Taller Analítica de Datos

Diplomado: Big Data & Business Analytics.

- Actividad M4-M5 (SCRIPT)

Modelos de Analítica Aplicados a los Negocios

Modelos Aplicados a Marketing y Finanzas

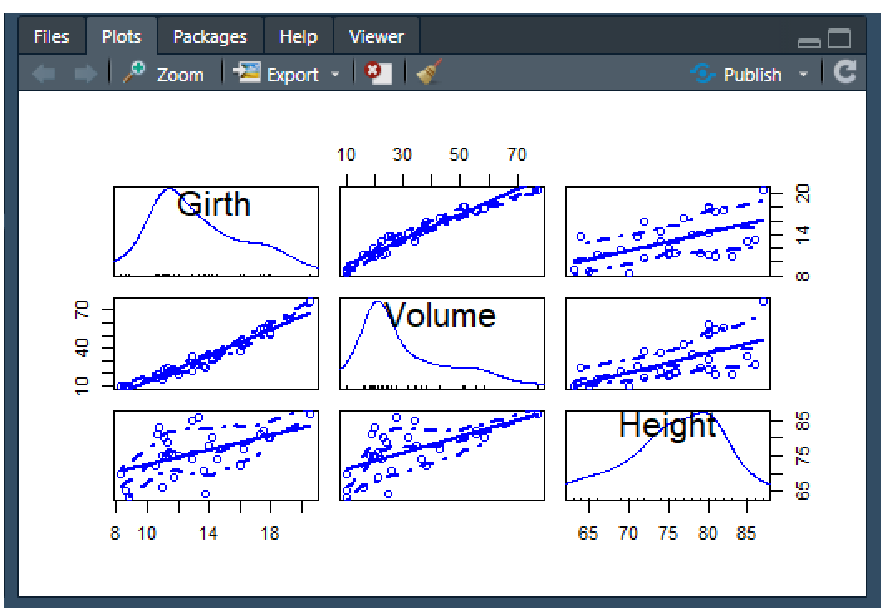
Integrantes: Miguel Ángel Ñustes Sánchez.

1° Ejercicio básico de regresión lineal múltiple ## - DATASET: TREES

Identificación de Variables (Significativas)

```
> round(cor(x = trees[c(1,2,3)]),2)
       Girth Height Volume
Girth   1.00   0.52   0.97
Height  0.52   1.00   0.60
Volume  0.97   0.60   1.00
>
```

```
> scatterplotMatrix(~Girth+Volume+Height, data = trees)
>
```

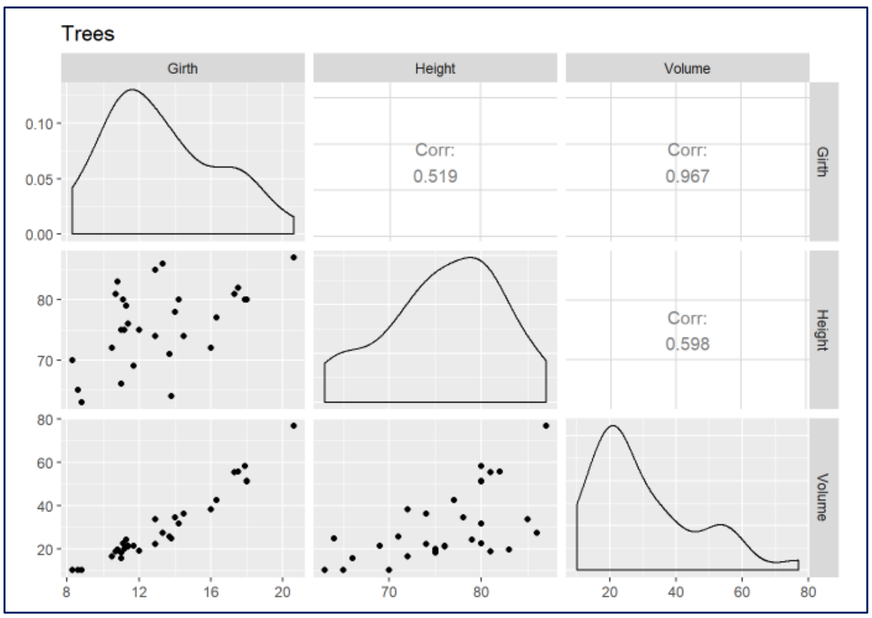


Análisis de Variables

Las estadísticas y la visualización muestran que existe una fuerte correlación entre la variable (Girth) circunferencia y la variable (Volume) volumen de los árboles.

Los resultados de las siguientes visualizaciones de correlaciones entre variables, reflejan una mayor correlación entre las variables Girth y Volume.

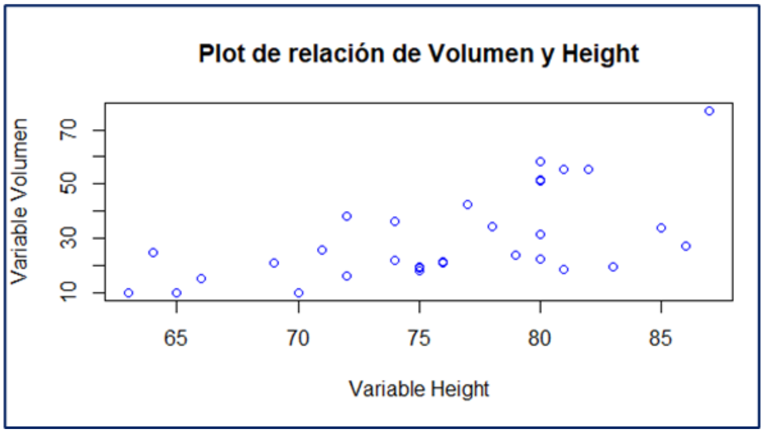
```
## Correlación
ggpairs(data = trees, columns = 1:3, title="Trees")
```



Se descarta mediante el análisis de correlación de variables, la variable **Height**. Debido a tener un valor menor (0,598) respecto al resultado de correlación de la variable **Girth** (0,967) en función de la variable **Volume.**

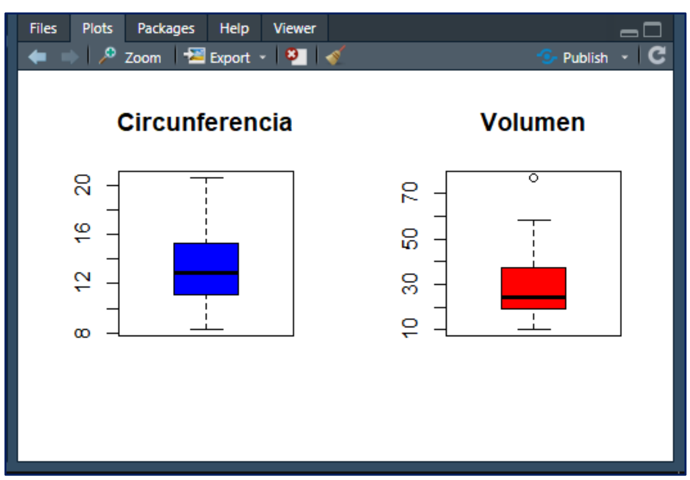
Con el objetivo de validar el análisis anterior, se realiza un diagrama de dispersión de la variable **Height** en función a la variable **Volume,** se corrobora estadísticamente que las variables **(Height – Volume)** no tienen una correlación superior a la expuesta anteriormente **(Girth – Volume).** Esto se visualiza en el siguiente diagrama de dispersión.

```r
plot(trees$Height, trees$Volume,col ="blue",
     ylab = "Variable Volumen", xlab = "Variable Height",
     main = "Plot de relación de Volumen y Height")
```



