

BITÁCORA 3 - GRUPO 2

Integrantes

- Cristhian Jimenez Campos C33973
- Olman Camacho Jerez C31523
- Jose Manuel Alfaro Monge C30244

Bitácora 1

Pregunta de investigación

¿Cuáles son los factores individuales, sociales y académicos que más inciden en la predicción de la deserción escolar en estudiantes de nivel secundario?

Objeto de estudio

El objeto de estudio es el fenómeno de la deserción escolar, entendido como el abandono prematuro del sistema educativo formal por parte de los estudiantes antes de completar el nivel educativo correspondiente.

Conceptos

Deserción escolar:

Según Spady, citado por Floricely Dzay Chulim (2012), se mencionan dos definiciones operacionales acerca de la deserción universitaria:

- a) Incluye a cualquier persona que abandona la institución de educación superior donde se encuentra registrada.
- b) Se refiere a aquellas personas que no reciben un título o grado de ninguna universidad.

Esta segunda definición sostiene que el o la estudiante que haya iniciado un proceso de aprendizaje superior y, en un determinado periodo de tiempo, no haya obtenido su respectivo título o grado, puede considerarse un desertor.

Factores personales:

Se entienden como factores personales todas aquellas características internas del estudiante, como la motivación, las actitudes y las habilidades cognitivas, que influyen directamente en su aprendizaje.

Factores sociales:

Los factores sociales se refieren a todas aquellas influencias externas relacionadas con el entorno del estudiante, que se presentan en su vida de manera inesperada o no planificada, y que provienen directamente de su círculo personal (familiares, amigos cercanos, entre otros). Estas influencias pueden afectar directamente su rendimiento académico.

Teorías

Teoría de la frustración:

Esta teoría es un aporte del investigador Abram Amsel. Según Alejandro Baquero (2007), la Teoría de la Frustración de Amsel expone una hipótesis acerca de la función de la omisión decepcionante de una recompensa en circunstancias de refuerzo no continuo.

De acuerdo con esta teoría, en la etapa de adquisición, el sujeto aprende a prever la recompensa obtenida en el contexto experimental debido a la presencia de claves contextuales que la anuncian. Luego, cuando la recompensa se omite inesperadamente, el sujeto experimenta una reacción emocional natural y aversiva denominada **frustración**, la cual llega a asociarse con las señales que previamente indicaban una recompensa.

Esto genera un conflicto al inicio del entrenamiento, ya que tanto la frustración como la recompensa son pronosticadas por condicionamiento clásico ante los mismos estímulos. A medida que avanza el entrenamiento, por efecto del contracondicionamiento, el conflicto se resuelve a favor de la respuesta, ya que el refuerzo no es predecible en una situación de refuerzo incompleto. De esta forma, la respuesta continúa incluso durante la extinción, puesto que se ha condicionado la expectativa de ausencia de recompensa. En cambio, los sujetos que reciben refuerzo constante no desarrollan este tipo de respuesta al no experimentar la frustración asociada a la omisión del refuerzo.

Teoría de los factores múltiples

La teoría de los factores múltiples sostiene que la deserción escolar responde a una combinación de factores individuales, sociales y académicos.

Basándonos en la base de datos con la que se trabajará, algunos de estos factores son el desempeño académico, el nivel educativo y los factores familiares, los cuales pueden aumentar el riesgo de que un alumno abandone la institución educativa.

La deserción no se ve afectada por un único factor; generalmente se debe a la interacción de múltiples elementos y, en ocasiones, es difícil identificar todas las causas. Es fundamental detectar la deserción académica a tiempo para poder brindar apoyo mediante tutorías, asistencia económica y programas de acompañamiento psicológico, los cuales pueden ayudar a prevenir el abandono. Además, es importante identificar qué factores tienen mayor peso en la decisión de desertar.

Teoría de los factores económicos y apoyo familiar en la deserción universitaria

Esta teoría, sustentada en diversos estudios, señala que las dificultades económicas, tanto familiares como individuales, tienen un impacto significativo en los estudiantes, aumentando la probabilidad de abandono universitario.

Los problemas económicos generan una gran presión sobre los estudiantes, quienes en muchos casos deben colaborar con el sustento familiar, lo que puede llevarlos a dejar de lado sus estudios.

Para aliviar esta presión, el acceso a becas y ayudas financieras se identifica como un factor protector, ya que brinda apoyo económico y emocional, permitiendo que los estudiantes continúen su formación y mejoren su rendimiento académico.

Bibliografía

Miranda, J., & Vázquez, J. (2023). *Análisis de la deserción estudiantil universitaria desde una perspectiva analítica*. Revista N° 24, Facultad de Economía y Negocios. Recuperado de https://admissionfen.cl/wp-content/uploads/2024/03/Revista-N%C2%B024_jaime_miranda.pdf

Poveda Velasco, I. M. (2019). *Los factores que influyen sobre la deserción universitaria: Estudio en la UMRPSFXCh* [Investigación y Negocios, 12(20), 63–80]. Investigación y Negocios. https://www.scielo.org.bo/scielo.php?script=sci_arttext&pid=S2521-27372019000200007

[Autor(es)]. (2021). *Título del artículo*. Revista de Educación (o nombre completo), volumen (número), páginas. Recuperado de https://www.scielo.sa.cr/scielo.php?script=sci_arttext&pid=S2215-34702021000100019

[Autor(es)]. (2023). *Título del artículo*. Revista (o nombre de la revista), volumen (número), páginas. Recuperado de http://www.scielo.edu.uy/scielo.php?script=sci_arttext&pid=S1688-93042023000301206#B56

[Autor(es)]. (2014). *Título del artículo*. Revista (o nombre de la revista), volumen (número), páginas. Recuperado de http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S2145-94442014000200010

[Autor(es)]. (Año). *Título del artículo*. Revista Revie, volumen (número), páginas. Recuperado de <https://revie.gob.do/index.php/revie/article/view/193/365>

[Autor(es)]. (Año). *Título del artículo*. Revista RIDE, volumen (número), páginas. Recuperado de <https://www.ride.org.mx/index.php/RIDE/article/view/1923/5020>

Dzay Chulim, F., & Narváez Trejo, O. M. (2012). *La deserción escolar desde la perspectiva estudiantil*. La Editorial Manda. <https://www.uv.mx/personal/onarvaez/files/2013/02/la-desercion-escolar.pdf>

Bitácora 2

Datos

Características de la tabla:

Esta base de datos contiene registros de 4,424 estudiantes, quienes serán clasificados de distintas maneras, desde su estado civil hasta los cursos que están cursando, entre otros. La base de datos fue publicada en 2021 y presenta diversos factores, incluyendo variables relacionadas con los padres, para analizar si influyen en la vida académica del estudiante. Fue creada por Valentim Realinho, Mónica Vieira Martins, Jorge Machado y Luís Baptista, investigadores del Instituto Politécnico de Portalegre en Portugal, y puede ser descargada desde el enlace [link](#). Los datos corresponden al segundo semestre, aunque no se especifica el año.

Las variables están organizadas en distintas categorías: variables relacionadas con la trayectoria académica, variables demográficas y variables socioeconómicas. Los tipos de datos incluyen variables continuas, categóricas y enteras.

Población de estudio:

Estudiantes matriculados en diferentes carreras de pregrado de una institución de educación superior.

Muestra observada:

4,424 estudiantes.

Unidad estadística o individuos:

Cada uno de los 4,424 estudiantes de educación superior durante determinados semestres.

Identificación de las variables de estudio:

Las variables de estudio incluyen información sobre la trayectoria académica, datos demográficos y factores socioeconómicos de los estudiantes, así como su rendimiento académico al final del primer y segundo semestre. El problema se plantea como una tarea de clasificación en tres categorías: abandono, matriculado y graduado.

