Análisis estadístico de la Encuesta Nacional de Salud y Nutrición (2012)

Equipo 14.

Cesar Pérez Apolinar (cesar.perez.apolinar@gmail.com)

Roberto Solís Robles (rsolis@uaz.edu.mx)

Humberto Guzmán Martínez (humberto.gmtz@outlook.com)

Victor Manuel Ramírez Alba (vm.ramirezalba@ugto.mx)

Ulises Osvaldo Tomás Canseco (tocu13@hotmail.com)

Marxil Sánchez García (marxilsg89@gmail.com)

Diana Palafox Monreal (dnplfx@hotmail.com)

Objetivo.

- Utilizando la siguiente base de datos. https://raw.githubusercontent.com/beduExpert/Programacion-R-Santander-2022/main/Sesion-06/data/advertising.csv

  Realizar los siguientes puntos:

I. Establecer si existe correlación entre los gastos en alimentos saludables (o no saludables) con respecto al nivel socio económico, así mismo, si hay recursos financieros extras y si existe inseguridad alimentaria.

II. Determinar las posibles causas de la inseguridad alimentaria (IA).

III. Comprobar si los hogares con menor nivel socioeconómico tienden a gastar más en productos no saludables que las personas con mayores niveles socioeconómicos por lo que presentan cierta inseguridad alimentaria.

Análisis Descriptivo

- Limpieza de datos y transformación de variables

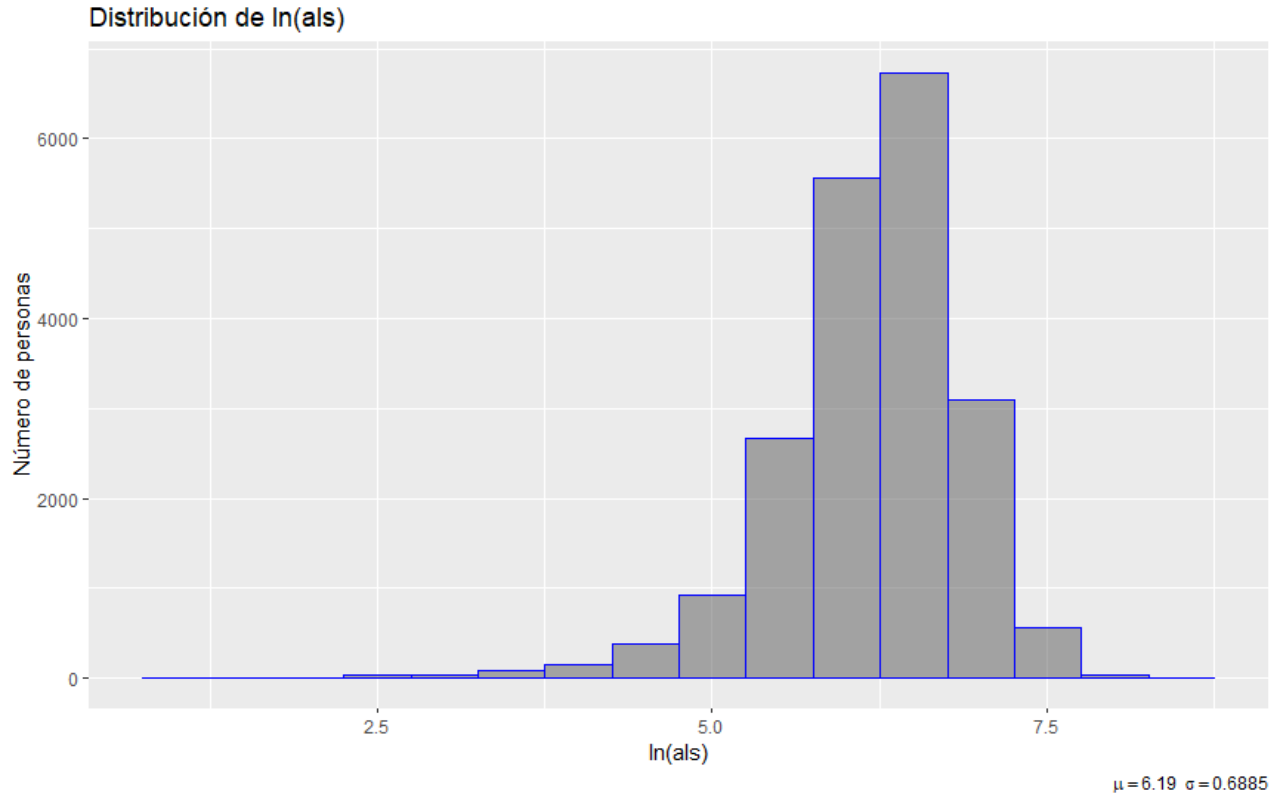
Summary original (40809) vs summary clean (20280)

