es predecir cuándo un hogar es de nivel socioeconómico Alto.

Para ello, luego de realizar el correspondiente preámbulo, creamos la variable dummy que deseamos predecir: Alto\_dum para la variable categórica Nivel S.E.

Luego creamos el dataset de training y testing, eliminando las variables individuales ID y Nivel S.E, teniendo en cuenta que la variable ID no aporta ningún dato relevante y la variable Nivel Socioeconómico la que deseamos predecir.

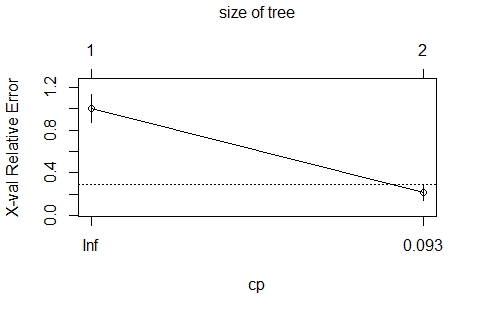
Posteriormente separamos el dataset en 70% de los datos para train y el 30% restante para test, con el fin de lograr un modelo de árbol de clasificación comparable con el modelo de regresión logística.

A continuación, estimamos nuestro primer modelo en training, bajo el método "class" correspondiente a clasificación, obteniendo el siguiente gráfico del árbol de clasificación:



De dicho árbol podemos concluir que el 38% de los hogares de la muestra tienen un nivel de educación terciaria mayor a 4,8 años y pertenecen al nivel socioeconómico alto, con un 10% de probabilidad de error. El 62 % de los hogares de la muestra presentan una educación terciaria menor a 4,8 años y pertenecen a un nivel socioeconómico bajo o medio, con un 2% de error.

Graficamos y analizamos la performance del modelo vs la complejidad:



Del gráfico concluimos que el mejor modelo de árbol debe contener 2 nodos,

cp= 0,093 ya que presente una tasa de error similar al mejor error, por lo tanto, comprendemos que nuestro árbol no debe continuar podándose.

Determinamos las reglas inducidas por el árbol:

Rule number: 3 [Alto\_dum=1 cover=40 (38%) prob=0.90]

Edu.Ter>=4.75

Rule number: 2 [Alto\_dum=0 cover=65 (62%) prob=0.02]

Edu.Ter< 4.75

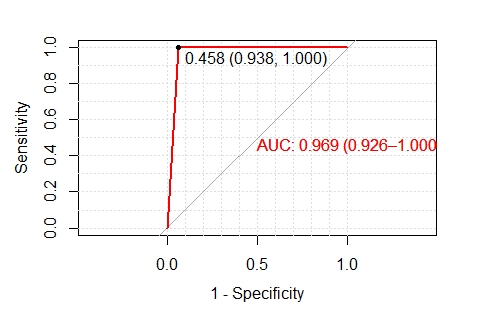
Las reglas extraídas del árbol inicial concuerdan con la lectura realizada del mismo.

El 38% del total de observaciones en train, correspondiente a 40 observaciones, tienen una educación terciaria mayor a 4,75 años y representan hogares con NSE Alto, con un error del 10%.

El 62% del total de observaciones en train, correspondiente a 65 observaciones, tienen una educación terciaria menor a 4,75 años y no son hogares con NSE Alto, con un error del 2%

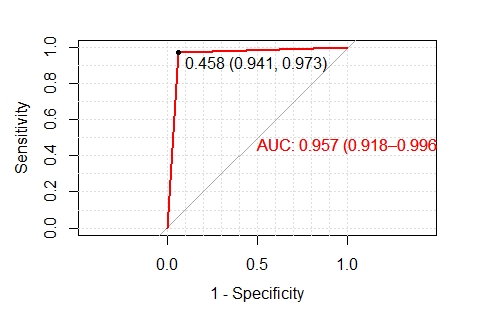
Teniendo nuestro árbol definido, procedemos a realizar la predicción del modelo sobre todo el dataset, graficamos la curva ROC en test y en train y hallamos las áreas debajo de sus respectivas curvas con el fin de evaluar la bondad de ajuste.

Curva ROC en Test:



El valor es aceptable, el modelo tiene buena bondad de ajuste.

Curva ROC en Train:



Comparando las AUC obtenidas en train y test, observamos que no hay indicio de overfitting, no habría sobreajuste de datos.

Realizamos la matriz de confusión utilizando como punto de corte la probabilidad 0,5 y obtenemos los siguientes resultados:

|  |  |  |  |
| --- | --- | --- | --- |
|  | 0 | 1 | Sum |
| FALSE | 30 | 0 | 30 |
| TRUE | 2 | 13 | 15 |
| Sum | 32 | 13 | 45 |

De acuerdo a la tabla de confusión obtenida observamos que:

El modelo predijo que 30 hogares no tenían NSE alto, y efectivamente no lo tenían.

El modelo predijo correctamente los hogares que no tenían NSE alto, no se encuentran valores falsos negativos.

El modelo predijo que 13 hogares tenían NSE alto, y efectivamente lo tenían.

El modelo predijo que 2 hogares tenían NSE alto, pero en realidad no lo tenían.

A partir de la tabla realizamos los cálculos correspondientes:

Precisión = (13+30)/45 = 0,9555556

Sensibilidad = 13/13 = 1

El modelo logra un valor perfecto de sensibilidad, lo cual nos indica la excelente capacidad del modelo de estimar hogares de Nivel Socioeconómico alto a los hogares que realmente cumplen con esta condición, la proporción de hogares con alto nivel socioeconómico está correctamente identificada.

Especificidad = 30/32 = 0,9375

El modelo logra una alta especificidad, lo cual nos indica la buena capacidad del modelo de estimar hogares que no tienen Nivel Socioeconómico Alto a los hogares que realmente tienen niveles SE inferiores; proporción de niveles socioeconómicos inferiores correctamente identificados.

Tasa de error = (0+2)/45 = 0,04444444

Se encuentra una tasa de error de clasificación minimizada.

## 3.3 Conclusión sobre los resultados y selección del mejor modelo

Para llevar adelante la selección del mejor modelo entre Regresión Logística vs. Árbol de clasificación, no deseamos realizarlo de manera absoluta ya que entendemos que podría haber causas que relativicen nuestra decisión. Estas radican en la explicación estadística-matemática que tiene cada algoritmo de fondo y a quien se la tengamos que explicar.