Primeras 5 filas de la tabla de datos:

```
library(dplyr)
library(ggplot2)
library(tidyr)
datos <- read.csv2("data.csv", sep = ";", header = TRUE, stringsAsFactors = FALSE)
datos <- datos %>%
  select(-GDP) ### El uso de esta variable es irrelevante
head(datos, 5)
```

	Marital.status	Application.mode	Application.order	Course
1	1	17	5	171
2	1	15	1	9254
3	1	1	5	9070
4	1	17	2	9773
5	2	39	1	8014

Daytime.evening.attendance. Previous.qualification

1	1	1
2	1	1
3	1	1
4	1	1
5	0	1

Previous.qualification..grade. Nacionality Mother.s.qualification

1	122.0	1	19
2	160.0	1	1
3	122.0	1	37
4	122.0	1	38
5	100.0	1	37

Father.s.qualification Mother.s.occupation Father.s.occupation

1	12	5	9
2	3	3	3
3	37	9	9
4	37	5	3
5	38	9	9

Admission.grade Displaced Educational.special.needs Debtor

1	127.3	1	0	0
2	142.5	1	0	0
3	124.8	1	0	0
4	119.6	1	0	0
5	141.5	0	0	0

Tuition.fees.up.to.date Gender Scholarship.holder Age.at.enrollment

1	1	1	0	20
2	0	1	0	19
3	0	1	0	19
4	1	0	0	20
5	1	0	0	45

International Curricular.units.1st.sem..credited.

1	0	0
2	0	0
3	0	0
4	0	0
5	0	0

Curricular.units.1st.sem..enrolled. Curricular.units.1st.sem..evaluations.

1	0	0
2	6	6
3	6	0
4	6	8
5	6	9

Curricular.units.1st.sem..approved. Curricular.units.1st.sem..grade.

1	0	0.0
2	6	14.0
3	0	0.0
4	6	13.428571428571429
5	5	12.333333333333334

Curricular.units.1st.sem..without.evaluations.

1	0
2	0
3	0
4	0
5	0

Curricular.units.2nd.sem..credited. Curricular.units.2nd.sem..enrolled.

1	0	0
2	0	6
3	0	6
4	0	6
5	0	6

	Curricular.units.2nd.sem..evaluations.	Curricular.units.2nd.sem..approved.
1	0	0
2	6	6
3	0	0
4	10	5
5	6	6
	Curricular.units.2nd.sem..grade.	
1	0.0	
2	13.666666666666666	
3	0.0	
4	12.4	
5	13.0	
	Curricular.units.2nd.sem..without.evaluations.	Unemployment.rate
1	0	10.8
2	0	13.9
3	0	10.8
4	0	9.4
5	0	13.9
	Inflation.rate	Target
1	1.4	Dropout
2	-0.3	Graduate
3	1.4	Dropout
4	-0.8	Graduate
5	-0.3	Graduate

La tabla se encuentra en formato tabular, esto se puede ver y también se comenta en la página de descarga

Resumen de 5 números de las variables cuantitativas y analizar el mismo:

```
library(dplyr)
# Se selecciona las variables cuantitativas
variables_cuantitativas <- select_if(datos, is.numeric)
# Calcular resumen de 5 números para cada variable
#'Vamos a usar sapply para aplicar la funcion fivenum a la base
#'el firenum es una funcion que nos ayuda a calcular el minimo y maximo, los Q1 y Q3, ad
resumen_5_numeros <- sapply(variables_cuantitativas, fivenum)
#para no tener problema trasponemos a resumen_5_numeros
resumen_5_numeros <- t(resumen_5_numeros)
#'Para facilitar la lectura vamos a ponerle nosmbres claros a las columnas
colnames(resumen_5_numeros) <- c("Minimo","Q1","Mediana","Q3","Máximo")
print(resumen_5_numeros)
```

	Minimo	Q1	Mediana	Q3	Máximo
Marital.status	1	1	1	1	6
Application.mode	1	1	17	39	57
Application.order	0	1	1	2	9
Course	33	9085	9238	9556	9991
Daytime.evening.attendance.	0	1	1	1	1
Previous.qualifikation	1	1	1	1	43
Nacionality	1	1	1	1	109
Mother.s.qualifikation	1	2	19	37	44
Father.s.qualifikation	1	3	19	37	44
Mother.s.occupation	0	4	5	9	194
Father.s.occupation	0	4	7	9	195
Displaced	0	0	1	1	1
Educational.special.needs	0	0	0	0	1

Debtor	0	0	0	0	1
Tuition.fees.up.to.date	0	1	1	1	1
Gender	0	0	0	1	1
Scholarship.holder	0	0	0	0	1
Age.at.enrollment	17	19	20	25	70
International	0	0	0	0	1
Curricular.units.1st.sem..credited.	0	0	0	0	20
Curricular.units.1st.sem..enrolled.	0	5	6	7	26
Curricular.units.1st.sem..evaluations.	0	6	8	10	45
Curricular.units.1st.sem..approved.	0	3	5	6	26
Curricular.units.1st.sem..without.evaluations.	0	0	0	0	12
Curricular.units.2nd.sem..credited.	0	0	0	0	19
Curricular.units.2nd.sem..enrolled.	0	5	6	7	23
Curricular.units.2nd.sem..evaluations.	0	6	8	10	33
Curricular.units.2nd.sem..approved.	0	2	5	6	20
Curricular.units.2nd.sem..without.evaluations.	0	0	0	0	12

La tabla muestra el resumen de cinco números de las variables cuantitativas de nuestra base de datos. En ella se puede observar que la mayoría de las variables tienen un valor mínimo de 0. La forma en que se evalúan nuestras variables es diferente, ya que cada una tiene rangos distintos para su asignación; por lo tanto, pueden ser continuas o presentar saltos en los valores, como pasar de 33 a 53, entre otros.

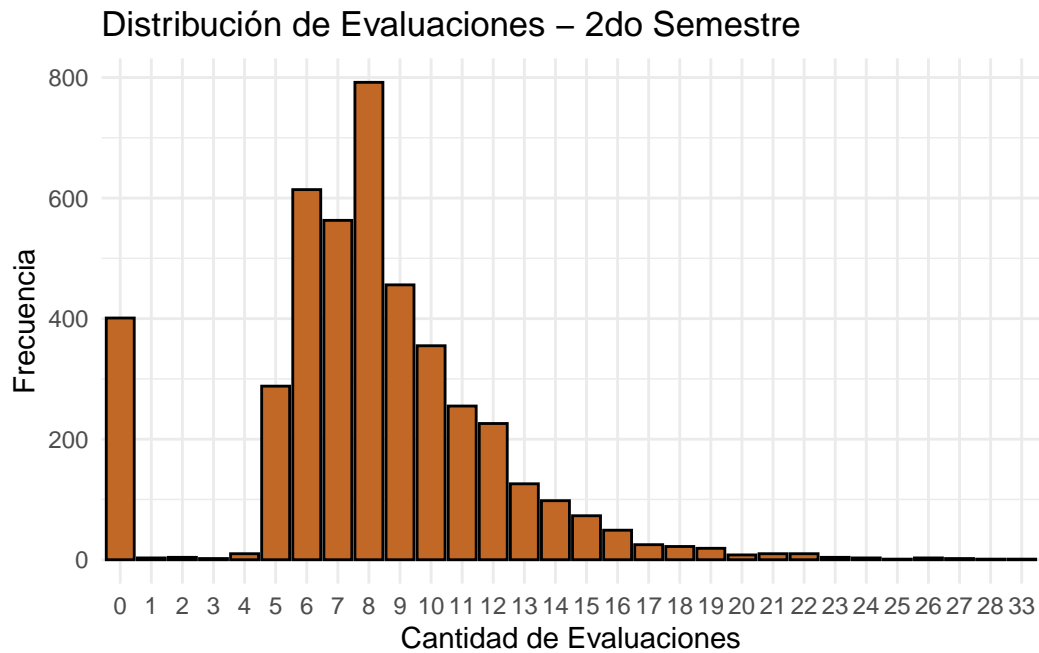
Hacer al menos un gráfico que describa la distribución para cada una de las variables cuantitativas:

```
ggplot(datos, aes(x = factor(Curricular.units.2nd.sem..evaluations.))) +
  geom_bar(fill = "#C26827", color = "#000000") +
  labs(
```

```

title = "Distribución de Evaluaciones - 2do Semestre",
x = "Cantidad de Evaluaciones",
y = "Frecuencia"
) +
theme_minimal()