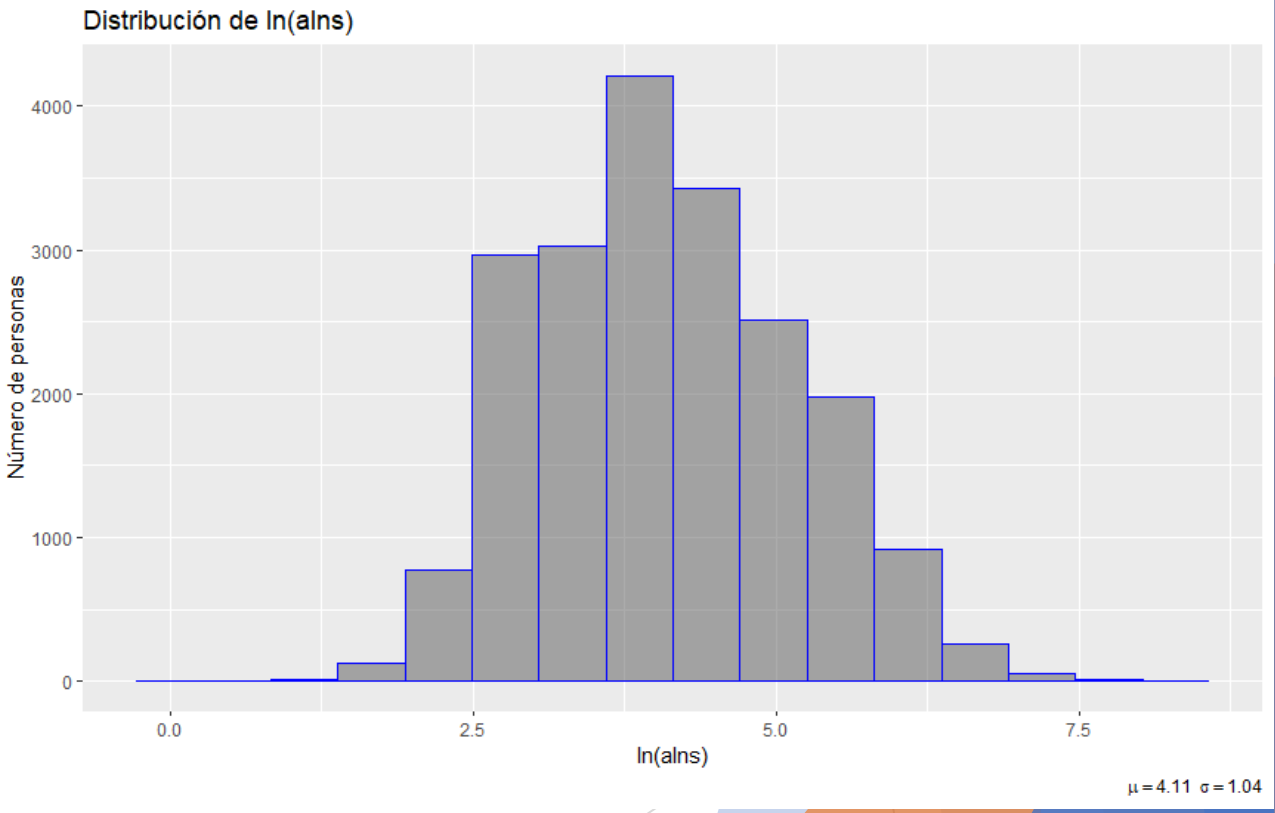
```
nse5f           area           numpeho          refin          edadjef          sexojef
Bajo     :3553   Urbana:13959   Min.   : 1.000   No:16421   Min.   : 18.00   Hombre:15887
MedioBajo:3927   Rural : 6321   1st Qu.: 3.000   Si: 3859   1st Qu.: 36.00   Mujer : 4393
Medio    :4119                  Median : 4.000              Median : 46.00
MedioAlto:4364                  Mean   : 3.991              Mean   : 47.32
Alto     :4317                  3rd Qu.: 5.000              3rd Qu.: 57.00
                                Max.   :19.000              Max.   :101.00
   añosedu         IA            ln_als          ln_alns
Min.   : 0.0   No: 5853   Min.   :1.099   Min.   :0.000
1st Qu.: 9.0   Si:14427   1st Qu.:5.844   1st Qu.:3.401
Median :12.0              Median :6.274   Median :4.007
Mean   :10.9              Mean   :6.192   Mean   :4.119
3rd Qu.:12.0              3rd Qu.:6.633   3rd Qu.:4.868
Max.   :24.0              Max.   :8.605   Max.   :8.298
```