Adicionalmente, se analiza las distribuciones de las dos variables significativas, se corrobora mediante diagramas de caja, que la variable **Volume** no se distribuye normalmente, pero la variable **Girth** sí. Reflejados en las siguientes visualizaciones.

```r
        ##  Boxplot : Dataset Trees ##
par(mfrow=c(1,2))
boxplot(trees$Girth, col="blue", main = "Circunferencia")
boxplot(trees$Volume, col="red", main = "Volumen")
boxplot(trees$Height, col="red", main = "Altura")
```



Debido a estos resultados, las pruebas de hipótesis sobre distribución no normal, se deben realizar a partir de pruebas no paramétricas.

```
> cor.test(trees$Girth, trees$Volume,method="pearson")

        Pearson's product-moment correlation

data:  trees$Girth and trees$Volume
t = 20.478, df = 29, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9322519 0.9841887
sample estimates:
      cor
0.9671194
```

La correlación visualiza un valor **p <2.2e-16**, por lo tanto, la correlación debe ser significativa. Se concluye así que, en la composición o compendio del modelo de regresión lineal, la variable **Volume** debe estar en ahí por su nivel de significancia.

Análisis del Modelo – Regresión Lineal

```
> print(Mod_1)

Call:
lm(formula = Volume ~ Girth + Height, data = trees)

Coefficients:
(Intercept)        Girth       Height
   -57.9877       4.7082       0.3393

>
```

```
> summary(Mod_1)

Call:
lm(formula = Volume ~ Girth + Height, data = trees)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Girth         4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared:  0.948,     Adjusted R-squared:  0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16

>
```

El volumen de los árboles se determina como la variable **Volume** dependiente y se pretende buscar las variables significativas que tengan un impacto dominante en el crecimiento de los árboles.

Se puede visualizar estadísticamente que la variable **Girth** y la variable **Volume** están correlacionadas.

Inicialmente el modelo de regresión se compone de la siguiente manera:

Mod_1 = lm (Volume ~ Girth+Height, data = trees) Generando un R-Cuadrado de 0.9242, es un resultado

el **Mod_1 está** constituido por la variable **Height,** generando un resultado 0.9442> R cuadrado ajustado [Mod_1]: 0.9242 por lo tanto, debemos incluir la variable **Heigth** en nuestro modelo.

Realizando un análisis de modelo inicial **Mod_1** contemplando el **R-Cuadrado de 0.9242.** Se propone realizar un modelo adicional con la finalidad de aumentar el valor del R-Cuadrado, reflejado en la siguiente expresión:

Mod_2 = lm(Volume~Height+Girth+Height*Girth, data = trees)

summary(Mod_2)

```
C:/Users/migue/Desktop/LABORAL/POSTGRADO - MIGUEL ÑUSTES - JAVERIANA/Diplomado - U.Sabana - Big Data/MOD - 4/

> Mod_2 = lm(Volume~Height+Girth+Height*Girth, data = trees)
> summary(Mod_2)

Call:
lm(formula = Volume ~ Height + Girth + Height * Girth, data = trees)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.5821 -1.0673  0.3026  1.5641  4.6649 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  69.39632   23.83575   2.911  0.00713 ** 
Height       -1.29708    0.30984  -4.186  0.00027 ***
Girth        -5.85585    1.92134  -3.048  0.00511 ** 
Height:Girth  0.13465    0.02438   5.524 7.48e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.709 on 27 degrees of freedom
Multiple R-squared:  0.9756,	Adjusted R-squared:  0.9728 
F-statistic: 359.3 on 3 and 27 DF,  p-value: < 2.2e-16

> 
```

Se determina que este modelo (Mod_2) con un R-Cuadrado ajustado = 0,9728, refleja un modelo mejor ajustado, por ende, es un modelo confiable para contemplarlo en las predicciones que se deseen realizar. Se debe esclarecer que no se puede afirmar que el modelo (Mod_2) es el último propuesto y el mejor, debido a que siempre habrá una mejor versión posible.

Análisis del Modelo de Predicciones –

- **Primer Caso:** altura 73.4 pies y diámetro 20 pulgadas.
  Predicción Inicial

```
## 1.5 Predecir un nuevo caso ##
## 1° Caso ##

## Predicción 1° Caso ##
PrimerCaso = pr                                                    ', level = 0.95)
print(PrimerCas
  ## Nota ##
  Segundo_Caso = predict (Mod_2, data.frame(Girth=18.1, Height=65.9))
  print(Segundo_Caso)
  summary(Segundo_Caso)
str(PrimerCaso)
## Estadísticos
summary(PrimerCaso)
```

Predicción Final

```
## Nota ##
Primer_Caso = predict (Mod_2, data.frame(Girth=20, Height=73.4))
print(Primer_Caso)
summary(Primer_Caso)
```

Basado en el segundo modelo de regresión lineal (Mod_2) que generó un ajuste R-Cuadrado de 0.97, se realiza la anterior predicción con los valores dados en las variables representativas.

- **Segundo Caso:** altura 65.9 pies y diámetro 18.1 pulgadas.
  Predicción Inicial

```
##  1.5  Predecir un nuevo caso  ##
## 2° Caso ##

## Predicción 2° Caso ##
SegundoCaso = predict(object = Mod_1, newdata = data.frame(Height=65.9, Girth=18.1), interval = "confidence", level = 0.95)

str(SegundoCaso)

## Estadísticos del modelo ##
summary(SegundoCaso)
```

Predicción Final

```
## Nota ##
Segundo_Caso = predict (Mod_2, data.frame(Girth=18.1, Height=65.9))
print(Segundo_Caso)
summary(Segundo_Caso)
```

- **Tercer Caso:** altura 53.7 pies y diámetro 15.4 pulgadas
  Predicción Inicial

```
##  1.5  Predecir un nuevo caso  ##
## 3° Caso ##

## Predicción 3° Caso ##
TercerCaso = predict(object = Mod_1, newdata = data.frame(Height=53.7, Girth=15.4), interval = "confidence", level = 0.95)

str(TercerCaso)

## Estadísticos del modelo ##
summary(TercerCaso)
```

Predicción Final

```
## Nota ##
Tercer_Caso = predict (Mod_2, data.frame(Girth=15.4, Height=53.7))
print(Tercer_Caso)
summary(Tercer_Caso)
```

- **Summary**

```
> summary(Primer_Caso)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  54.75   54.75   54.75   54.75   54.75   54.75 
> summary(Segundo_Caso)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  38.54   38.54   38.54   38.54   38.54   38.54 
> summary(Tercer_Caso)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  20.92   20.92   20.92   20.92   20.92   20.92 
> 
```