```



Hacer al menos dos gráficos que describan la relación entre las variables:

```

grafico_internacional_genero <- function(datos, genero, internacional, pais) {
  df_plot <- datos %>%
    mutate(
      genero = factor({{genero}}, levels = c(1, 0), labels = c("Hombres", "Mujeres")),
      tipo = factor({{internacional}}, levels = c(0, 1), labels = c("Nacional", "Internacional")),
    ) %>%
    group_by(genero, tipo) %>%

```

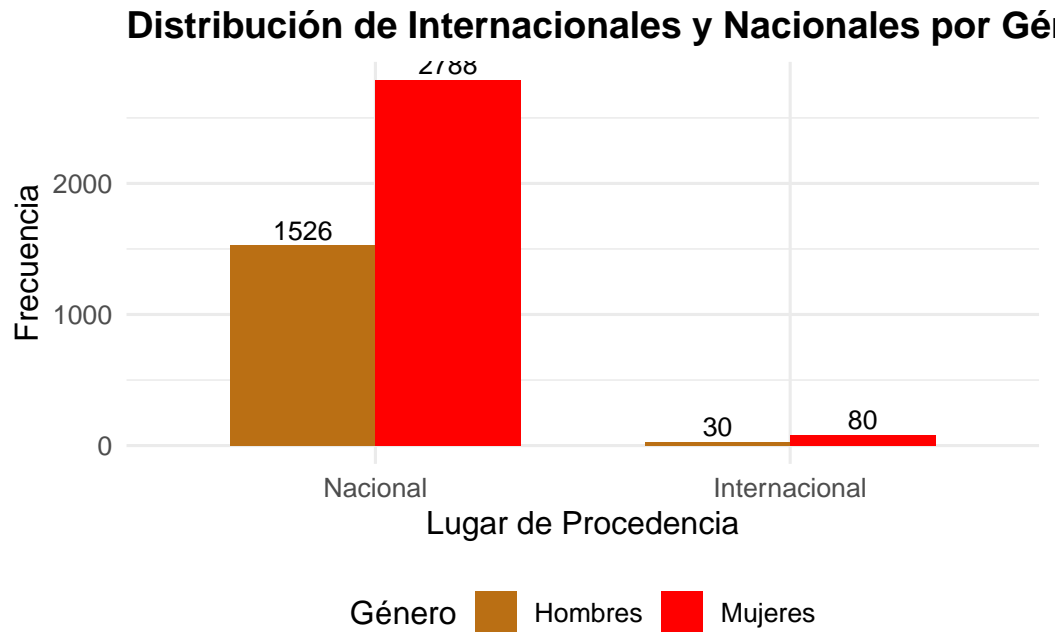
```

    summarise(Frecuencia = n(), .groups = "drop")

ggplot(df_plot, aes(x = tipo, y = Frecuencia, fill = genero)) +
  geom_col(position = "dodge", width = 0.7) +
  geom_text(aes(label = Frecuencia),
            position = position_dodge(width = 0.7),
            vjust = -0.3, size = 3.5) +
  scale_fill_manual(values = c("#BA6F14", "#FF0000")) +
  labs(
    title = "Distribución de Internacionales y Nacionales por Género",
    x = "Lugar de Procedencia",
    y = "Frecuencia",
    fill = "Género"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    legend.position = "bottom",
    plot.title = element_text(face = "bold"),
    axis.text.x = element_text(angle = 0, hjust = 0.5)
  )
}

grafico_internacional_genero(datos, datos$Gender, datos$International, datos$Nationality

```



```

gráfico_necesidad_genero <- function(df, genero, needs){
  df_plot <- datos %>%
    mutate(
      genero = factor({{genero}}, levels = c(1, 0), labels = c("Hombres", "Mujeres")),
      necesidades = factor({{needs}}, levels = c(0, 1),
        labels = c("Sin necesidades especiales", "Con necesidades especiales"))
    ) %>%
    group_by(genero, necesidades) %>%
    summarise(Frecuencia = n(), .groups = "drop")

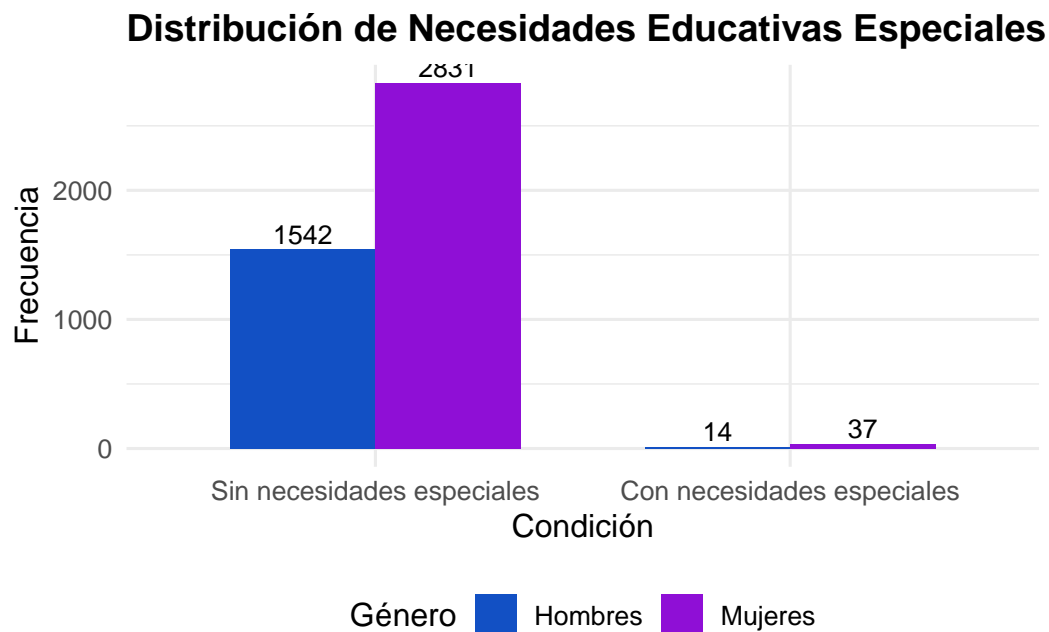
  ggplot(df_plot, aes(x = necesidades, y = Frecuencia, fill = genero)) +
    geom_col(position = "dodge", width = 0.7) +
    geom_text(aes(label = Frecuencia),
      position = position_dodge(width = 0.7),
      vjust = -0.3, size = 3.5) +
    scale_fill_manual(values = c("#1250C4", "#8F0FD6")) +
    labs(