Análisis Descriptivo

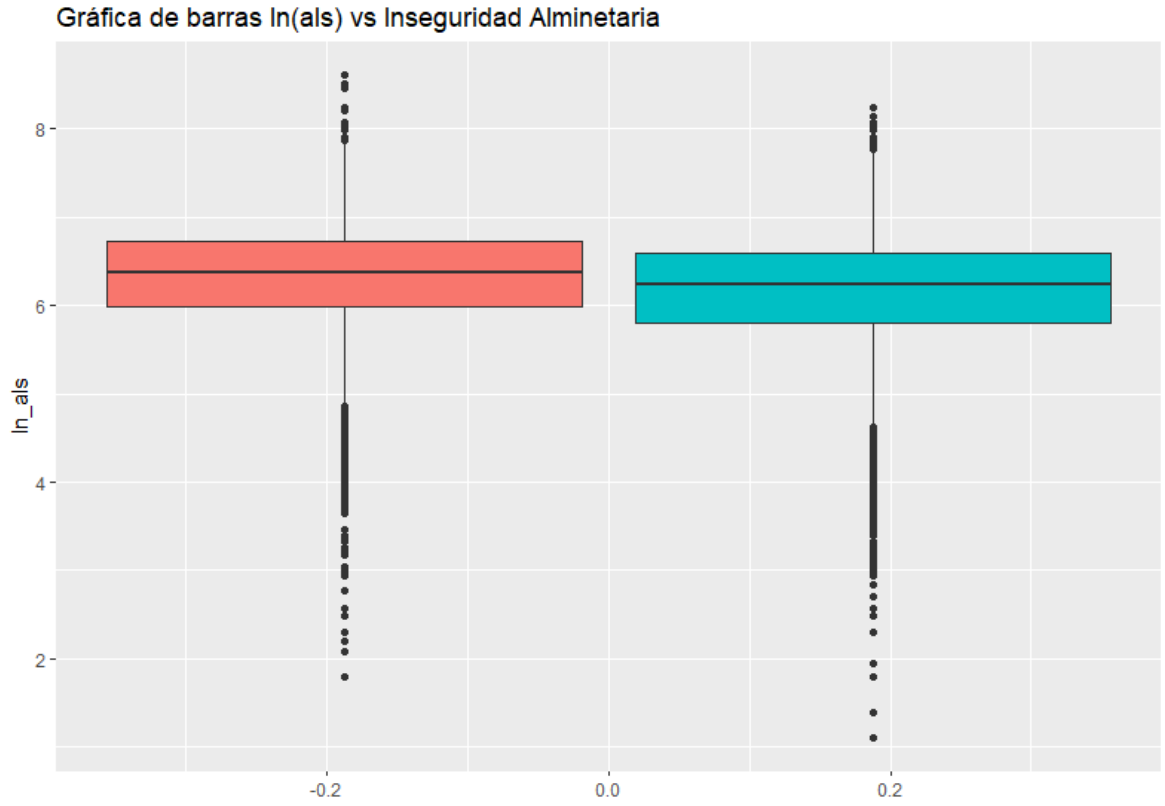
Inseguridad alimetaria	No (N=5853)	Si (N=14427)	Overall (N=20280)
nse5f			
Bajo	499 (8.5%)	3054 (21.2%)	3553 (17.5%)
MedioBajo	761 (13.0%)	3166 (21.9%)	3927 (19.4%)
Medio	989 (16.9%)	3130 (21.7%)	4119 (20.3%)
MedioAlto	1431 (24.4%)	2933 (20.3%)	4364 (21.5%)
Alto	2173 (37.1%)	2144 (14.9%)	4317 (21.3%)
refin			
No	5007 (85.5%)	11414 (79.1%)	16421 (81.0%)
Si	846 (14.5%)	3013 (20.9%)	3859 (19.0%)
ln_alns			
Mean (SD)	4 (1)	4 (1)	4 (1)
Median [Min, Max]	4 [0, 8]	4 [0, 8]	4 [0, 8]
ln_als			
Mean (SD)	6 (0.7)	6 (0.7)	6 (0.7)
Median [Min, Max]	6 [2, 9]	6 [1, 8]	6 [1, 9]
area			
Urbana	4492 (76.7%)	9467 (65.6%)	13959 (68.8%)
Rural	1361 (23.3%)	4960 (34.4%)	6321 (31.2%)
numpeho			
Mean (SD)	4 (2)	4 (2)	4 (2)
Median [Min, Max]	4 [1, 20]	4 [1, 20]	4 [1, 20]
edadjef			
Mean (SD)	50 (20)	50 (20)	50 (20)
Median [Min, Max]	50 [20, 100]	50 [20, 100]	50 [20, 100]