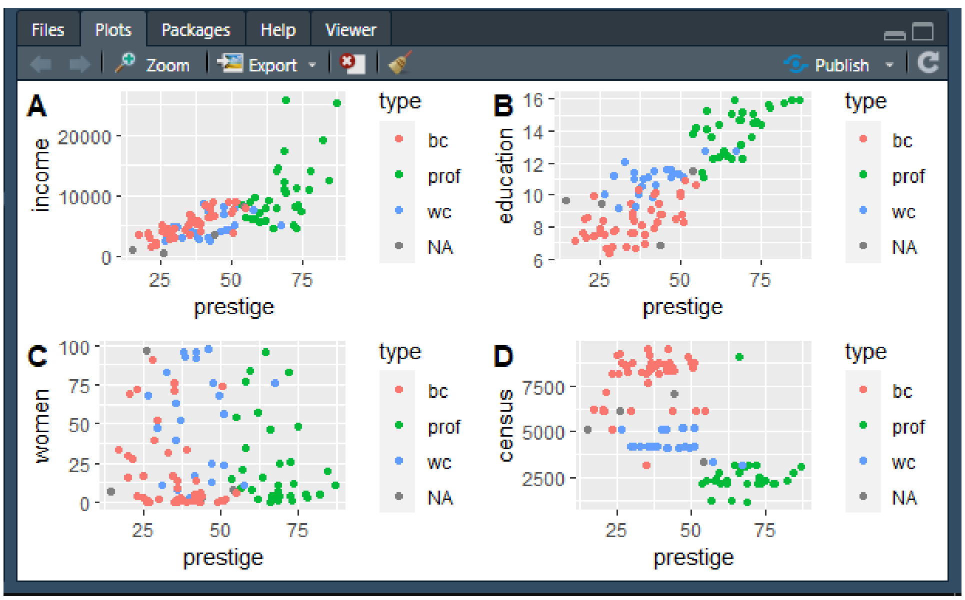
Se determina la predicción para el primer, segundo y tercer caso en función a la variable **(Volume).** Concluyendo este modelo (Mod_2) con un R-Cuadrado ajustado = 0,9728, refleja un modelo mejor ajustado, por ende, es un modelo confiable para contemplarlo en las predicciones realizadas. Se debe esclarecer que no se puede afirmar que el modelo (Mod_2) es el último propuesto y el mejor, debido a que siempre habrá una mejor versión posible.

2° Ejercicio de regresión lineal múltiple ## - DATASET: PRESTIGE

Análisis Gráficos – Variables

Se realiza el análisis de gráficos, identificando el comportamiento de los datos, con esto ver si es necesario realizar algún tipo de preprocesamiento de la información, mediante las siguientes funciones y gráficos.

```
##  2.3  Análisis gráfico ##
## Relación
plot_income <- ggplot(data = Prestige, aes(x = prestige, y = income, col = type)) + geom_point()
plot_education <- ggplot(data = Prestige, aes(x = prestige, y = education, col = type)) + geom_point()
plot_women <- ggplot(data = Prestige, aes(x = prestige, y = women, col = type)) + geom_point()
plot_census <- ggplot(data = Prestige, aes(x = prestige, y = census, col = type)) + geom_point()

plot_grid(plot_income, plot_education, plot_women, plot_census, labels = "AUTO")
```

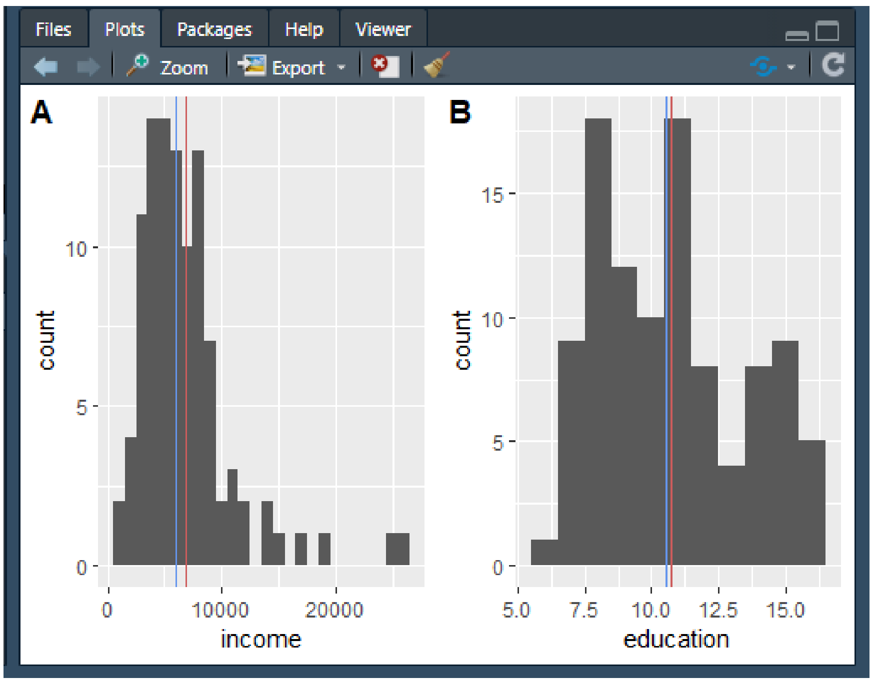


Se observa una fuerte relación lineal positiva de la variable **Prestige** con la variable **Income** (ingresos) y la variable **Education** (educación) más que con la variable **Women** (mujeres) y la variable **Census** (censo). Se interpreta una tendencia en esta relación lineal **(Education – Prestige)** cuando ambas variables aumentan o disminuyen simultáneamente a un ritmo constante.

En la relación de la Variable **Census** en función o respecto a la variable **Prestige** se identifica una relación no lineal, reflejando un comportamiento de aumentos/disminuciones entre ellas que no se dan con la misma intensidad.

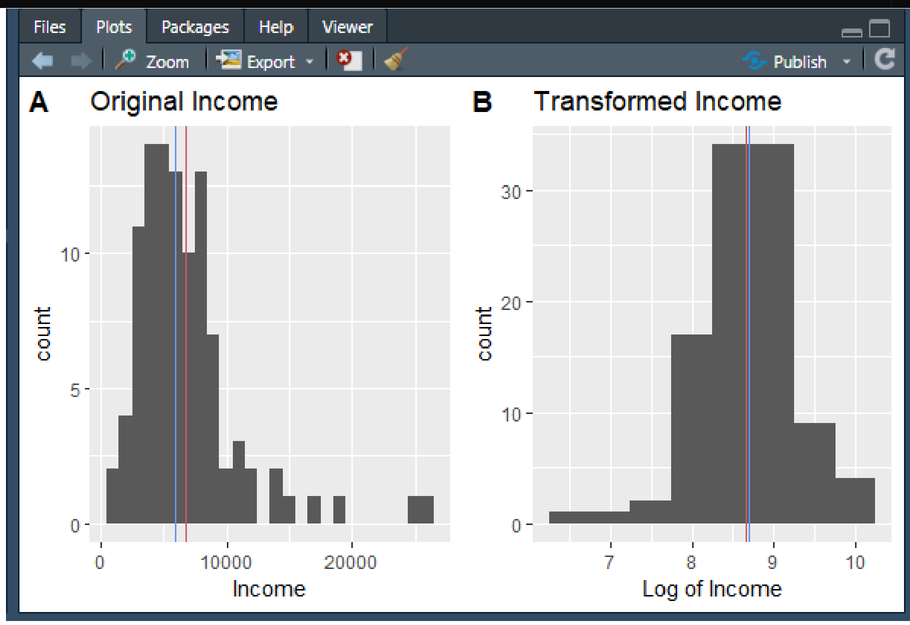
Después de identificar la relación lineal más fuerte entre variables **(Education – Prestige)**, se procede a realizar un análisis a la distribución de datos de las variables de ingresos y educación a través del diagrama de historial y compararlas con los valores medios y medianos, presentados a continuación.

```
## Histogramas
hist_income <- ggplot(Prestige, aes(x = income)) + geom_histogram(binwidth = 1000) +
  geom_vline(xintercept = mean(Prestige$income), color = "indianred") +
  geom_vline(xintercept = median(Prestige$income), color = "cornflowerblue")
hist_education <- ggplot(Prestige, aes(x = education)) + geom_histogram(binwidth = 1) +
  geom_vline(xintercept = mean(Prestige$education), color = "indianred") +
  geom_vline(xintercept = median(Prestige$education), color = "cornflowerblue")

plot_grid(hist_income, hist_education, labels = "AUTO")
```