Comentado esto, realizamos la selección del mejor modelo considerando dos dimensiones:

1. Complejidad del algoritmo: Independiente del resultado en la bondad de ajuste es más sencillo explicar el modelo de Clasificación de Árboles.
2. Bondad de ajuste de cada modelo.

2.a) Bondad de ajuste la Regresión Logística

AUC Curva Roc Regresión Logística Test = 0,962

Especificidad= 1

Sensitividad = 0,9230769

Tasa de error= 0,02222222

2.b) Bondad de ajuste la Clasificación de Arboles

AUC Curva Roc Clasificación de Arboles Test = 0,969

Especificidad= 0,9375

Sensitividad = 1

Tasa de error = 0,04444444

Si bien el modelo de Regresión Logística presenta menos tasa de error y mayor especificidad, el modelo de Clasificación de Arboles presenta mayor AUC y mejor sensitividad.

Realizando un trade-off, entre la complejidad de modelo y bondad de ajuste nos parece lo más acertado seleccionar el modelo de Clasificación de Árboles.

# 4 – Parte 3

## 4.1 Modelo K-MEANS

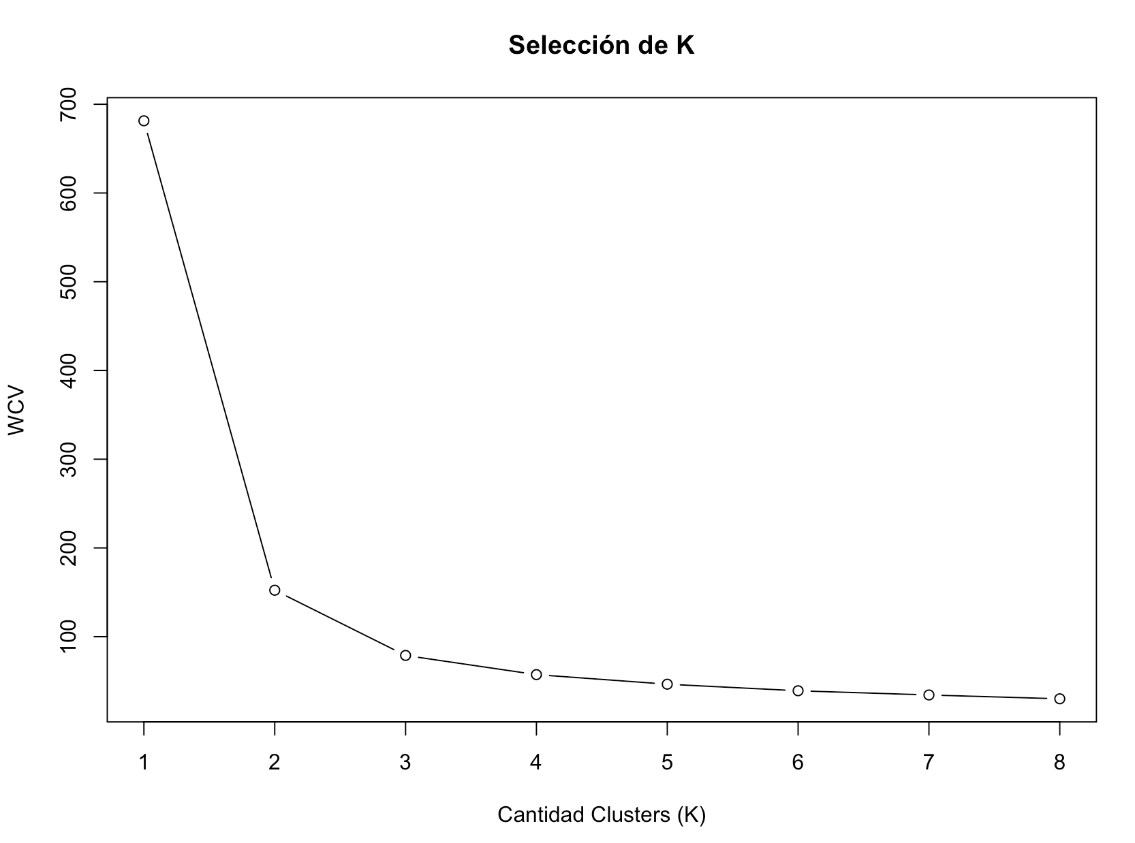
* Estandarización

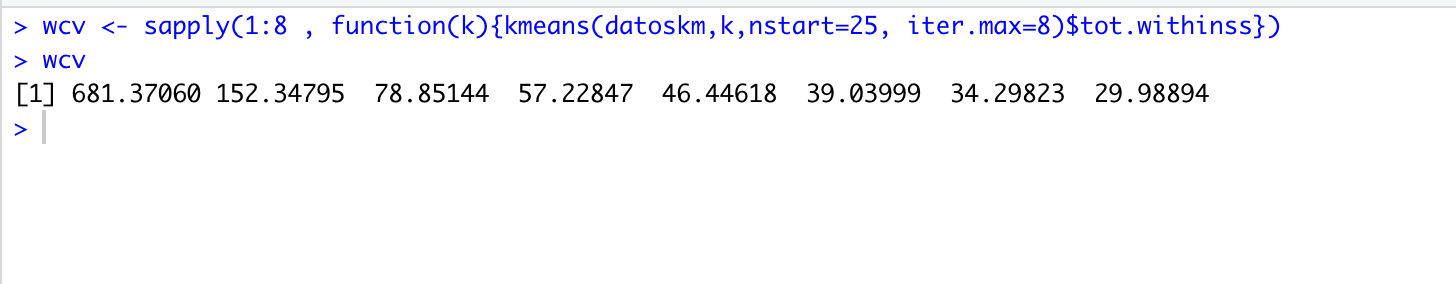
En este caso observamos que todas nuestras variables presentan diferente unidad de medida y escalas. Dada esta situación estandarizamos las variables para evitar la influencia de la unidad de medida.

* Selección de K (contexto y WCV)

Contexto, dada la ciencia en la que estamos trabajando, en este caso el desarrollo de un modelo que clasifique el NSE, intuimos que una discriminación básica para esta variable derivara en al menos 3 conjuntos: NSE Alto, Medio y Bajo.