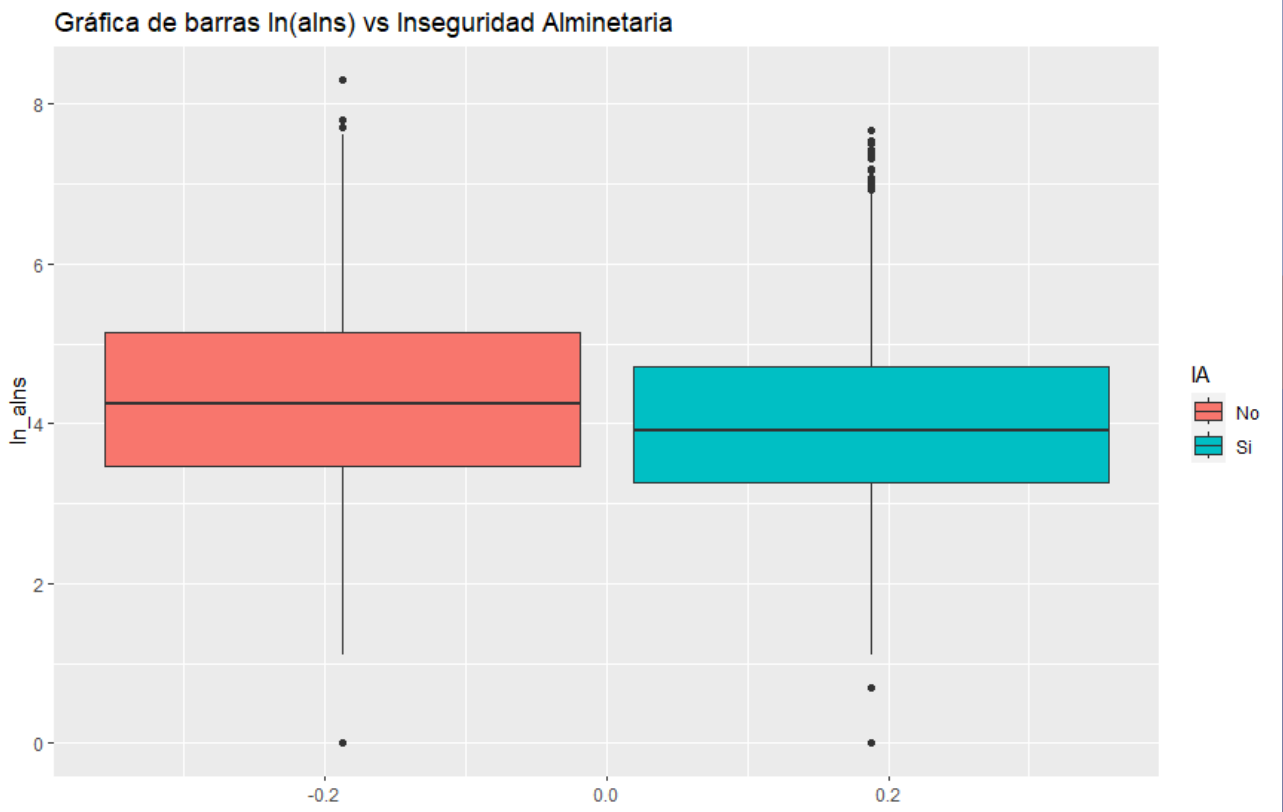
Análisis Descriptivo





Análisis Descriptivo





Cálculo de Probabilidades

- De acuerdo a la tabla de frecuencias del porcentaje de población Mexicana que presenta inseguridad alimentaria vs Nivel Socioeconómico, si se selecciona un grupo de 100 personas, ¿Cuál es la probabilidad de que 70 personas o mas de clase alta, de clase media y de clase baja presente inseguridad alimentaria?

	Bajo (N=3553)	MedioBajo (N=3927)	Medio (N=4119)	MedioAlto (N=4364)	Alto (N=4317)	Overall (N=20280)
IA						
No	499 (14.0%)	761 (19.4%)	989 (24.0%)	1431 (32.8%)	2173 (50.3%)	5853 (28.9%)
Si	3054 (86.0%)	3166 (80.6%)	3130 (76.0%)	2933 (67.2%)	2144 (49.7%)	14427 (71.1%)

P(70 personas o más de clase alta) = 0.0012%

P(70 personas o más de clase media) = 89.91%

P(70 personas o más de clase baja) = 99.26%

Planteamiento de Hipótesis Estadísticas

La mayoría de las personas afirman que los hogares con menor nivel socioeconómico tienden a gastar más en productos no saludables que las personas con mayores niveles socioeconómicos

Planteamiento de hipótesis:

H_o: prom_ln_alns_baja <= prom_ln_alns_alta
H_a: prom_ln_alns_baja > prom_ln_alns_alta

F test to compare two variances

```
data:  df.clean[df.clean$nse5f == "Bajo", "ln_alns"] and
df.clean[df.clean$nse5f == "Alto", "ln_alns"]
F = 0.79317, num df = 3552, denom df = 4316, p-value =
6.199e-13
alternative hypothesis: true ratio of variances is not equal
to 1
95 percent confidence interval:
 0.7449630 0.8446658
sample estimates:
ratio of variances
         0.7931725
```

Welch Two Sample t-test

```
data:  df.clean[df.clean$nse5f == "Bajo", "ln_alns"] and
df.clean[df.clean$nse5f == "Alto", "ln_alns"]
t = -40.863, df = 7819.2, p-value = 1
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.9596575        Inf
sample estimates:
mean of x mean of y
 3.688413  4.610932
```

Planteamiento de Hipótesis Estadísticas

El promedio de gasto en alimentos no saludables en familias que presentan IA es mayor que en familias que no tienen IA.

Planteamiento de hipótesis:

Ho: prom_ln_alns_IA <= prom_ln_alns_NIA
Ha: prom_ln_alns_IA > prom_ln_alns_NIA"

F test to compare two variances

```
data:  df.clean[df.clean$IA == "Si", "ln_alns"] and
df.clean[df.clean$IA == "No", "ln_alns"]
F = 0.89267, num df = 14426, denom df = 5852, p-value = 1.755e-07
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8549641 0.9316744
sample estimates:
ratio of variances
          0.892672
```

Welch Two Sample t-test

```
data:  df.clean[df.clean$IA == "Si", "ln_alns"] and
df.clean[df.clean$IA == "No", "ln_alns"]
t = -18.178, df = 10310, p-value = 1
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.3249486        Inf
sample estimates:
mean of x mean of y
 4.032844  4.330827
```

Planteamiento de Hipótesis Estadísticas

El promedio de gasto en alimentos no saludables en familias cuando el jefe de familia es mujer es mayor que donde el jefe de familia es hombre.