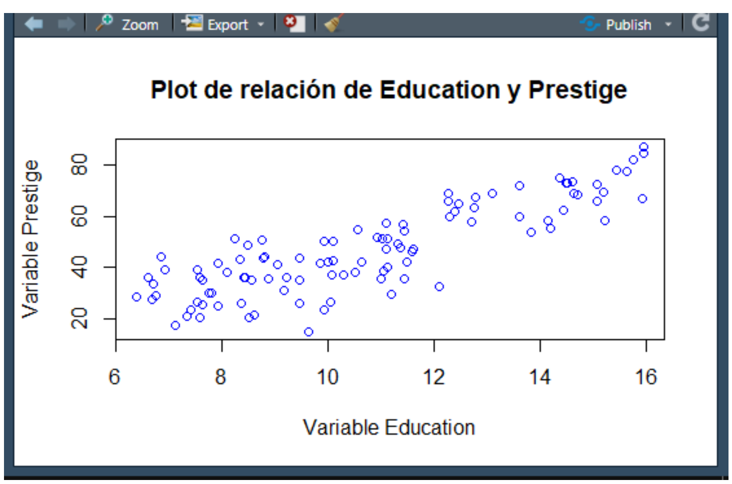
Se observa que la variable **Income** (ingreso) es una distribución sesgada a la derecha y que la educación tampoco representa la distribución normal. Debido a la interpretación de las visualizaciones anteriores, se transforma esto en una distribución normal si es posible. Mediante el uso de Log2 para la variable **Income (**ingreso) y escalaremos el valor de la variable educación en su media. Generando así una comparación entre el histograma de la variable **Income** (Original) **y** el histograma de la variable **Income** sin sesgos (Modificado).

```r
# ------------------------------------------------------------------------------------
##Histograma Original Variable Income / Histograma Modificado Variable Income

  hist_income <- ggplot(Prestige, aes(x = income)) + geom_histogram(binwidth = 1000) +
  labs(title = "Original Income") + labs(x ="Income") +
  geom_vline(xintercept = mean(Prestige$income), color = "indianred") +
  geom_vline(xintercept = median(Prestige$income), color = "cornflowerblue")
hist_Update_income <- ggplot(Prestige, aes(x = log(income))) + geom_histogram(binwidth = 0.5) +
  labs(title = "Transformed Income") + labs(x ="Log of Income") +
  geom_vline(xintercept = mean(log(Prestige$income)), color = "indianred") +
  geom_vline(xintercept = median(log(Prestige$income)), color = "cornflowerblue")

plot_grid(hist_income, hist_Update_income, labels = "AUTO")
```



Finalmente se valida con un diagrama de dispersión, graficando la variable **Income** en función de la variable **Prestige**, mostrado a continuación.

```r
plot( Prestige$education, Prestige$prestige, col ="blue",
      ylab = "Variable Prestige", xlab = "Variable Education",
      main = "Plot de relación de Education y Prestige")
```

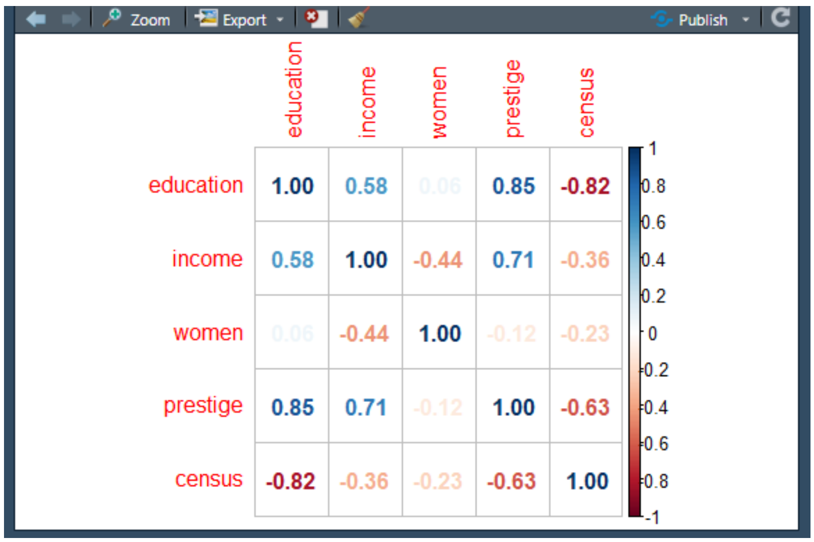


Análisis de los Modelos – Regresión Lineal Múltiple

Inicialmente para el desarrollo y análisis del modelo de regresión lineal múltiple, se identifican las variables mas representativas del conjunto de datos.

Las siguientes visualizaciones expresan las variables mas representativas, estudiadas en un modelo de correlación en función de la variable **Prestige.**

```
cor(Prestige[,-6])
corrplot(cor(Prestige[,-6]) , method = "number")
```



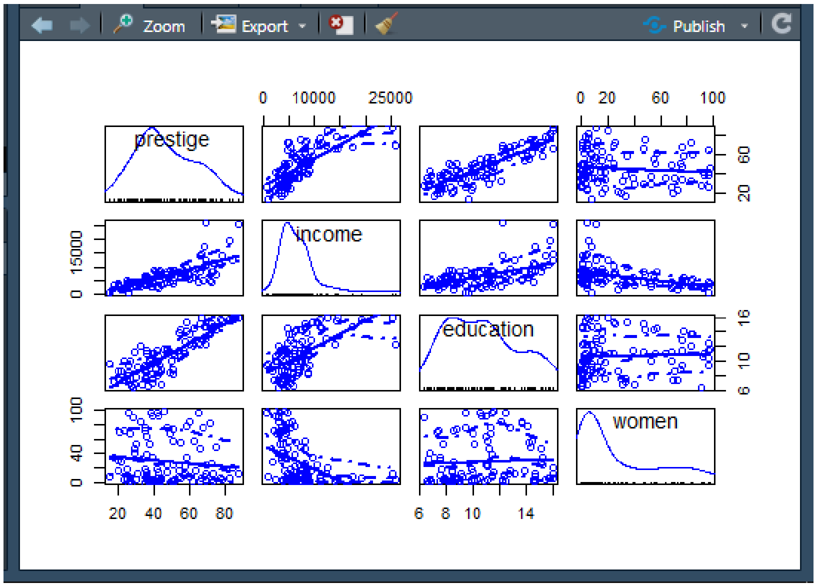
El resultado obtenido en la correlación mas fuerte es **(Education vs Prestige) = 0.85**

Modelo 1

Se procede al análisis del Modelo de regresión lineal múltiple, con el fin de determinar cómo se relaciona la calificación de la variable Prestige con las variables **Income**, **Education** y **Women**, para ello se realiza una regresión lineal múltiple.