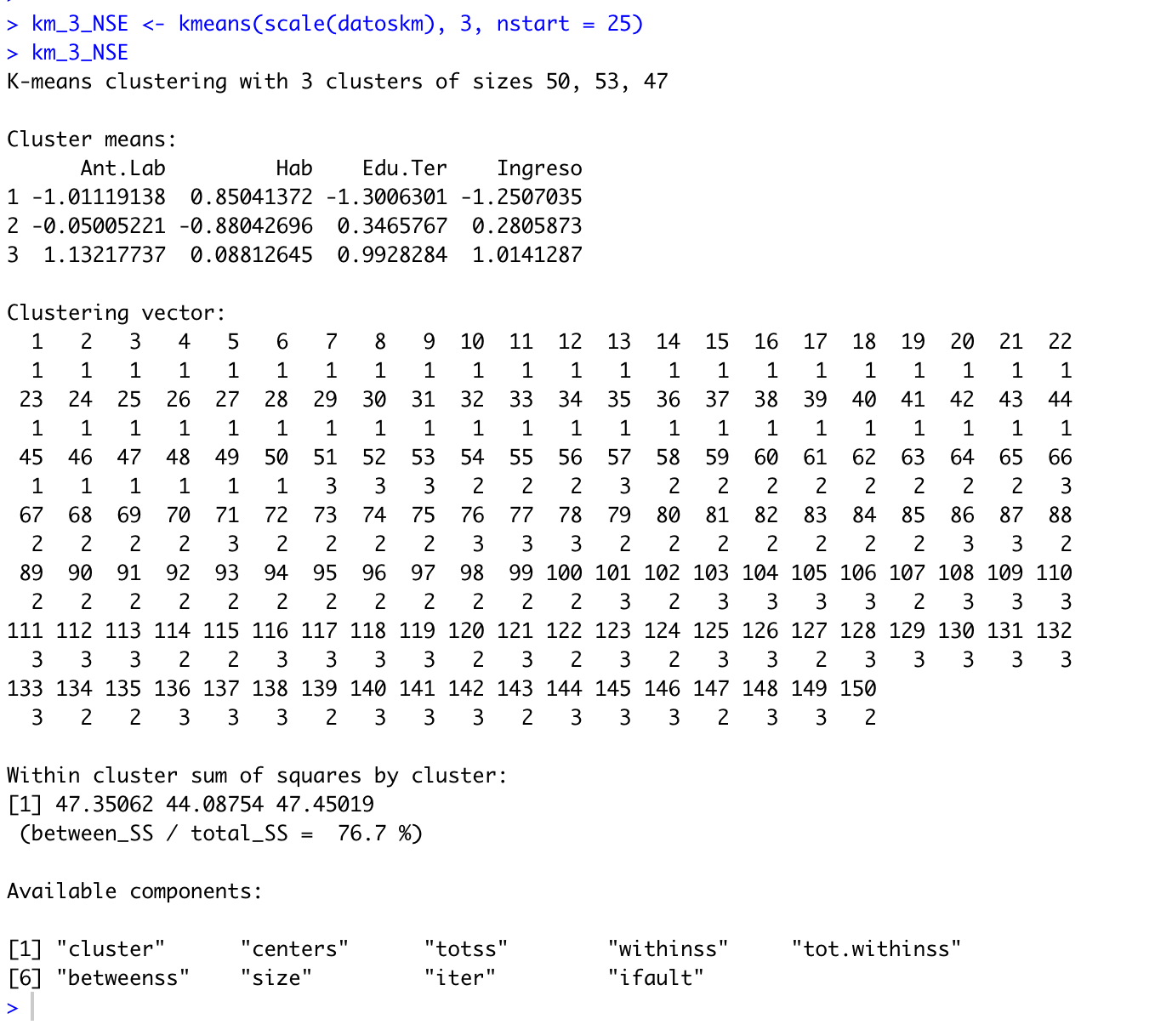
Teniendo esta percepción tratamos de validarlo con el apoyo de Rstudio realizando un gráfico que nos indique la variación del WCV al agregar un cluster adicional.

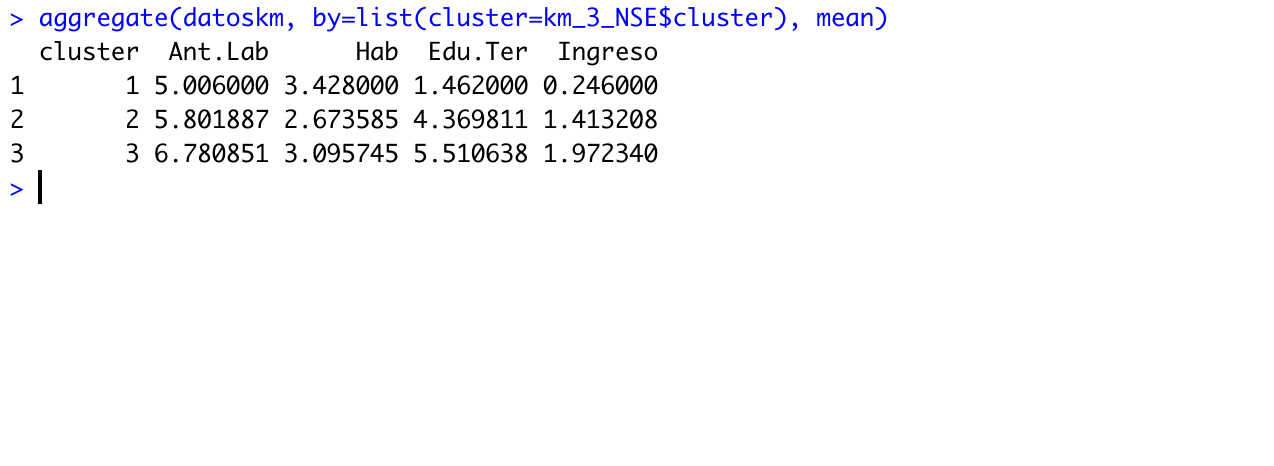




El grafico “Selección de K” nos valida nuestra percepción por lo que avanzamos el modelo con K=3.

El trade off entre el WCV y agregar un conjunto adicional no es relevante.

* Interpretación de los clusters.