```

```

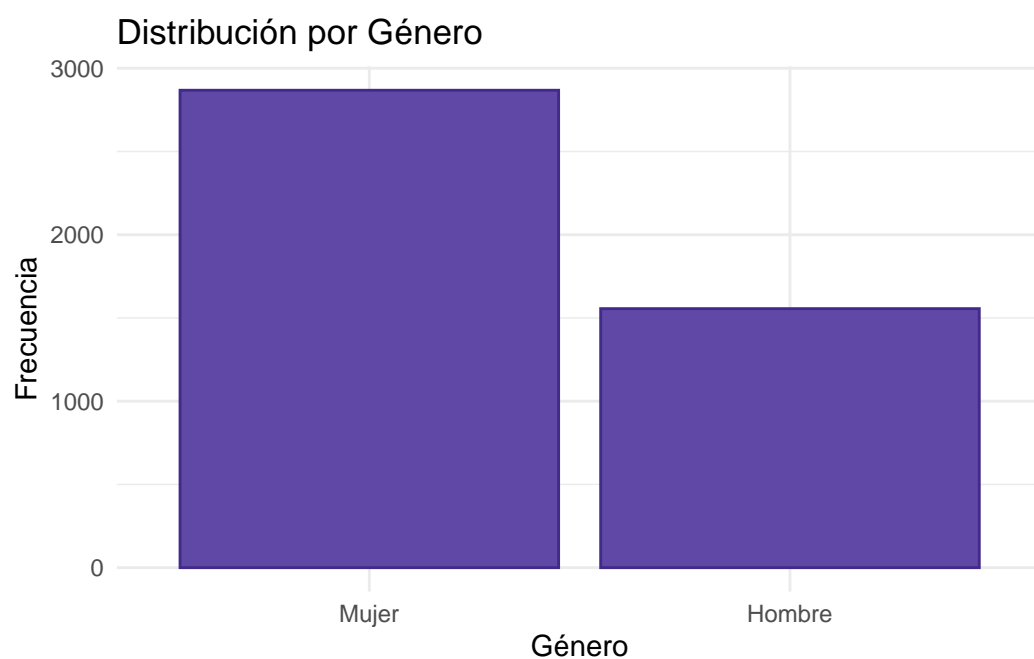
    title = "Distribución de Necesidades Educativas Especiales por Género",
    x = "Condición",
    y = "Frecuencia",
    fill = "Género"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    legend.position = "bottom",
    plot.title = element_text(face = "bold"),
    axis.text.x = element_text(angle = 0, hjust = 0.5)
  )
}
gráfico_necesidad_genero(datos, datos$Gender, datos$Educational.special.needs)

```



Hacer al menos un gráfico que muestre la distribución de las variables categóricas:

```
ggplot(datos, aes(x = factor(Gender,  
                           levels = c(0, 1),  
                           labels = c("Mujer", "Hombre")))) +  
geom_bar(fill = "#6049A6", color = "#42298A") +  
labs(  
  title = "Distribución por Género",  
  x = "Género",  
  y = "Frecuencia"  
) +  
theme_minimal()
```



Identificar valores faltantes y posibles outliers:

```
datos_faltantes <- datos %>%  
  filter(if_any(everything(), is.na))  
head(datos_faltantes)
```

```
[1] Marital.status  
[2] Application.mode  
[3] Application.order  
[4] Course  
[5] Daytime.evening.attendance.  
[6] Previous.qualification  
[7] Previous.qualification..grade.  
[8] Nationality  
[9] Mother.s.qualification  
[10] Father.s.qualification  
[11] Mother.s.occupation  
[12] Father.s.occupation  
[13] Admission.grade  
[14] Displaced  
[15] Educational.special.needs  
[16] Debtor  
[17] Tuition.fees.up.to.date  
[18] Gender  
[19] Scholarship.holder  
[20] Age.at.enrollment  
[21] International  
[22] Curricular.units.1st.sem..credited.  
[23] Curricular.units.1st.sem..enrolled.  
[24] Curricular.units.1st.sem..evaluations.
```

```

[25] Curricular.units.1st.sem..approved.
[26] Curricular.units.1st.sem..grade.
[27] Curricular.units.1st.sem..without.evaluations.
[28] Curricular.units.2nd.sem..credited.
[29] Curricular.units.2nd.sem..enrolled.
[30] Curricular.units.2nd.sem..evaluations.
[31] Curricular.units.2nd.sem..approved.
[32] Curricular.units.2nd.sem..grade.
[33] Curricular.units.2nd.sem..without.evaluations.
[34] Unemployment.rate
[35] Inflation.rate
[36] Target
<0 rows> (o 0- extensión row.names)

```

```

datos %>%
  summarise(
    across(
      where(is.numeric),
      ~sum(
        .<quantile(.,0.25,na.rm=TRUE)-1.5*IQR(.)|
        .>quantile(.,0.75,na.rm=TRUE)+1.5*IQR(.),na.rm = TRUE
      )
    )
  )#cantidad de outliers por variable

```

	Marital.status	Application.mode	Application.order	Course
1	505	0	541	442

	Daytime.evening.attendance.	Previous.qualification	Nacionality
1	483	707	110

	Mother.s.qualification	Father.s.qualification	Mother.s.occupation	
1	0	0	182	
	Father.s.occupation	Displaced	Educational.special.needs	Debtor
1	177	0	51	503
	Tuition.fees.up.to.date	Gender	Scholarship.holder	Age.at.enrollment
1	528	0	1099	441
	International	Curricular.units.1st.sem..credited.		
1	110		577	
	Curricular.units.1st.sem..enrolled.	Curricular.units.1st.sem..evaluations.		
1		424		158
	Curricular.units.1st.sem..approved.			
1		180		
	Curricular.units.1st.sem..without.evaluations.			
1		294		
	Curricular.units.2nd.sem..credited.	Curricular.units.2nd.sem..enrolled.		
1		530		369
	Curricular.units.2nd.sem..evaluations.	Curricular.units.2nd.sem..approved.		
1		109		44
	Curricular.units.2nd.sem..without.evaluations.			
1		282		

```

es_outlier <- function(x) {

  if(!is.numeric(x)) return(rep(FALSE,length(x)))
  q1 <- quantile(x,0.25,na.rm = TRUE)
  q3 <- quantile(x,0.75,na.rm = TRUE)

  resultado <- x<(q1-1.5*IQR(x,na.rm = TRUE))|
               x>(q3+1.5*IQR(x,na.rm = TRUE))
  return(resultado)
}

```

```
}
```

```
outliers <- as.data.frame(sapply(datos,es_outlier))  
head(outliers)
```

	Marital.status	Application.mode	Application.order	Course
1	FALSE	FALSE	TRUE	TRUE
2	FALSE	FALSE	FALSE	FALSE
3	FALSE	FALSE	TRUE	FALSE
4	FALSE	FALSE	FALSE	FALSE
5	TRUE	FALSE	FALSE	TRUE
6	TRUE	FALSE	FALSE	FALSE

	Daytime.evening.attendance.	Previous.qualification
1	FALSE	FALSE
2	FALSE	FALSE
3	FALSE	FALSE
4	FALSE	FALSE
5	TRUE	FALSE
6	TRUE	TRUE

	Previous.qualification..grade.	Nacionality	Mother.s.qualification
1	FALSE	FALSE	FALSE
2	FALSE	FALSE	FALSE
3	FALSE	FALSE	FALSE
4	FALSE	FALSE	FALSE
5	FALSE	FALSE	FALSE
6	FALSE	FALSE	FALSE

	Father.s.qualification	Mother.s.occupation	Father.s.occupation
1	FALSE	FALSE	FALSE
2	FALSE	FALSE	FALSE
3	FALSE	FALSE	FALSE

4	FALSE	FALSE	FALSE
5	FALSE	FALSE	FALSE
6	FALSE	FALSE	FALSE

Admission.grade Displaced Educational.special.needs Debtor

1	FALSE	FALSE	FALSE	FALSE
2	FALSE	FALSE	FALSE	FALSE
3	FALSE	FALSE	FALSE	FALSE
4	FALSE	FALSE	FALSE	FALSE
5	FALSE	FALSE	FALSE	FALSE
6	FALSE	FALSE	FALSE	TRUE

Tuition.fees.up.to.date Gender Scholarship.holder Age.at.enrollment

1	FALSE	FALSE	FALSE	FALSE
2	TRUE	FALSE	FALSE	FALSE
3	TRUE	FALSE	FALSE	FALSE
4	FALSE	FALSE	FALSE	FALSE
5	FALSE	FALSE	FALSE	TRUE
6	FALSE	FALSE	FALSE	TRUE

International Curricular.units.1st.sem..credited.