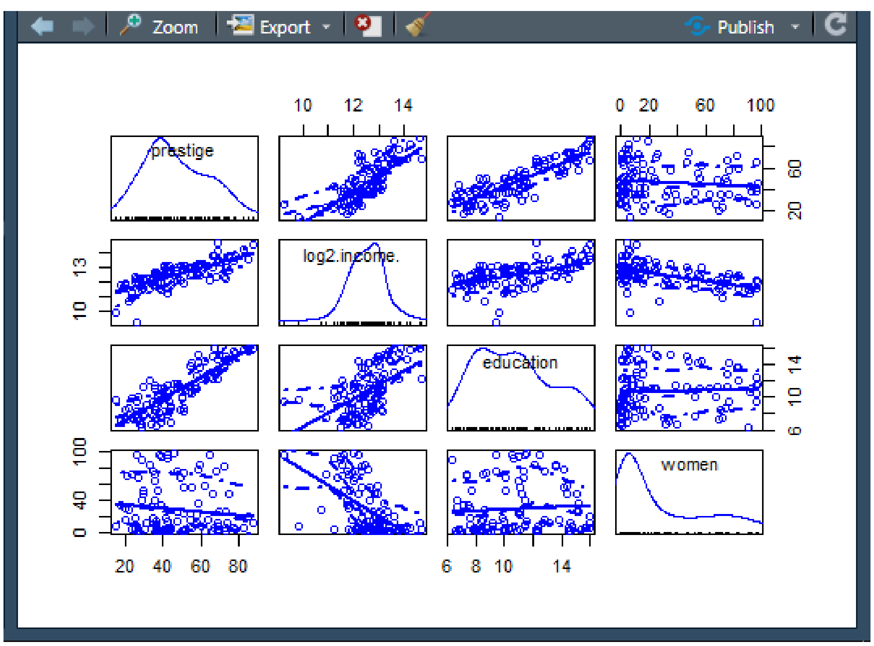
```
## 2.5 Análisis construción de modelos de regresión lineal múltiple ##
scatterplotMatrix(~ prestige + income +education + women, span =0.7, data = Prestige)
```

A partir de esta expresión se visualiza el siguiente comportamiento de las variables en el diagrama de dispersión.



El diagrama de dispersión anterior de la variable dependiente prestigio y el ingreso del predictor muestra una forma no lineal de puntos de datos. Debido a esto se realiza una modificación en lugar de utilizar la variable **Income** directamente, se utiliza el logaritmo de la variable **Income** con base 2 para transformar la forma de curvatura de los datos.

```
##  2.5  Análisis construción de modelos de regresión lineal múltiple ##
scatterplotMatrix(~ prestige + log2(income) +education + women, span =0.7, data = Prestige)
```



Como resultado, el diagrama de dispersión entre el prestigio y log2 (ingresos) no muestra una forma de curvatura a continuación.

```
##  1° Análisis regresión lineal múltiple  ##
prestige.mod1 <- lm(prestige ~ education + log2(income) + women, data= Prestige)

summary(prestige.mod1)
```

```
There were 50 or more warnings (use warnings() to see the first 50)
> summary(prestige.mod1)

Call:
lm(formula = prestige ~ education + log2(income) + women, data = Prestige)

Residuals:
    Min      1Q  Median      3Q     Max
-17.364  -4.429  -0.101   4.316  19.179

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -110.9658    14.8429  -7.476 3.27e-11 ***
education       3.7305     0.3544  10.527  < 2e-16 ***
log2(income)    9.3147     1.3265   7.022 2.90e-10 ***
women           0.0469     0.0299   1.568     0.12
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.093 on 98 degrees of freedom
Multiple R-squared:  0.8351,	Adjusted R-squared:   0.83
F-statistic: 165.4 on 3 and 98 DF,  p-value: < 2.2e-16
```

A partir de **Estimate (Education), b1 = 3.7305** implica que se espera que la calificación de prestigio aumente en 3.7305 unidades por un año adicional de educación. Además, la hipótesis de que la calificación de prestigio se relaciona linealmente con el nivel educativo con otros predictores siendo constantes es la siguiente:

Ho: b1 es igual a 0 (sin relación lineal)

Ha: b1 no es igual a 0 (relación lineal significativa)

Entonces, para el estadístico de prueba t = 10.527 y el valor p para el estadístico de prueba (t = 10.527) es menor que 2 * 10 ^ (- 16). Lo que significa que la probabilidad de obtener el estadístico de prueba 10.527 por casualidad bajo el supuesto de b1 = 0 es extremadamente rara.

Entonces rechazamos la hipótesis nula b1 = 0 y muestra la evidencia de una relación lineal positiva entre el nivel de educación y el nivel de calificación de preevaluación.

Modelo 2

A continuación, se calcula el segundo modelo de regresión sin las variables **Women**. Solo 2 predictores, la variable Education y log2 (variable **Income**), se utilizan para la regresión.

```
  ##  2° Análisis regresión lineal múltiple  ##
prestige.mod2 <- lm(prestige ~ education + log2(income), data= Prestige)

summary(prestige.mod2)
```

```
> summary(prestige.mod2)

Call:
lm(formula = prestige ~ education + log2(income), data = Prestige)

Residuals:
     Min       1Q   Median       3Q      Max 
-17.0346  -4.5657  -0.1857   4.0577  18.1270 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -95.1940    10.9979  -8.656 9.27e-14 ***
education      4.0020     0.3115  12.846  < 2e-16 ***
log2(income)   7.9278     0.9961   7.959 2.94e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.145 on 99 degrees of freedom
Multiple R-squared:  0.831,	Adjusted R-squared:  0.8275 
F-statistic: 243.3 on 2 and 99 DF,  p-value: < 2.2e-16
```

Se identifica en el análisis del segundo modelo, que no hay una diferencia significativa entre el modelo de regresión1 con variable **Women** y el modelo de regresión2 sin variable **Women**, aunque la intersección con el eje y y las pendientes de la variable **Education** y log2 (variable **Income**) han cambiado ligeramente del modelo1 al modelo2. Finalmente se contempla que el modelo susceptible de tener un mejor ajuste.

Modelo 3

Inicialmente en este modelo se crea un nuevo dataset, con el fin de poder administrar los datos que se contienen en el dataset **Prestige,** se procede a escalar el valor de la variable **Education** a su valor medio.

```
prstg_df = Prestige

# Escalar el valor de la educación a su valor medio
set.seed(1)
education.c = scale(prstg_df$education, center=TRUE, scale=FALSE)
prstg_df = cbind(prstg_df, education.c)
```

El modelo 3 se denota en la siguiente expresión (formula). Ajustando el modelo de regresión lineal con las variables **(Income y Education)** centradas. Obteniendo las siguientes estadísticas.

```
## Modelo 3 ##
lm_mod3 = lm(prestige ~ education.c + log(income), data = prstg_df)
summary(lm_mod3)
```