Comentarios:

* Cluster 1, 50 observaciones – WCV 47,35062

Cluster 2, 53 observaciones – WCV 44,08754

Cluster 3, 47 observaciones – WCV 47,45019

WCV entre Clusters no presenta variaciones significativas.

* El mejor cluster es el número 3 y el peor el 2. Esto lo determinamos en relación al WCV.
* Con k=3 explicamos el 76.7 de la volatilidad en los datos.
* Nombramiento de Clusters

Cluster 1 - NSE Bajo

Cluster 2 - NSE Medio

Cluster 3 - NSE Alto

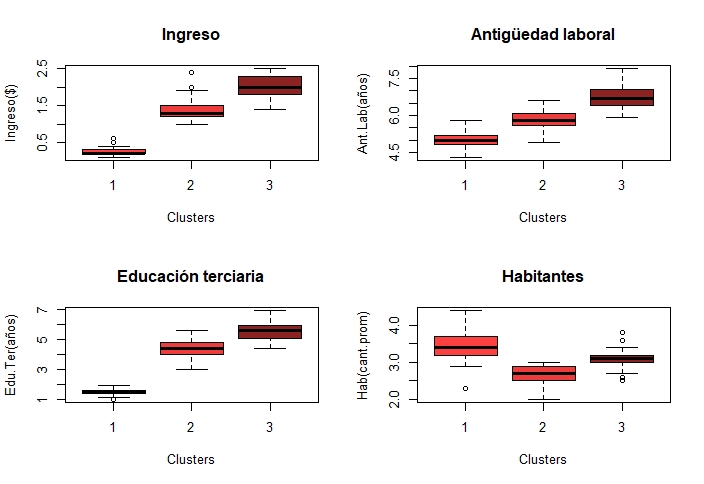
Justificación:

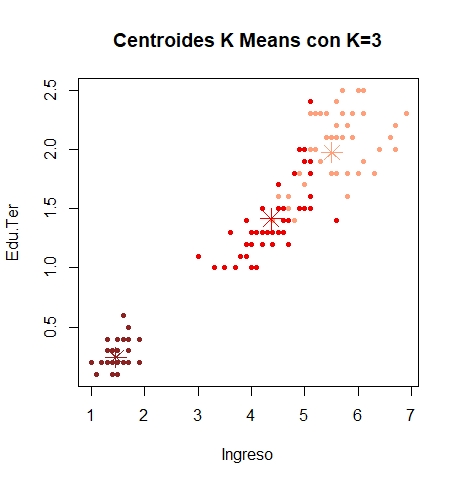
Como punto de partida notamos que el conjunto que presenta valores superiores o en el entorno al promedio en todas las variables es el grupo 3, el cual nos sugiere que el conjunto 3 es el NSE alto.

Quedando dos conjuntos 1 y 2, entendemos que el que tenga mejores promedios de ambos clusters determinara ser el NSE Medio

Para leerlo con mayor facilidad desnormalizamos los datos y notamos que nuestra lectura inicial esta correcta.

El conjunto 3 representa valores superiores en todas las variables, esto nos indica que es conjunto de NSE Alto, el que lo sigue es el conjunto 2, el cual denominamos NSE Medio y como último el conjunto 1 lo denominamos NSE Bajo

Diagramas de caja



### 4.1.1 - Conclusión final:

Los diagramas de caja refuerzan nuestro nombramiento a los conjuntos, es la misma información con una visualización diferente.

La caja de Ingresos y Educación Terciaria nos otorga una información que parece ser intuitiva Mayor Nivel SE, mayor ingreso. Mayor cantidad de años de estudio en educación terciaria, mayor nivel SE.

Antigüedad laboral no parecía ser tan intuitiva pero valida la misma tendencia, a mayor promedio en cantidad de años trabajados en el hogar, mayor NSE.

La variable Cantidad de habitantes promedio en el hogar parece contrastar correctamente con los datos demográficos y tendencias en que los hogares con NSE bajo son los que tienen en promedio mayor cantidad de habitantes.

En su conjunto observamos que el NSE Alto en 3 variables: Ingreso, Edu. Ter y Ant. Lab está por encima del resto, lo contrario sucede con el NSE bajo.

# 5 – Script

#======================================================================

# Universidad ORT Uruguay

# Facultad de Administración y Ciencias Sociales

# Obligatorio de Analítica de Negocios y Big Data

# Docente: Mag. Guillermo Magnou

# Estudiantes: Cinthia Amorin - Nº 188817, Guillermo Trifoglio - Nº 162229, Cecilia Machado - N°213640

#======================================================================

#\*\*\*\*\*\*\*\*\*\*\*\*

# PARTE UNO

#\*\*\*\*\*\*\*\*\*\*\*\*

#======================================================================

# Inicio de preámbulo

# Borrar toda al memoria de trabajo

rm(list=ls())

# Cargamos librerías

library(rio)

library(tidyverse)

library(ggplot2)

library(AppliedPredictiveModeling)

library(caret)

library(Hmisc)

library(funModeling)

library(fastDummies)

library(ISLR)

# Establecemos el directorio de trabajo

setwd('C:/Users/Cecilia Machado/Desktop/Obligatorio Analitica de Negocios y Big Data')

#Chequeamos que haya sido correctamente ejecutado

getwd()

# Cargamos base de datos

datos <- import('base.csv')

# Chequeamos que la cantidad de filas y columnas sean los correctos

dim(datos)

# Visualizamos base de datos

View(datos)

# Visualizamos tipo de datos

str(datos)

head(datos)

#Fin de preámbulo

#======================================================================

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

# Análisis descriptivo

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

# Histogramas

hist(datos$Ingreso, col = "brown3", main = "Histograma de Ingreso", xlab = "Ingreso", ylab = "Frecuencia absoluta" )

hist(datos$Ant.Lab, col = "brown3", main = "Histograma de Antigüedad Laboral", xlab = "Antigüedad Laboral", ylab = "Frecuencia absoluta" )

hist(datos$Hab, col = "brown3", main = "Histograma de Habitantes", xlab = "Habitantes", ylab = "Frecuencia absoluta" )

hist(datos$Edu.Ter, col = "brown3", main = "Histograma de Educación Terciaria", xlab = "Educación Terciaria", ylab = "Frecuencia absoluta" )