1	FALSE	FALSE
2	FALSE	FALSE
3	FALSE	FALSE
4	FALSE	FALSE
5	FALSE	FALSE
6	FALSE	FALSE

Curricular.units.1st.sem..enrolled. Curricular.units.1st.sem..evaluations.

1	TRUE	FALSE
2	FALSE	FALSE
3	FALSE	FALSE
4	FALSE	FALSE
5	FALSE	FALSE

6	FALSE	FALSE
Curricular.units.1st.sem..approved. Curricular.units.1st.sem..grade.		
1	FALSE	FALSE
2	FALSE	FALSE
3	FALSE	FALSE
4	FALSE	FALSE
5	FALSE	FALSE
6	FALSE	FALSE
Curricular.units.1st.sem..without.evaluations.		
1	FALSE	
2	FALSE	
3	FALSE	
4	FALSE	
5	FALSE	
6	FALSE	
Curricular.units.2nd.sem..credited. Curricular.units.2nd.sem..enrolled.		
1	FALSE	TRUE
2	FALSE	FALSE
3	FALSE	FALSE
4	FALSE	FALSE
5	FALSE	FALSE
6	FALSE	FALSE
Curricular.units.2nd.sem..evaluations. Curricular.units.2nd.sem..approved.		
1	FALSE	FALSE
2	FALSE	FALSE
3	FALSE	FALSE
4	FALSE	FALSE
5	FALSE	FALSE
6	TRUE	FALSE
Curricular.units.2nd.sem..grade.		

1	FALSE
2	FALSE
3	FALSE
4	FALSE
5	FALSE
6	FALSE

Curricular.units.2nd.sem..without.evaluations. Unemployment.rate

1	FALSE	FALSE
2	FALSE	FALSE
3	FALSE	FALSE
4	FALSE	FALSE
5	FALSE	FALSE
6	TRUE	FALSE

Inflation.rate Target

1	FALSE	FALSE
2	FALSE	FALSE
3	FALSE	FALSE
4	FALSE	FALSE
5	FALSE	FALSE
6	FALSE	FALSE

Investigar técnicas que permitan subsanar los valores perdidos y outliers

Manejo de valores faltantes y outliers

Valores faltantes:

Los valores faltantes pueden ser tratados mediante diferentes métodos que modifican los datos de distinta manera. Por eso, a la hora de utilizarlos, sus resultados se verán afectados de mayor o menor forma dependiendo de la técnica utilizada. Existen técnicas como la eliminación de los

datos faltantes, sin embargo, esta práctica puede perjudicar el análisis de los datos, por lo que nos centraremos en los métodos que buscan modificarlos de manera que aporten información.

Se pueden utilizar métodos simples como la imputación por la media, mediana o moda, que son fáciles de implementar; sin embargo, estos datos pueden verse distorsionados por la presencia de valores atípicos. Por eso existen métodos más robustos como la imputación por vecinos más cercanos, imputación por regresión, imputación múltiple o imputación por series de tiempo. Estas técnicas conservan de mejor manera la estructura de los datos.

Outliers:

Hay tres opciones posibles a la hora de enfrentar un outlier, tanto univariable como multivariable. Estas opciones son eliminarlo, mantenerlo o reemplazarlo. Al igual que con los valores faltantes, se pueden eliminar todos los outliers, pero este método provoca pérdida de información que normalmente es valiosa y aporta a la investigación.

También se tiene la opción de utilizar métodos simples como reemplazarlos por la media o la moda, aunque el uso de estas técnicas no es recomendado porque puede generar resultados artificiales y sesgados. Por lo tanto, se recomienda utilizar técnicas más robustas que reduzcan su influencia sin eliminarlos por completo.

Ahora veremos algunas formas de enfrentar los outliers dependiendo de si son univariantes o multivariantes.

Outliers univariable

Detectar y clasificar: Se realiza una revisión manual para determinar si son errores, casos aleatorios o datos de interés que deberían analizarse por separado. Luego se decide individualmente qué hacer con cada uno.

Winsorización: Se sustituye el dato por el valor del percentil más cercano. Esto conserva la estructura de los datos y limita los efectos sobre las estadísticas principales.

Outliers multivariable

Revisión contextual: Se verifica que las variables que se estén comparando no tengan una relación incoherente o anómala que genere valores extraños o fuera de contexto.

Análisis por separado: Se analizan los casos atípicos como un conjunto separado de los datos comunes para identificar patrones sin alterar los datos principales.

Regresión: Son métodos que consisten en modelar la relación de una variable dependiente con otras independientes, para evaluar la influencia de los outliers sobre las relaciones entre variables.

Bibliografía:

Rosas, J. F. M. (2009). *Métodos de imputación para el tratamiento de datos faltantes*. Revista de Métodos Cuantitativos para la Economía y la Empresa. Recuperado de <https://www.redalyc.org/pdf/2331/23311722>

Medina, F., & Galván, M. (2007). *Imputación de datos: Teoría y práctica*. Serie Estudios Estadísticos y Prospectivos, N.º 54. Santiago de Chile: ONU-CEPAL. <https://repositorio.cepal.org/server/api/core/bitstreams/02dd479f-fae2-43c4-b5ec-5419fa7f6190/content>

Rosas, J. F. M., & Álvarez Verdejo, E. (2009). *Métodos de imputación para el tratamiento de datos faltantes: Aplicación mediante R/S-PLUS*. Revista de Métodos Cuantitativos para la Economía y la Empresa, (07), 03-30. <https://www.redalyc.org/pdf/2331/233117228001.pdf>

Cousineau, D. (2010). *Outliers detection and treatment: A review*. International Journal of Psychological Research, 3(1), 58-67. <https://www.redalyc.org/pdf/2990/299023509004.pdf>

Leys, C., Delacre, M., Mora, Y. L., Lakens, D., & Ley, C. (2019). How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration. *International Review of Social Psychology*, 32(1). <https://doi.org/10.5334/irsp.289>

Bitácora 3

Para facilitar la revisión de la bitácora 3, se adjunta los paquetes que se tienen que cargar, además de la carga del archivo

```

#install.packages("rcompanion")
#install.packages("ResourceSelection")
#install.packages("pROC")
#install.packages("car")

library(rcompanion)
library(ResourceSelection)
library(pROC)
library(car)
library(dplyr)
library(tidyr)
library(ggplot2)

datos <- read.csv2("data.csv", sep = ";", header = TRUE, stringsAsFactors = FALSE)
datos <- datos %>%
  select(-GDP) ### El uso de esta variable es irrelevante

```

1. Técnicas o Métodos Estadísticos Seleccionados

Regresión Logística Binaria: para modelar la probabilidad de deserción en función de factores individuales, sociales y académicos.

2. Marco Metodológico

Etapas preliminares del análisis

Estadística descriptiva: Nos ayuda a recolectar, organizar, analizar y presentar un conjunto de datos que nos ayuda a conocer sus principales características.

Análisis Exploratorio de Datos (EDA): Para comprender la distribución y relaciones básicas entre variables.