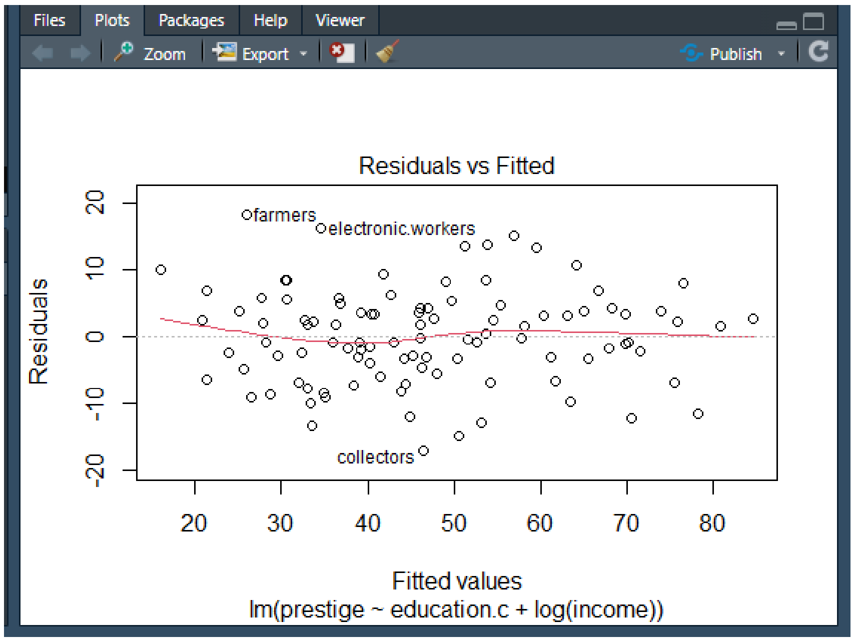
```
> summary(lm_mod3)

Call:
lm(formula = prestige ~ education.c + log(income), data = prstg_df)

Residuals:
     Min       1Q   Median       3Q      Max 
-17.0346  -4.5657  -0.1857   4.0577  18.1270 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -52.2209    12.4661  -4.189 6.09e-05 ***
education.c   4.0020     0.3115  12.846  < 2e-16 ***
log(income)  11.4375     1.4371   7.959 2.94e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.145 on 99 degrees of freedom
Multiple R-squared:  0.831,	Adjusted R-squared:  0.8275 
F-statistic: 243.3 on 2 and 99 DF,  p-value: < 2.2e-16
```

Se observa un ajuste mejorado del modelo de regresión, identificando un R-Cuadrado = 0.8275.

Se grafican los valores residuales y la gráfica del último modelo de regresión construido. Los residuos son la diferencia entre los valores reales y predichos.

```
## Modelo 3 ##
lm_mod3 = lm(prestige ~ education.c + log(income), data = prstg_df)
summary(lm_mod3)
par(mfrow = c(2, 2))
plot(lm_mod3)
```

Se identifican los valores residuales y la gráfica del último modelo de regresión construido. Los residuos son básicamente la diferencia entre los valores reales y predichos. Desde el resumen anterior del modelo, los residuos varían de -17 a 18 y puede ver en el gráfico a continuación que están distribuidos uniformemente.



Finalmente, en este Modelo 3, se observa que $R2$ ha aumentado a el 82% y los residuos oscilan entre -17,5 y 18. Debido a esto aún se procede a realizar otro ajuste al modelo el cual se contemplara en el Modelo 4.

Modelo 4

En el siguiente modelo, se procede a agregar la variable **Type** al modelo de regresión lineal y se examinará si el modelo ha mejorado y tiene sentido, de la siguiente manera.

```
# ----------------------------------------------------------------------

## Modelo 4 ##
lm_mod4 = lm(prestige ~ education.c + log(income) + type, data = prstg_df)
summary(lm_mod4)

# ----------------------------------------------------------------------
```

Se obtienen los siguientes resultados e indicadores estadísticos.

```
> summary(lm_mod4)

Call:
lm(formula = prestige ~ education.c + log(income) + type, data = prstg_df)

Residuals:
    Min      1Q  Median      3Q     Max
-13.511  -3.746   1.011   4.356  18.438

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -45.9329    15.2089  -3.020  0.00326 **
education.c   3.2845     0.6081   5.401 5.06e-07 ***
log(income)  10.4875     1.7167   6.109 2.31e-08 ***
typeprof      6.7509     3.6185   1.866  0.06524 .
typewc       -1.4394     2.3780  -0.605  0.54645
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.637 on 93 degrees of freedom
Multiple R-squared:  0.8555,	Adjusted R-squared:  0.8493
F-statistic: 137.6 on 4 and 93 DF,  p-value: < 2.2e-16
```

Sin embargo, la variable **Type** tienen un coeficiente b3 = 6.7509 y un valor t = 1.886. El valor p de obtener el estadístico t 1.568 es 0.06524 que es menor que el nivel alfa = 0.05. Lo que implica que existe una relación lineal significativa.

El valor múltiple de R-cuadrado es 0,8493 lo que implica que aproximadamente el 84,9% de la variabilidad de la variable dependiente se explica por la línea de regresión ajustada. Entonces, la combinación ponderada de las 3 variables predictoras explicó aproximadamente el 84,9% de la varianza de la variable dependiente.

Predicciones

```
> summary(Prediccion_1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  60.79   60.79   60.79   60.79   60.79   60.79
> summary(Prediccion_2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  52.47   52.47   52.47   52.47   52.47   52.47
> summary(Prediccion_3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  97.84   97.84   97.84   97.84   97.84   97.84
> summary(Prediccion_4)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  84.52   84.52   84.52   84.52   84.52   84.52
> 
```

```
##  2.6  Realizar predicciones ##

## Primera Predicción

lm_mod3 = lm(prestige ~ education.c + log(income), data = prstg_df)
Prediccion_1 = predict (lm_mod3, data.frame(education.c=2.2, income=8979))

summary(Prediccion_1)


-----------------------------------------------------------------------------

## Segunda Predicción

lm_mod3 = lm(prestige ~ education.c + log(income), data = prstg_df)
Prediccion_2 = predict (lm_mod3, data.frame(education.c=8.5, income=5692))

summary(Prediccion_2)

-----------------------------------------------------------------------------
```

```
-----------------------------------------------------------------------------

## Tercera Predicción

lm_mod3 = lm(prestige ~ education.c + log(income), data = prstg_df)
Prediccion_3 = predict (lm_mod3, data.frame(education.c=14.2, income=3486))

summary(Prediccion_3)


-----------------------------------------------------------------------------

## Cuarta Predicción

lm_mod3 = lm(prestige ~ education.c + log(income), data = prstg_df)
Prediccion_4 = predict (lm_mod3, data.frame(education.c=8.40, income=8054))

summary(Prediccion_4)
```

Se realiza las predicciones basadas en el modelo 3 de regresión lineal, El valor múltiple de R-cuadrado es 0,8275 lo que implica que aproximadamente el 82,7% de la variabilidad de la variable dependiente se explica por la línea de regresión ajustada. Entonces, los residuos oscilan entre -17,5 y 18. La combinación ponderada de las 3 variables predictoras explicó aproximadamente el 82,7% de la varianza de la variable dependiente.