# Diagrama de cajas

boxplot(datos$Ingreso, col = "brown3", main = "Diagrama de caja de Ingreso", ylab="Ingreso" )

boxplot(datos$Ant.Lab,col = "brown3", main = "Diagrama de caja de Antigüedad Laboral", ylab="Ant.Lab" )

boxplot(datos$Hab, col = "brown3",main = "Diagrama de caja de Habitantes", ylab="Hab" )

boxplot(datos$Edu.Ter,col = "brown3", main = "Diagrama de caja de Educación Terciaria", ylab="Edu.Ter" )

# Medidas de Resumen, separación y dispersión:

# Resumen de las variables del archivo

summary(datos[,2:6])

# Varianza de variables cuantitativas

Varianza\_Ant.Lab <- var(datos$Ant.Lab)

Varianza\_Ant.Lab

Varianza\_Hab <- var(datos$Hab)

Varianza\_Hab

Varianza\_Edu.Ter <- var(datos$Edu.Ter)

Varianza\_Edu.Ter

Varianza\_Ingreso <- var(datos$Ingreso)

Varianza\_Ingreso

lista\_varianza <- matrix(c(Varianza\_Ant.Lab, Varianza\_Edu.Ter, Varianza\_Hab, Varianza\_Ingreso),ncol=4,byrow=TRUE)

colnames(lista\_varianza) <- c("Ant.Lab","Edu.Ter","Hab", "Ingreso")

rownames(lista\_varianza) <- c("Varianza")

lista\_varianza <- as.table(lista\_varianza)

lista\_varianza

# Desviación estándar

Desviación\_Ant.Lab <- sd(datos$Ant.Lab)

Desviación\_Ant.Lab

Desviación\_Edu.Ter <- sd(datos$Edu.Ter)

Desviación\_Edu.Ter

Desviación\_Hab <- sd(datos$Hab)

Desviación\_Hab

Desviación\_Ingreso <- sd(datos$Ingreso)

Desviación\_Ingreso

lista\_desvio <- matrix(c(Desviación\_Ant.Lab, Desviación\_Edu.Ter, Desviación\_Hab, Desviación\_Ingreso),ncol=4,byrow=TRUE)

colnames(lista\_desvio) <- c("Ant.Lab","Edu.Ter","Hab", "Ingreso")

rownames(lista\_desvio) <- c("Desviación estándar")

lista\_desvio <- as.table(lista\_desvio)

lista\_desvio

# Coeficiente de variación

CoefVariacion\_Ant.Lab <- (Desviación\_Ant.Lab /mean(datos$Ant.Lab ))\*100

CoefVariacion\_Ant.Lab

CoefVariacion\_Edu.Ter <- (Desviación\_Edu.Ter/mean(datos$Edu.Ter))\*100

CoefVariacion\_Edu.Ter

CoefVariacion\_Hab <- (Desviación\_Hab/mean(datos$Hab))\*100

CoefVariacion\_Hab

CoefVariacion\_Ingreso <- (Desviación\_Ingreso/mean(datos$Ingreso))\*100

CoefVariacion\_Ingreso

lista\_cv <- matrix(c(CoefVariacion\_Ant.Lab, CoefVariacion\_Edu.Ter, CoefVariacion\_Hab, CoefVariacion\_Ingreso),ncol=4,byrow=TRUE)

colnames(lista\_cv) <- c("Ant.Lab","Edu.Ter","Hab", "Ingreso")

rownames(lista\_cv) <- c("Desviación estándar")

lista\_cv <- as.table(lista\_cv)

lista\_cv

# Rango

Rango\_Ant.Lab <- (max(datos$Ant.Lab))-(min(datos$Ant.Lab))

Rango\_Ant.Lab

Rango\_Edu.Ter <- (max(datos$Edu.Ter))-(min(datos$Edu.Ter))

Rango\_Edu.Ter

Rango\_Hab <- (max(datos$Hab))-(min(datos$Hab))

Rango\_Hab

Rango\_Ingreso <- (max(datos$Ingreso))-(min(datos$Ingreso))

Rango\_Ingreso

lista\_Rango <- matrix(c(Rango\_Ant.Lab, Rango\_Edu.Ter, Rango\_Hab, Rango\_Ingreso),ncol=4,byrow=TRUE)

colnames(lista\_Rango) <- c("Ant.Lab","Edu.Ter","Hab", "Ingreso")

rownames(lista\_Rango) <- c("Rango")

lista\_Rango <- as.table(lista\_Rango)

lista\_Rango

# Rango Intercualtílico

RIC\_Ant.Lab <- IQR(datos$Ant.Lab)

RIC\_Ant.Lab

RIC\_Edu.Ter <- IQR(datos$Edu.Ter)

RIC\_Edu.Ter

RIC\_Hab <- IQR(datos$Hab)

RIC\_Hab

RIC\_Ingreso <- IQR(datos$Ingreso)

RIC\_Ingreso

lista\_RIC <- matrix(c(RIC\_Ant.Lab, RIC\_Edu.Ter, RIC\_Hab, RIC\_Ingreso),ncol=4,byrow=TRUE)

colnames(lista\_RIC) <- c("Ant.Lab","Edu.Ter","Hab", "Ingreso")

rownames(lista\_RIC) <- c("RIC")

lista\_RIC <- as.table(lista\_RIC)

lista\_RIC

# Lista final con datos obtenidos

lista\_final <- matrix(c(lista\_varianza, lista\_desvio, lista\_cv, lista\_RIC, lista\_Rango),ncol=4,byrow=TRUE)

colnames(lista\_final) <- c("Ant.Lab","Edu.Ter","Hab", "Ingreso")

rownames(lista\_final) <- c("Varianza", "Desviación Estándar", "Coef. Variación","Rango", "RIC")

lista\_final <- as.table(lista\_final)

lista\_final

# Creación de tabla de frecuencias de clases de la variable Ingreso

# Ancho de clases = (max - min)/Número de clases

(2.5-0.1)/3

# Ancho de clases = 0.8

val\_ini\_Ingreso <- 0.1

val\_fin\_Ingreso <- 2.5

salto\_Ingreso <- 0.8

clasesIngreso <- seq(val\_ini\_Ingreso,val\_fin\_Ingreso,salto\_Ingreso)

clasesIngreso

# Se genera una variable tal que cada valor sea a qué clase pertenece cada observación de Ingreso

clases\_Ingreso <- cut(datos$Ingreso, breaks = clasesIngreso)

print(clases\_Ingreso)