Método estadístico principal: Regresión Logística Binaria La regresión logística modela la probabilidad de que un evento ocurra (en este caso, deserción) mediante la función logística:

$P(Y=1/X)=1/(1+e^{-(\beta_0+\beta_1X_1+\dots+\beta_pX_p)})$ Donde:

Y es la variable dependiente binaria (Dropout vs Graduate/Enrolled) X_1, \dots, X_p son las variables predictoras β_0, \dots, β_p son los coeficientes a estimar La estimación se realiza mediante máxima verosimilitud, maximizando la función:

$$\ell(\beta) = \sum_{i=1}^n [y_i \log(p_i) + (1-y_i) \log(1-p_i)]$$

Fundamento estadístico: La regresión logística modela la relación entre un conjunto de variables predictoras y la probabilidad de ocurrencia del evento.

Justificación: En nuestro caso, la regresión logística binaria es apropiada ya que nuestro análisis está centrado en una variable binaria, que toma el valor 1 en caso de “Dropout” y 0 en caso contrario. El uso de esta técnica permite evaluar cómo se relacionan múltiples factores, que en nuestro caso pueden ser académicos y socioeconómicos. Nos facilita la interpretación mediante odds ratios, además de ser un método confiable y simple, ampliamente utilizado en investigación.

3- Aplique dichas técnicas o métodos. Presente los resultados, analícelos y responda su pregunta de investigación.

Estadística descriptiva Permite recolectar, organizar, analizar y presentar conjuntos de datos para identificar y comprender sus características principales, patrones y tendencias.

```
columnas_seleccionadas <- datos[, c(
  "Admission.grade",
  "Age.at.enrollment",
  "Previous.qualification..grade.",
  "Curricular.units.1st.sem..grade.",
  "Curricular.units.2nd.sem..grade."
)]

#pasa r chr a num y pasar int a num
columnas_seleccionadas <- as.data.frame(lapply(columnas_seleccionadas, function(x) {
  if(is.character(x)) {
    # Convertir a numérico (maneja decimales)
```

```

    return(as.numeric(as.character(x)))
  } else {
    return(x)
  }
}))
#str(columnas_seleccionadas)
#Se imprimió cuando se estaba analizando para tener mejor comprensión

```

```

calcular_moda <- function(x){
  x <- x[!is.na(x)]
  if(length(x) == 0) return(NA)

  tabla <- table(x)
  moda <- as.numeric((names(tabla)[tabla == max(tabla)]))
  return(moda[1])
}

```

```

estadisticas_descrip <- data.frame(
  Media = sapply(columnas_seleccionadas, mean, na.rm = TRUE),
  Mediana = sapply(columnas_seleccionadas, median, na.rm = TRUE),
  Moda = sapply(columnas_seleccionadas, calcular_moda),
  Desviacion_Estandar = sapply(columnas_seleccionadas, sd, na.rm = TRUE),
  Minimo = sapply(columnas_seleccionadas, min, na.rm = TRUE),
  Maximo = sapply(columnas_seleccionadas, max, na.rm = TRUE),
  Rango = sapply(columnas_seleccionadas, function(x) diff(range(x, na.rm = TRUE)))
)
print(estadisticas_descrip)

```

	Media	Mediana	Moda	Desviacion_Estandar
Admission.grade	126.97812	126.10000	130.0	14.482001

Age.at.enrollment	23.26514	20.00000	18.0	7.587816
Previous.qualification..grade.	132.61331	133.10000	133.1	13.188332
Curricular.units.1st.sem..grade.	10.64082	12.28571	0.0	4.843663
Curricular.units.2nd.sem..grade.	10.23021	12.20000	0.0	5.210808
	Minimo	Maximo	Rango	
Admission.grade	95	190.00000	95.00000	
Age.at.enrollment	17	70.00000	53.00000	
Previous.qualification..grade.	95	190.00000	95.00000	
Curricular.units.1st.sem..grade.	0	18.87500	18.87500	
Curricular.units.2nd.sem..grade.	0	18.57143	18.57143	

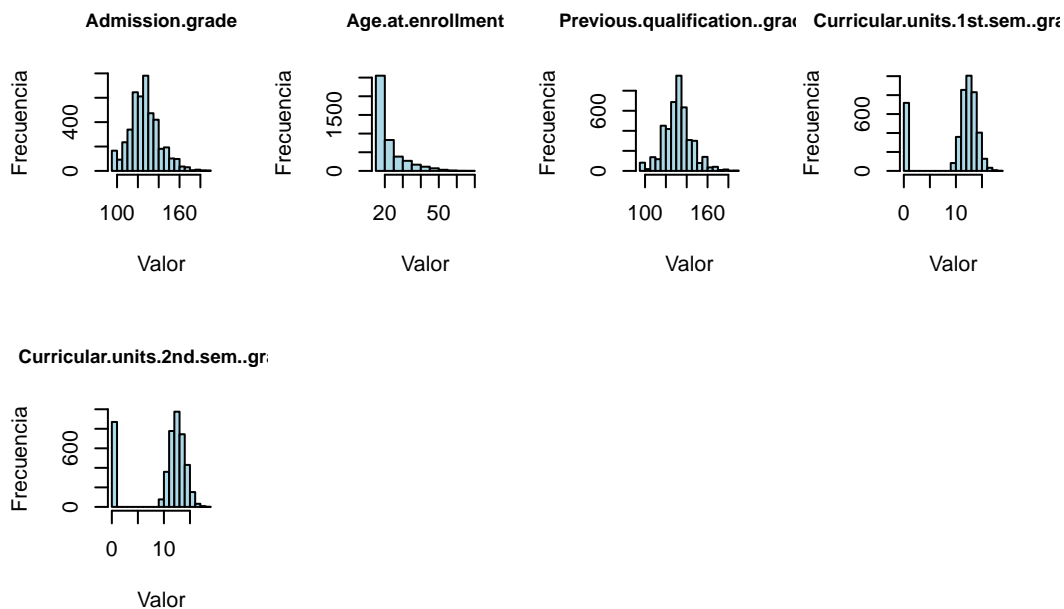
```

par(mfrow = c(2, 4)) # Parámetros gráficos

nombres_var <- names(columnas_seleccionadas) # Extrae los nombres y los asigna al vector

for(i in 1:ncol(columnas_seleccionadas)) { # Bucle para generar los histogramas, iterand
  hist(columnas_seleccionadas[[i]], #Estructura que va a tener el histograma
    main = nombres_var[i],
    cex.main = 0.9,
    xlab = "Valor",
    ylab = "Frecuencia",
    col = "lightblue",
    border = "black")
}

```



análisis En el análisis estadístico nos ayuda a poder ver las variables, pudiendo analizar y comparar, tipo se puede ver uno como la moda, lo cual sería los valores más frecuentes, el rango difiere entre valores máximos y mínimos. En las gráficas se puede ver los histogramas, en el cual se puede ver las variables que se seleccionaron y se le signaron en columnas_seleccionadas, se grafica esto, para tener una idea del comportamiento de estas variables.