3° Ejercicio de Clasificación ## - DATASET: CHURN

Procesamiento de Datos

Se identifican la data que es representativa para inicializar los modelos.

```
sapply(churn_2, function(x) sum(is.na(x)))

churn_2[is.na(churn_2$TotalCharges),]
------------------------------------------
```

```
C:/Users/migue/Desktop/LABORAL/POSTGRADO - MIGUEL ÑUSTES - JAVERIANA/Diplomado - U.Sabana - Big Data/MOD - 4/
>  sapply(churn_2, function(x) sum(is.na(x)))
Tiene_plan_internacional                  Minutos_dia             Llamadas_dia
                       0                            0                        0
 Minutos_internacionales                Reclamaciones Llamadas_internacionales
                       0                            0                        0
             Cancelacion
                       0
>
```

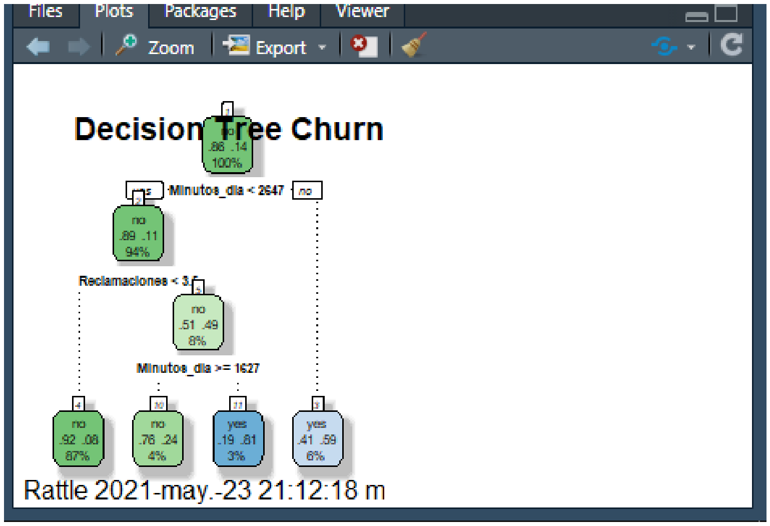
```
>  churn_2[is.na(churn_2$TotalCharges),]
[1] Tiene_plan_internacional Minutos_dia             Llamadas_dia
[4] Minutos_internacionales  Reclamaciones           Llamadas_internacionales
[7] Cancelacion
<0 rows> (or 0-length row.names)
>
```

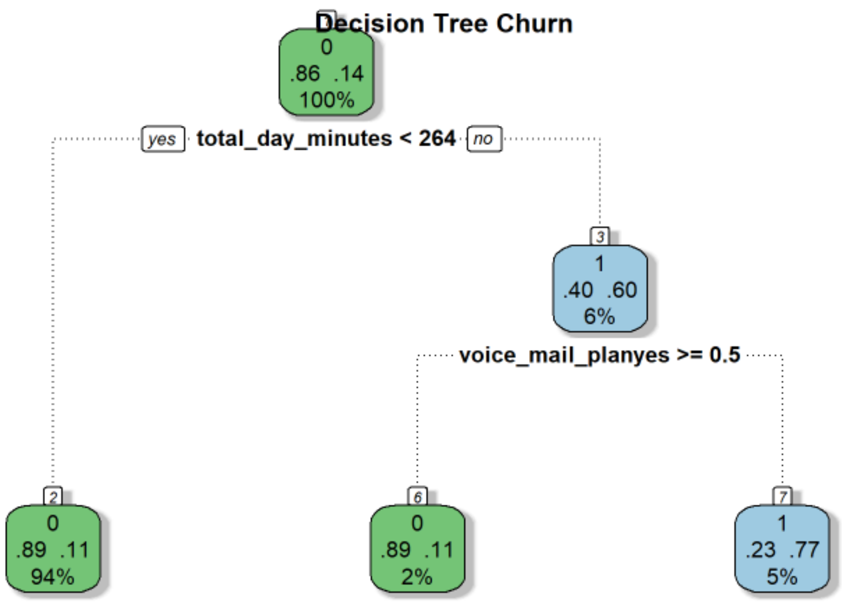
Modelo de Árbol de decisión

Se procede a realizar el modelamiento. El análisis del árbol de decisiones es un método de clasificación que utiliza modelos de decisiones en forma de árbol y sus posibles resultados. Es una de las herramientas más utilizadas en el análisis del machine learning.

```

## 3.5 Modelo de Árbol de Decisión
library(rpart)
library(rattle)

tree = train(factor(Cancelacion) ~., method = "rpart", data = churn_2)

fancyRpartPlot(tree$finalModel, main="Decision Tree Churn")

```

Se visualiza el siguiente diagrama de árbol de decisión.





Se genera la matriz de confusión, buscando analizar los estadísticos de validación.

```
## Matriz de confusión
treePred <- predict(tree, newdata = test_)
confusionMatrix(test_$Cancelacion, treePred)
```

Obteniendo **un Accuracy de 0.89**

```
Console   Terminal   Jobs
C:/Users/migue/Desktop/LABORAL/POSTGRADO - MIGUEL ÑUSTES - JAVERIANA/Diplomado - U.Sabana - E
> confusionMatrix(test_$Cancelacion, treePred)
Confusion Matrix and Statistics

          Reference
Prediction  no yes
       no  828  21
       yes  88  53

               Accuracy : 0.8899
                 95% CI : (0.8687, 0.9087)
    No Information Rate : 0.9253
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4379

 Mcnemar's Test P-Value : 2.588e-10

            Sensitivity : 0.9039
            Specificity : 0.7162
         Pos Pred Value : 0.9753
         Neg Pred Value : 0.3759
             Prevalence : 0.9253
         Detection Rate : 0.8364
   Detection Prevalence : 0.8576
      Balanced Accuracy : 0.8101

       'Positive' Class : no

>
```

Modelo Random Forest

Se procede a realizar el modelo de Random Forest en la siguiente expresión (Fórmula).

```
# --------------------------------------------------------------


## 3.6 Modelo Random Forest

library(randomForest)

modRF = randomForest(factor(Cancelacion) ~.,
                     data = train_,
                     ntree = 500,
                     mtry = 5,
                     importance = TRUE,
                     na.action=randomForest::na.roughfix,
                     replace=FALSE)
modRF
```

```
> modRF = randomForest(factor(Cancelacion) ~.,
+                         data = train_,
+                         ntree = 500,
+                         mtry = 5,
+                         importance = TRUE,
+                         na.action=randomForest::na.roughfix,
+                         replace=FALSE)
> modRF

Call:
 randomForest(formula = factor(Cancelacion) ~ ., data = train_,      ntree = 500, mtry
 = 5, importance = TRUE, replace = FALSE,      na.action = randomForest::na.roughfix)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 5

        OOB estimate of  error rate: 9.35%
Confusion matrix:
      no yes class.error
no  3322 122  0.03542393
yes  253 313  0.44699647
>
```

Se comprueba mediante la matriz de confusión la precisión **(Accuracy) = 0.91**

```
> confusionMatrix(test_$Cancelacion, RFPred)
Confusion Matrix and Statistics

          Reference
Prediction  no yes
       no  829  20
       yes  68  73

               Accuracy : 0.9111
                 95% CI : (0.8916, 0.9281)
    No Information Rate : 0.9061
    P-Value [Acc > NIR] : 0.316

                  Kappa : 0.5759

 Mcnemar's Test P-Value : 5.437e-07

            Sensitivity : 0.9242
            Specificity : 0.7849
         Pos Pred Value : 0.9764
         Neg Pred Value : 0.5177
             Prevalence : 0.9061
         Detection Rate : 0.8374
   Detection Prevalence : 0.8576
      Balanced Accuracy : 0.8546

       'Positive' Class : no

>
```