# A esa nueva variable, calcularle las frecuencias absolutas:

Frec\_Abs\_Clases\_Ingreso <- table(clases\_Ingreso)

Frec\_Abs\_Clases\_Ingreso

# Expresarlo como un "data frame" que es nuestra tabla deseada a la que le vamos a ir

# agregando columnas

tabla\_frecuencia\_clases\_Ingreso <- data.frame(Frec\_Abs\_Clases\_Ingreso)

tabla\_frecuencia\_clases\_Ingreso

# Agregar al data frame una columna de Frecuencias relativas

Frec\_rel\_Ingreso <- tabla\_frecuencia\_clases\_Ingreso$Freq/sum(tabla\_frecuencia\_clases\_Ingreso$Freq)

tabla\_frecuencia\_clases\_Ingreso$Frec\_rel\_Ingreso <- Frec\_rel\_Ingreso

tabla\_frecuencia\_clases\_Ingreso

# Agregar al data frame una columna de Frecuencias porcentuales

Frec\_por\_Ingreso <- tabla\_frecuencia\_clases\_Ingreso$Freq/sum(tabla\_frecuencia\_clases\_Ingreso$Freq)\*100

tabla\_frecuencia\_clases\_Ingreso$Frec\_por\_Ingreso <- Frec\_por\_Ingreso

tabla\_frecuencia\_clases\_Ingreso

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

#Análisis de correlación.

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

tabla\_correlación <- cor(datos[,2:5])

tabla\_correlación

#Gráficos de dispersión.

Graf\_Ingreso\_vs\_AntLab <- plot(datos$Ant.Lab, datos$Ingreso, col = "brown2", main = 'Ingreso vs Ant. Lab', xlab = 'Ingreso', ylab = 'Antiguedad Laboral')

abline(lm(Ingreso~Ant.Lab, data = datos))

Graf\_Ingreso\_vs\_Hab <- plot(datos$Ingreso, datos$Hab , col = 'brown2', main = 'Ingreso vs Hab', xlab = 'Ingreso', ylab = 'Habitantes')

abline(lm(Hab~Ingreso, data = datos))

Graf\_Ingreso\_vs\_EduTer <- plot(datos$Ingreso, datos$Edu.Ter , col = 'brown2', main = 'Ingreso vs Edu Ter', xlab = 'Ingreso', ylab = 'Educación Terciaria')

abline(lm(Edu.Ter~Ingreso, data = datos))

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

#Regresión lineal múltiple

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

#Creamos las variables dummies para la variable categórica Nivel S.E

datos$Bajo\_dum <- ifelse(datos$Nivel.SE == 'Bajo', 1, 0)

table(datos$Bajo\_dum, datos$Nivel.SE)

datos$Alto\_dum <- ifelse(datos$Nivel.SE == 'Alto', 1, 0)

table(datos$Alto\_dum, datos$Nivel.SE)

#Creamos el dataset de traininig y testing

set.seed(1111)

#Separamos el dataset en 70% para train y 30% para test

train <- sample(nrow(datos), nrow(datos)\*0.7)

test <- (-train)

#Creamos regresión lineal con todas las variables

reg1 <- lm(Ingreso ~ Ant.Lab+Hab+Edu.Ter+Bajo\_dum+Alto\_dum , data = datos,subset = train)

summary(reg1)

vif(reg1)

#Trabajamos con un nivel de significancia, ?? = 0.10

#Regresión lineal aplicando método automático

reg1.step <- step(reg1, direction = "backward")

#Creamos regresión lineal sin las variables: Ant.Lab y Edu.Ter

reg2 <- lm(Ingreso ~ Hab+Bajo\_dum+Alto\_dum , data = datos,subset = train)

summary(reg2)

vif(reg2)

#ECM en testing

mean((datos$Ingreso[test] - predict(reg2, datos[test, ]))\*\*2)

# R-cuadrado en testing

corel <- cor(datos$Ingreso[test], predict(reg2, datos[test, ]))

corel\*\*2

#ECM en train

mean((datos$Ingreso[train] - predict(reg2, datos[train, ]))\*\*2)

# R-cuadrado en train

corel <- cor(datos$Ingreso[train], predict(reg2, datos[train, ]))

corel\*\*2

# Interpretación del modelo

# Predecimos el ingreso de una persona con una cantidad de 3 Habitantes, y nivel socioeconómico Alto

0.57806+0.27217\*3-1.26481\*0+0.66290\*1

# Predecimos que una persona con las características mencionadas anteriormente tendría un ingreso de $20.574,7

#======================================================================

#\*\*\*\*\*\*\*\*\*\*\*\*

#PARTE DOS

#\*\*\*\*\*\*\*\*\*\*\*\*

#======================================================================# Inicio de preámbulo

# Borrar toda al memoria de trabajo

rm(list=ls())

# Cargamos librerías

library(rio)

library(tidyverse)

library(ggplot2)

library(AppliedPredictiveModeling)

library(caret)

library(Hmisc)

library(funModeling)

library(fastDummies)

library(ISLR)

# Establecemos el directorio de trabajo

setwd('C:/Users/Cecilia Machado/Desktop/Obligatorio Analitica de Negocios y Big Data')

#Chequeamos que haya sido correctamente ejecutado

getwd()

# Cargamos base de datos

datos <- import('base.csv')

# Chequeamos que la cantidad de filas y columnas sean los correctos

dim(datos)

# Visualizamos base de datos

View(datos)

# Visualizamos tipo de datos

str(datos)

head(datos)