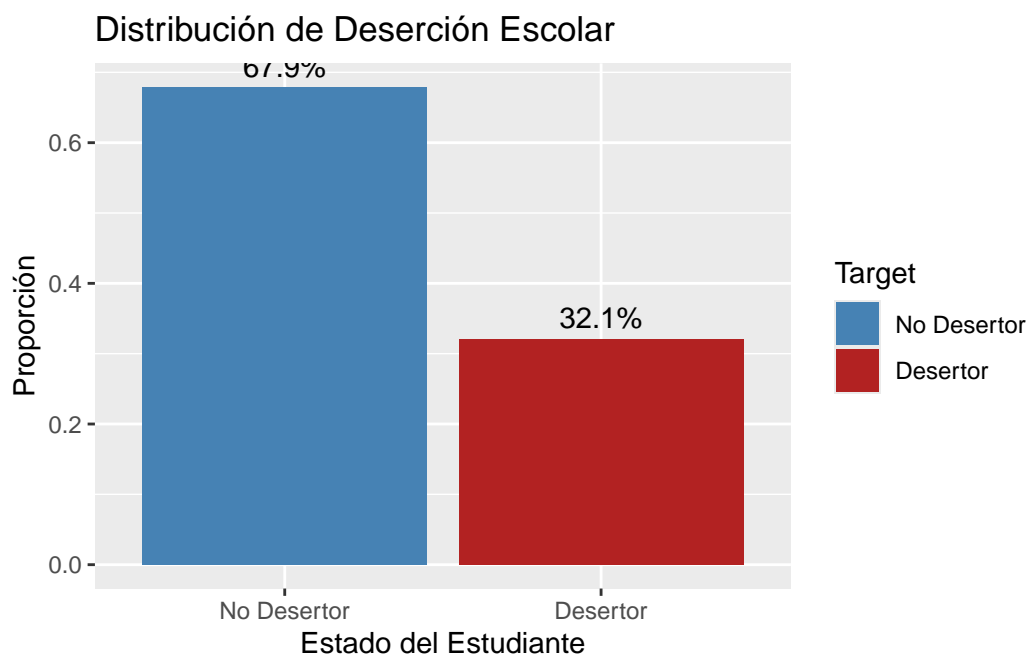
Gráficos de las variables que tienen relación directa con Target:

```
datos %>%
  mutate(
    Target = factor(
      Target,
      levels = c("Graduate", "Enrolled", "Dropout"),
      labels = c("No Desertor", "No Desertor", "Desertor")
    )
  ) %>%
  ggplot(aes(x = Target, fill = Target)) +
  geom_bar(aes(y = after_stat(count / sum(count)))) +
  geom_text(
```

```

aes(
  y = after_stat(count / sum(count)),
  label = paste0(round(after_stat(count / sum(count)) * 100, 1), "%"),
),
stat = "count",
vjust = -0.5,
size = 4
) +
scale_fill_manual(
  values = c("No Desertor" = "steelblue",
            "Desertor"      = "firebrick")
) +
labs(
  title = "Distribución de Deserción Escolar",
  x = "Estado del Estudiante",
  y = "Proporción"
)

```

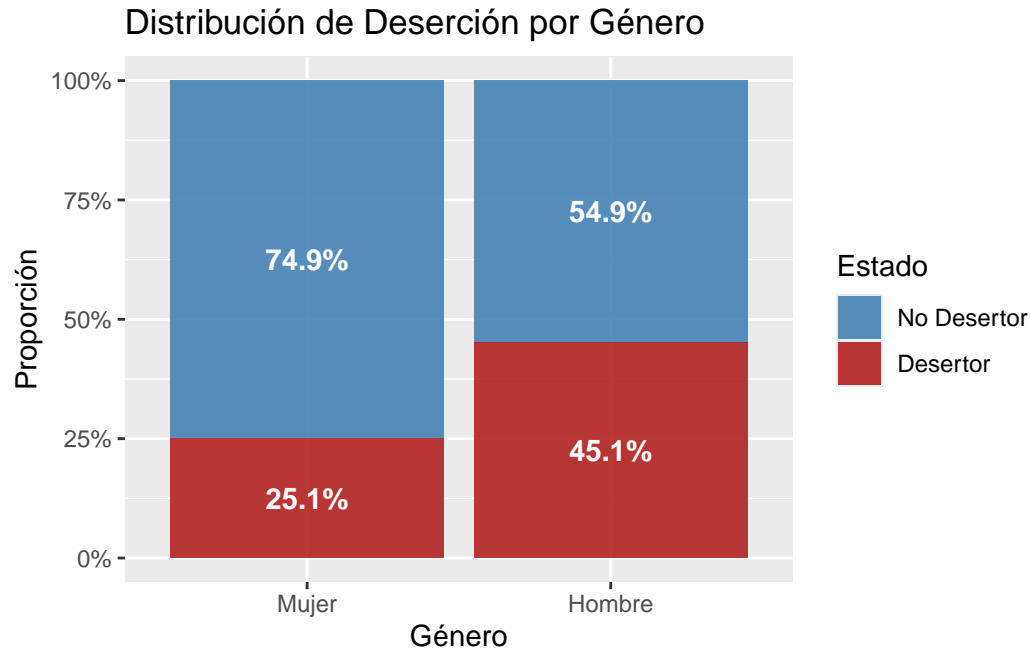


```

datos %>%
  mutate(
    Target = factor(Target,
                     levels = c("Graduate", "Enrolled", "Dropout"),
                     labels = c("No Desertor", "No Desertor", "Desertor")),
    Gender = factor(Gender, levels = c(0, 1), labels = c("Mujer", "Hombre"))
  ) %>%
  group_by(Gender, Target) %>%
  summarise(n = n(), .groups = "drop") %>%
  group_by(Gender) %>%
  mutate(prop = n / sum(n)) %>%

  ggplot(aes(x = Gender, y = prop, fill = Target)) +
  geom_bar(stat = "identity", position = "stack", alpha = 0.9) +
  geom_text(
    aes(label = paste0(round(prop * 100, 1), "%")),
    position = position_stack(vjust = 0.5),
    color = "white",
    fontface = "bold",
    size = 4
  ) +
  scale_fill_manual(values = c("steelblue", "firebrick")) +
  scale_y_continuous(labels = scales::percent) +
  labs(
    title = "Distribución de Deserción por Género",
    x = "Género",
    y = "Proporción",
    fill = "Estado"
  )

```

```

datos %>%
  mutate(
    Target = factor(
      Target,
      levels = c("Graduate", "Enrolled", "Dropout"),
      labels = c("No Desertor", "No Desertor", "Desertor")
    )
  ) %>%
  count(Age.at.enrollment, Target) %>%
  group_by(Age.at.enrollment) %>%
  mutate(prop = n / sum(n)) %>%

  ggplot(aes(x = Age.at.enrollment, y = prop, color = Target)) +
  geom_line(size = 1.3, alpha = 0.9) +
  geom_point(size = 3) +
  scale_color_manual(
    values = c("No Desertor" = "steelblue",

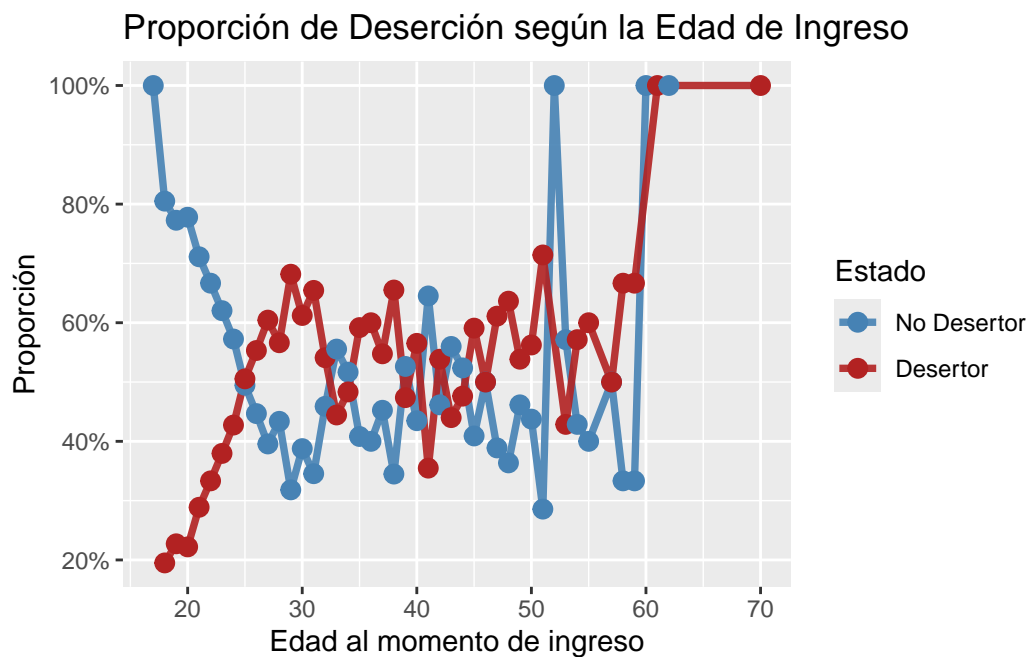
```

```

    "Desertor"      = "firebrick")

) +
scale_y_continuous(labels = scales::percent_format()) +
labs(
  title = "Proporción de Deserción según la Edad de Ingreso",
  x = "Edad al momento de ingreso",
  y = "Proporción",
  color = "Estado"
)