Planteamiento de hipótesis:

Ho: prom_ln_alns_m <= prom_ln_alns_h
Ha: prom_ln_alns_m > prom_ln_alns_h

F test to compare two variances

```
data:  df.clean[df.clean$sexojef == "Mujer", "ln_alns"] and
df.clean[df.clean$sexojef == "Hombre", "ln_alns"]
F = 0.97877, num df = 4392, denom df = 15886, p-value = 0.377
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9338932 1.0264683
sample estimates:
ratio of variances
         0.9787731
```

Two Sample t-test

```
data:  df.clean[df.clean$sexojef == "Mujer", "ln_alns"] and
df.clean[df.clean$sexojef == "Hombre", "ln_alns"]
t = -11.384, df = 20278, p-value = 1
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.2305803        Inf
sample estimates:
mean of x mean of y
 3.961017  4.162486
```

Modelo de Regresión Logístico

- Modelo para determinar factores sobre Inseguridad Alimentaria

Dado que la variable dependiente es cualitativa y el número de predictores es amplio se usa un modelo de regresión logística múltiple.

Modelo 1 (todas las variables)

La variable ln_als, no se tomara en cuenta debido a que tiene muchos outliers, lo que puede complicar la creación de el modelo. Además de que esta variable es asimétrica.

Modelo de Regresión Logístico

- Las Bi contrastan la siguiente hipótesis

  Ho: Bi = 0
  Ha: Bi =! 0

Si p-value >= significancia - no rechazo Ho
Si p-value < significancia - rechazo Ho

Se tomará una significancia de 0.05

A cualquier nivel de confianza existe evidencia estadística para no rechazar hipótesis nula por lo tanto se determina edadjef como variable no significativa.

```
Call:
glm(formula = IA ~ nse5f + area + numpeho + refin + edadjef +
    sexojef + añosedu + ln_alns, family = binomial, data = df.clean)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6481  -1.0554   0.6106   0.8053   1.6718

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.944211   0.128332  15.150  < 2e-16 ***
nse5fMedioBajo -0.320961   0.064394  -4.984 6.22e-07 ***
nse5fMedio     -0.555149   0.063512  -8.741  < 2e-16 ***
nse5fMedioAlto -0.932981   0.063743 -14.637  < 2e-16 ***
nse5fAlto      -1.538133   0.068077 -22.594  < 2e-16 ***
areaRural      -0.081577   0.041089  -1.985 0.047104 *
numpeho         0.165837   0.010129  16.373  < 2e-16 ***
refinSi         0.388896   0.044613   8.717  < 2e-16 ***
edadjef         0.001223   0.001238   0.988 0.323317
sexojefMujer    0.146485   0.041319   3.545 0.000392 ***
añosedu        -0.052456   0.004499 -11.659  < 2e-16 ***
ln_alns        -0.109353   0.016499  -6.628 3.41e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24373  on 20279  degrees of freedom
Residual deviance: 22128  on 20268  degrees of freedom
AIC: 22152

Number of Fisher Scoring iterations: 4
```

Modelo de Regresión Logístico

Modelo 2

Beta (area)
p-value = 0.049 < significancia (0.05)
A un nivel de confianza de 95% existe evidencia estadística para rechazar hipótesis nula.

Por lo tanto se determina la variable área como significativa.