#Fin de preámbulo

#======================================================================

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

#Modelo de Regresión Logística

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

#Tabla de frecuencias absolutas de variable categórica Nivel S.E Vs Ingreso

datos$Ing\_cat = ifelse(datos$Ingreso <1 , "Menor a 1", ifelse(datos$Ingreso >2, "Mayor a 2", "Entre 1 y 2"))

tabla\_frec\_abs\_IngVSNivelSE <- table(datos$Ing\_cat, datos$Nivel.SE)

tabla\_frec\_abs\_IngVSNivelSE

tabla\_frec\_abs\_IngVSNivelSE\_marg <- addmargins(tabla\_frec\_abs\_IngVSNivelSE)

tabla\_frec\_abs\_IngVSNivelSE\_marg

#Tabla de frecuencias porcentuales de la variable categórica Nivel S.E Vs Ingreso

tabla\_frec\_relPor\_IngVSNivelSE <- addmargins(prop.table(table(datos$Ing\_cat, datos$Nivel.SE), 1), 2)\*100

tabla\_frec\_relPor\_IngVSNivelSE

tabla\_frec\_relPor\_IngVSNivelSE\_red <- round(tabla\_frec\_relPor\_IngVSNivelSE)

tabla\_frec\_relPor\_IngVSNivelSE\_red

#Creamos la variable dummy para la variable categórica Nivel S.E

datos$Alto\_dum <- ifelse(datos$Nivel.SE == 'Alto', 1, 0)

table(datos$Alto\_dum, datos$Nivel.SE)

addmargins(table(datos$Alto\_dum, datos$Nivel.SE))

# Diagrama de caja e histograma

boxplot(datos$Edu.Ter~datos$Alto\_dum, xlab = 'Nivel socioeconómico: 0 = Bajo y Medio, 1 = Alto',

ylab = 'Educación terciaria (años)',main = 'Edu.Ter respecto al Nivel.SE',col= c("brown4", "brown2"))

boxplot(datos$Ingreso~datos$Alto\_dum, xlab = 'Nivel socioeconómico: 0 = Bajo y Medio, 1 = Alto',

ylab = 'Ingreso ($)',main = 'Ingresos respecto al Nivel.SE',col= c("brown4", "brown2"))

par(mfrow = c(2, 2))

hist(datos$Edu.Ter[datos$Alto\_dum == 1], col = 'brown2', main = 'Edu.Ter y Nivel.SE Alto', ylab = "Frecuencia", xlab = "Educación terciaria (años)")

hist(datos$Edu.Ter[datos$Alto\_dum == 0], col = 'brown4', main = 'Edu.Ter y Nivel.SE Medio y Bajo', ylab = "Frecuencia", xlab = "Educación terciaria (años)")

hist(datos$Ingreso[datos$Alto\_dum == 1], col = 'brown2', main = 'Ingreso y Nivel.SE Alto', ylab = "Frecuencia", xlab = "Ingreso ($)")

hist(datos$Ingreso[datos$Alto\_dum == 0], col = 'brown4', main = 'Ingreso y Nivel.SE Medio y Bajo', ylab = "Frecuencia", xlab = "Ingreso ($)")

#Creamos el dataset de traininig y testing

set.seed(1111)

#Separamos el dataset en 70% para train y 30% para test

train <- sample(nrow(datos), nrow(datos)\*0.7)

test <- (-train)

# Regresión Logística con todas las variables

glm.fit <- glm(Alto\_dum ~ Ingreso + Ant.Lab + Hab + Edu.Ter , datos[train, ], family=binomial)

summary(glm.fit)

# Regresión Logística aplicando método atutomático

glm.fit.back <- step(glm.fit, direction = 'backward')

summary(glm.fit.back)

# Regresión logística final sin Hab (nivel de significancia mayor que 0.10)

glm.fit2 <- glm(Alto\_dum ~ Ingreso + Edu.Ter , datos[train, ], family=binomial)

summary(glm.fit2)

vif(glm.fit2)

#Predicción en toda la base de datos

datos$Alto\_dum\_pred <- predict(glm.fit2, datos, type ="response")

summary(datos$Alto\_dum\_pred)

#Predicción en test

datos$Alto\_dum\_pred\_test <- ifelse(datos$Alto\_dum\_pred > 0.5, 1, 0)

table(datos$Alto\_dum\_pred\_test[test])

addmargins(table(datos$Alto\_dum\_pred\_test[test], datos$Alto\_dum[test]))

#Evaluación en test - VER VALORES

# Precisión

(12+32)/45

# Especificidad

32/32

# Sensivilidad

12/13

# Tasa de error

(0+1)/45

# Curva ROC

rocobj <- roc( datos$Alto\_dum[test], datos$Alto\_dum\_pred\_test[test], auc = TRUE, ci = TRUE )

print(rocobj)

plot.roc( rocobj, legacy.axes = TRUE, print.thres = "best", print.auc = TRUE,

auc.polygon = FALSE, max.auc.polygon = FALSE, auc.polygon.col = "gainsboro",

col = 2, grid = TRUE )

# Multicolinealidad

cor(datos$Ingreso,datos$Edu.Ter)

vif(glm.fit2)

#======================================================================

# Inicio de preámbulo

# Borrar toda al memoria de trabajo

rm(list=ls())

# Cargamos librerías

library(rio)

library(rpart)

library(rattle)

library(corrplot)

library(pROC)

# Establecemos el directorio de trabajo

setwd('C:/Users/Cecilia Machado/Desktop/Obligatorio Analitica de Negocios y Big Data')

#Chequeamos que haya sido correctamente ejecutado

getwd()

# Cargamos base de datos

datos <- import('base.csv')

# Chequeamos nombre de las variables

names(datos)

# Chequeamos que la cantidad de filas y columnas sean los correctos

dim(datos)

# Visualizamos base de datos

View(datos)

# Visualizamos tipo de datos

str(datos)

head(datos)

#Resumen de los datos

summary(datos)

#Fin de preámbulo

#======================================================================

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

#Modelo de Árbol de clasificación

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

#Creamos las variables dummies para la variable categórica Nivel S.E

datos$Alto\_dum <- ifelse(datos$Nivel.SE == 'Alto', 1, 0)

table(datos$Alto\_dum, datos$Nivel.SE)

#Creamos el dataset de traininig y testing

set.seed(1111)

#Sacamos las variables individuales ID y Nivel S.E

datos2 <- datos [,-c(1,6)]

#Separamos el dataset en 70% para train y 30% para test

train <- sample(nrow(datos2), nrow(datos2)\*0.7)

test <- (-train)

#Estimamos nuestro primer modelo en training, con el método "class"

arbol.inicial <- rpart(Alto\_dum ~ Ant.Lab + Hab + Edu.Ter + Ingreso , datos2[train, ], method = 'class')

arbol.inicial

#Graficamos nuestro arbol inicial

fancyRpartPlot(arbol.inicial)

#Graficamos la performance del modelo vs. la complejidad

plotcp(arbol.inicial)