Modelo de SVM

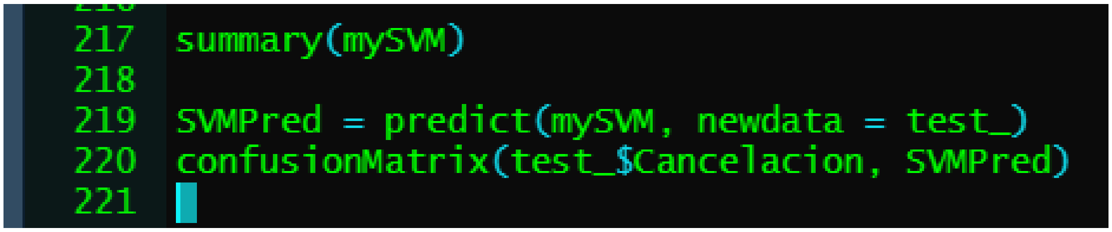
Se procede a realizar el modelo de SVM en la siguiente expresión (Fórmula).

```r
-------------------------------------

## 3.7 Modelo SVM

library(e1071)

mySVM = svm(factor(Cancelacion) ~.,
                data = train_,
                scale = TRUE,
                kernel = "radial",
                cachesize = 3000,
                shrinking = T,
                cost = 4,
                epsilon = 0.2)

summary(mySVM)
```

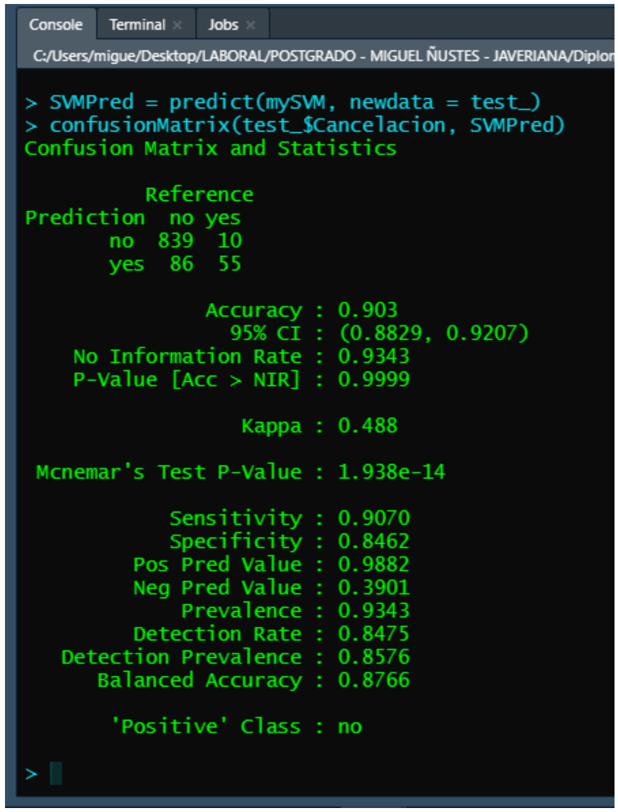
Se obtiene los siguientes estadísticos

```
Console   Terminal ×   Jobs ×
C:/Users/migue/Desktop/LABORAL/POSTGRADO - MIGUEL ÑUSTES - JAVERIANA/Diplomado - U.Sabana - Big Data/MOD - 4/
> summary(mySVM)

Call:
svm(formula = factor(Cancelacion) ~ ., data = train_, kernel = "radial",
    cachesize = 3000, shrinking = T, cost = 4, epsilon = 0.2, scale = TRUE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  4

Number of Support Vectors:  1030

 ( 548 482 )


Number of Classes:  2

Levels:
 no yes
```

Se realiza la validación del modelo, mediante la matriz de confusión, generando una precisión **(Acurancy = 0.90)**



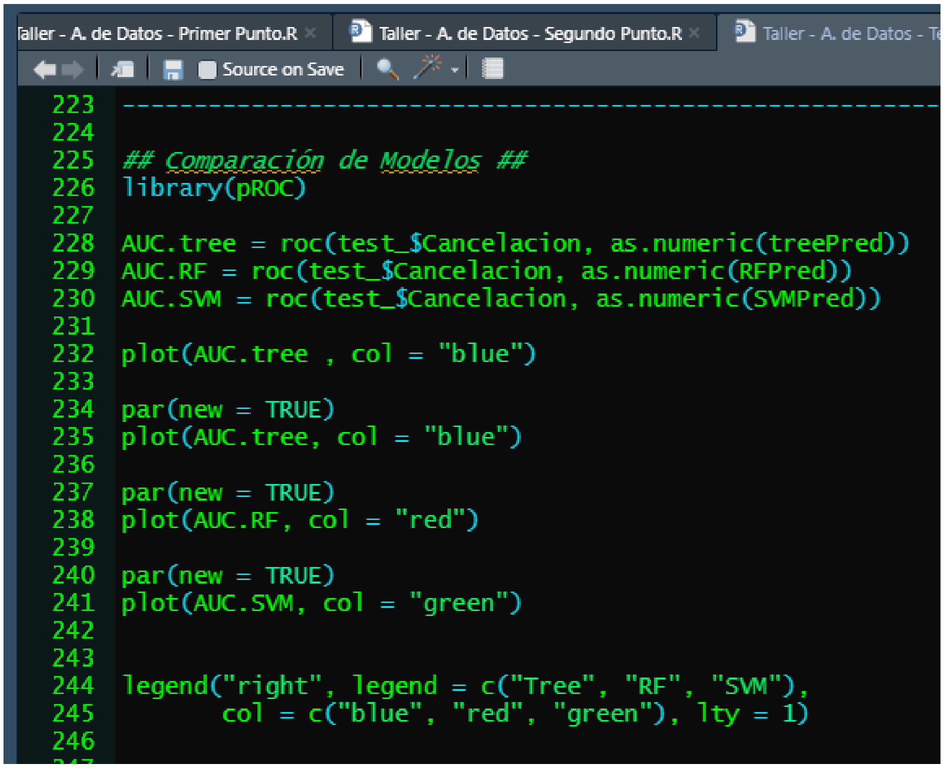
```
summary(mySVM)

SVMPred = predict(mySVM, newdata = test_)
confusionMatrix(test_$Cancelacion, SVMPred)
```



```
Console  Terminal  Jobs
C:/Users/migue/Desktop/LABORAL/POSTGRADO - MIGUEL ÑUSTES - JAVERIANA/Diplon

> SVMPred = predict(mySVM, newdata = test_)
> confusionMatrix(test_$Cancelacion, SVMPred)
Confusion Matrix and Statistics

          Reference
Prediction  no yes
       no  839  10
       yes  86  55

               Accuracy : 0.903
                 95% CI : (0.8829, 0.9207)
    No Information Rate : 0.9343
    P-Value [Acc > NIR] : 0.9999

                  Kappa : 0.488

 Mcnemar's Test P-Value : 1.938e-14

            Sensitivity : 0.9070
            Specificity : 0.8462
         Pos Pred Value : 0.9882
         Neg Pred Value : 0.3901
             Prevalence : 0.9343
         Detection Rate : 0.8475
   Detection Prevalence : 0.8576
      Balanced Accuracy : 0.8766

       'Positive' Class : no

>
```

Resumen / Comparación de Modelos

Se realiza la comparación de los modelos expuestos.



```r
# ----------------------------------------------------------------

## Comparación de Modelos ##
library(pROC)

AUC.tree = roc(test_$Cancelacion, as.numeric(treePred))
AUC.RF = roc(test_$Cancelacion, as.numeric(RFPred))
AUC.SVM = roc(test_$Cancelacion, as.numeric(SVMPred))

plot(AUC.tree , col = "blue")

par(new = TRUE)
plot(AUC.tree, col = "blue")

par(new = TRUE)
plot(AUC.RF, col = "red")

par(new = TRUE)
plot(AUC.SVM, col = "green")


legend("right", legend = c("Tree", "RF", "SVM"),
       col = c("blue", "red", "green"), lty = 1)
```