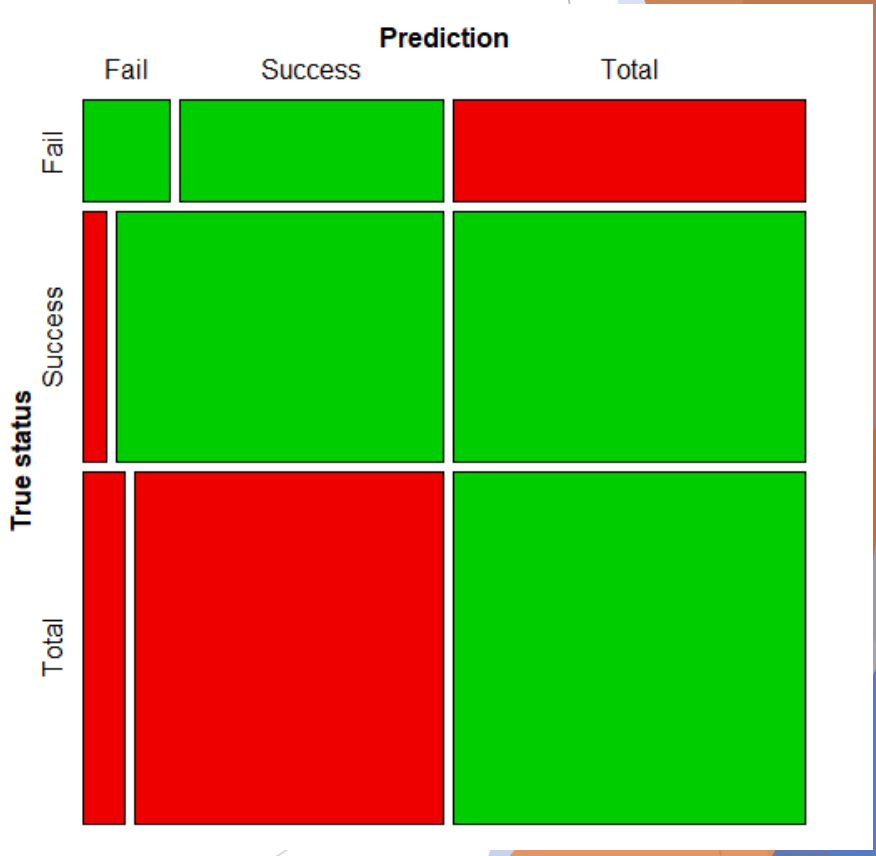
```
Call:
glm(formula = IA ~ nse5f + area + numpeho + refin + sexojef +
    añosedu + ln_alns, family = binomial, data = df.clean)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6395  -1.0554   0.6103   0.8050   1.6586

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     2.026193   0.097946  20.687  < 2e-16 ***
nse5fMedioBajo -0.318853   0.064358  -4.954 7.26e-07 ***
nse5fMedio     -0.551218   0.063387  -8.696  < 2e-16 ***
nse5fMedioAlto -0.925194   0.063254 -14.627  < 2e-16 ***
nse5fAlto      -1.523028   0.066326 -22.963  < 2e-16 ***
areaRural      -0.080633   0.041078  -1.963 0.049657 *
numpeho         0.164207   0.009978  16.457  < 2e-16 ***
refinSi         0.387753   0.044600   8.694  < 2e-16 ***
sexojefMujer    0.148135   0.041281   3.588 0.000333 ***
añosedu        -0.054213   0.004134 -13.114  < 2e-16 ***
ln_alns        -0.110618   0.016450  -6.725 1.76e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24373  on 20279  degrees of freedom
Residual deviance: 22129  on 20269  degrees of freedom
AIC: 22151

Number of Fisher Scoring iterations: 4
```

Modelo de Regresión Logístico

- Comparación de clasificación predicha y observaciones

Para este estudio se va a emplear un threshold de 0.5

Prediction			
True status	Fail	Success	Total
Fail	1461	4392	5853
Success	1015	13412	14427
Total	2476	17804	20280



Precisión
(verdadero.negativos + verdadero.positivo) / total.muestra
0.7333 (73.33%)

Porcentaje de falsos negativos
falso.negativo / (falso.negativo + verdadero.positivo)
0.0703 (7.03 %)

Conclusiones

1) Mientras mayor sea la clase social se tiene menos probabilidad de presentar IA
2) La probabilidad de IA en área rural es menor que en área urbana
3) A mayor número de personas en el hogar, mayor probabilidad de presentar IA
4) Hay mayor probabilidad de presentar IA si las familias cuentan con recursos financieros adicionales o si la mujer funge como jefe de familia
5) Hay menor probabilidad de presentar IA si el jefe de familia cuenta con mas años de educación o si aumentamos el gasto en alimentos no saludable.

Link al GitHub

https://github.com/humbertogmtz/bedu-m2-r-team14

No hubo problemas al hacer el código