#Ver reglas

asRules(arbol.inicial)

#Hacemos la prediccion del modelo sobre todo el dataset

datos2$pred\_arbol\_NSEAlto = predict(arbol.inicial, datos2)[,2]

summary(datos2$pred\_arbol\_NSEAlto)

# Curva ROC en test

roc\_test <- roc(datos2$Alto\_dum[test], datos2$pred\_arbol\_NSEAlto [test], auc = TRUE, ci = TRUE )

print(roc\_test)

plot.roc( roc\_test, legacy.axes = TRUE, print.thres = "best", print.auc = TRUE,

auc.polygon = FALSE, max.auc.polygon = FALSE, auc.polygon.col = "gainsboro",

col = 2, grid = TRUE )

# Curva ROC en train

roc\_train <- roc(datos2$Alto\_dum[train], datos2$pred\_arbol\_NSEAlto [train], auc = TRUE, ci = TRUE )

print(roc\_train)

plot.roc( roc\_train, legacy.axes = TRUE, print.thres = "best", print.auc = TRUE,

auc.polygon = FALSE, max.auc.polygon = FALSE, auc.polygon.col = "gainsboro",

col = 2, grid = TRUE )

# Tabla de confusion con probabilidad > 0.5 como 'punto de corte'

addmargins(table(datos2$pred\_arbol\_NSEAlto[test] > 0.5, datos2$Alto\_dum[test]))

# Precisión

(13+30)/45

# Sensibilidad

13/13

# Especificidad

30/32

# Tasa de error

(0+2)/45

#=====================================================================

#\*\*\*\*\*\*\*\*\*\*\*\*

#PARTE TRES

#\*\*\*\*\*\*\*\*\*\*\*\*

#======================================================================

# Inicio de preámbulo

# Borrar toda al memoria de trabajo

rm(list=ls())

library(rio)

library(tidyverse)

library(ggplot2)

library(AppliedPredictiveModeling)

library(caret)

library(Hmisc)

library(funModeling)

library(fastDummies)

library(ISLR)

# Establecemos el directorio de trabajo

setwd('C:/Users/Cecilia Machado/Desktop/Obligatorio Analitica de Negocios y Big Data')

#Chequeamos que haya sido correctamente ejecutado

getwd()

# Cargamos base de datos

datos <- import('base.csv')

# Chequeamos nombre de las variables

names(datos)

# Chequeamos que la cantidad de filas y columnas sean los correctos

dim(datos)

# Visualizamos base de datos

View(datos)

# Visualizamos tipo de datos

str(datos)

head(datos)

#Resumen de los datos

summary(datos)

#Fin de preámbulo

#======================================================================

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*

#Modelo K-Means

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*

# Sacamos las variables ID y NivelSE del dataset

datoskm <- datos[,2:5]

view(datoskm)

summary(datoskm)

# Justificación de cantidad de clusters

# Calculamos el WSS para distintos valores de K desde 1 hasta 8

wcv <- sapply(1:8 , function(k){kmeans(datoskm,k,nstart=25, iter.max=8)$tot.withinss})

wcv

plot(1:8, wcv, type="b", main = 'Seleccionar K', xlab = "Cantidad Clusters (K)", ylab = "WCV")

# Cluster con K=3

set.seed(123)

km\_3\_NSE <- kmeans(scale(datoskm), 3, nstart = 25)

km\_3\_NSE

# Cantidad de observaciones por clusters

table(km\_3\_NSE$cluster)

# Centroides con K=3

Centroides <- aggregate(datoskm, by=list(cluster=km\_3\_NSE$cluster), mean)

# Incorporo los cluster a los datos

NSE\_c <- cbind(datoskm, cluster = km\_3\_NSE$cluster)

# Visualizamos la información de cada cluster

NSE\_c[NSE\_c$cluster==1,]

NSE\_c[NSE\_c$cluster==2,]

NSE\_c[NSE\_c$cluster==3,]

# Interpretación gráfica de los clusters

par(mfrow = c(2, 2))

boxplot(Ingreso~km\_3\_NSE$cluster,data = datos,main = 'Ingreso',xlab = "Clusters",ylab = "Ingreso($)",col = c('brown1', 'brown2', 'brown4'))

boxplot(Ant.Lab~km\_3\_NSE$cluster,data = datos,main = 'Antigüedad laboral',xlab = "Clusters",ylab = "Ant.Lab(años)",col = c('brown1', 'brown2', 'brown4'))

boxplot(Edu.Ter~km\_3\_NSE$cluster,data = datos,main = 'Educación terciaria',xlab = "Clusters",ylab = "Edu.Ter(años)",col = c('brown1', 'brown2', 'brown4'))

boxplot(Hab~km\_3\_NSE$cluster,data = datos,main = 'Habitantes',xlab = "Clusters",ylab = "Hab(cant.prom)",col = c('brown1', 'brown2', 'brown4'))

# Gráfico con centroides

vcol <- c('brown4', 'red2', "lightsalmon")

plot(datoskm$Edu.Ter, datoskm$Ingreso, main = "Centroides K Means con K=3", xlab = "Ingreso", ylab = "Edu.Ter",

col= vcol[km\_3\_NSE$cluster], pch = 20, cex = 1)

points(Centroides$Edu.Ter, Centroides$Ingreso, col= vcol , pch = 8, cex = 2)

plot(datoskm$Ant.Lab, datoskm$Ingreso, main = "Centroides K Means con K=3", xlab = "Ingreso", ylab = "Ant.Lab",

col= vcol[km\_3\_NSE$cluster], pch = 20, cex = 1)

points(Centroides$Ant.Lab, Centroides$Ingreso, col= vcol, pch = 8, cex = 2)

plot(datoskm$Hab, datoskm$Ingreso, main = "Centroides K Means con K=3", xlab = "Ingreso", ylab = "Hab",

col= vcol[km\_3\_NSE$cluster], pch = 20, cex = 1)

points(Centroides$Hab, Centroides$Ingreso, col = vcol , pch = 8, cex = 2)

1. Hay multicolinealidad cuando existe una alta correlación entre las variables explicativas. Se calcula a través del factor de inflación de la varianza [↑](#footnote-ref-1)
2. 5 ^2 = 100 combinaciones diferentes de modelos. [↑](#footnote-ref-2)