```



```

datos %>%
  mutate(
    Target = factor(
      Target,
      levels = c("Graduate", "Enrolled", "Dropout"),
      labels = c("No desertor", "No desertor", "Desertor")
    )
  ) %>%

```

```

select(
  Target,
  Matriculados = Curricular.units.1st.sem..enrolled.,
  Aprobados = Curricular.units.1st.sem..approved.
) %>%
pivot_longer(
  cols = c(Matriculados, Aprobados),
  names_to = "Variable",
  values_to = "Valor"
) %>%

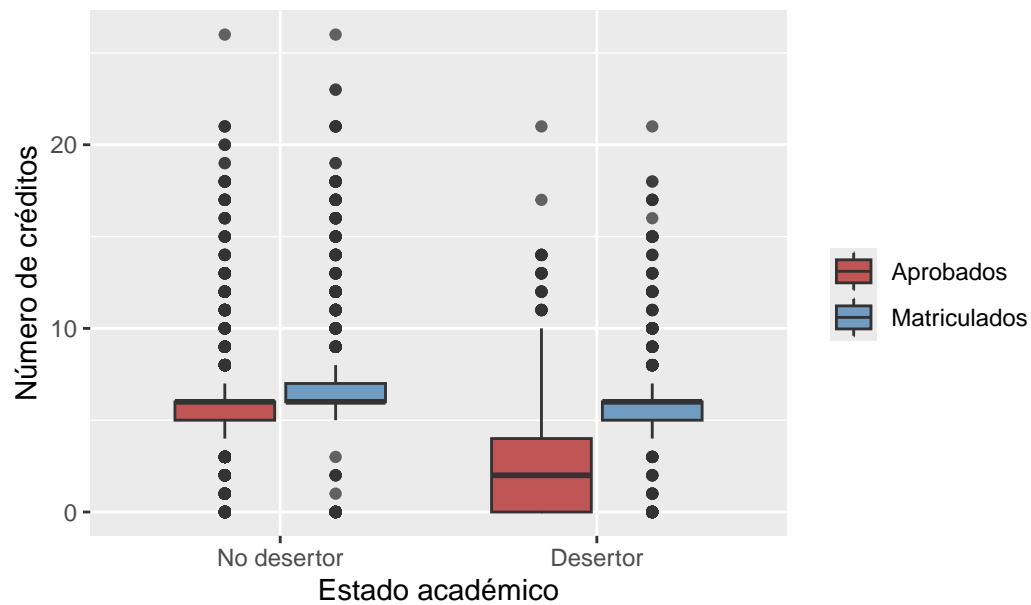
ggplot(aes(x = Target, y = Valor, fill = Variable)) +
geom_boxplot(alpha = 0.75, width = 0.7, outlier.size = 1.5) +

scale_fill_manual(
  values = c("Matriculados" = "steelblue",
             "Aprobados" = "firebrick")
) +

labs(
  title = "Créditos 1er Semestre: Matriculados y Aprobados según Estado Académico",
  x = "Estado académico",
  y = "Número de créditos",
  fill = ""
)

```

Créditos 1er Semestre: Matriculados y Aprobados según Estad



```
datos %>%
  mutate(
    Target = factor(
      Target,
      levels = c("Graduate", "Enrolled", "Dropout"),
      labels = c("No desertor", "No desertor", "Desertor")
    )
  ) %>%
  select(
    Target,
    Matriculados = Curricular.units.2nd.sem..enrolled.,
    Aprobados = Curricular.units.2nd.sem..approved.
  ) %>%
  pivot_longer(
    cols = c(Matriculados, Aprobados),
    names_to = "Variable",
    values_to = "Valor"
```

```

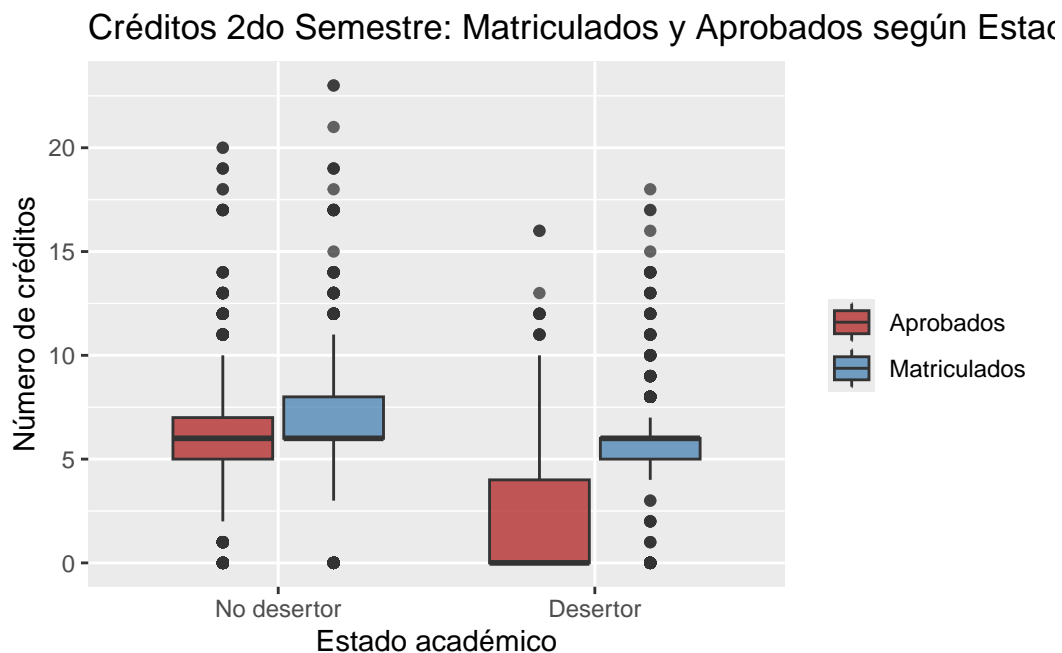
) %>%

ggplot(aes(x = Target, y = Valor, fill = Variable)) +
geom_boxplot(alpha = 0.75, width = 0.7, outlier.size = 1.5) +

scale_fill_manual(
  values = c("Matriculados" = "steelblue",
            "Aprobados" = "firebrick")
) +

labs(
  title = "Créditos 2do Semestre: Matriculados y Aprobados según Estado Académico",
  x = "Estado académico",
  y = "Número de créditos",
  fill = ""
)

```



```

datos %>%
  mutate(
    Target = factor(
      Target,
      levels = c("Graduate", "Enrolled", "Dropout"),
      labels = c("No Desertor", "No Desertor", "Desertor")
    ),
    Marital.status = factor(
      Marital.status,
      levels = c(1, 2, 3, 4, 5, 6),
      labels = c("Soltero", "Casado", "Viudo", "Divorciado", "Unión Lib", "Separado")
    )
  ) %>%
  group_by(Marital.status, Target) %>%
  summarise(n = n(), .groups = "drop") %>%
  group_by(Marital.status) %>%
  mutate(prop = n / sum(n)) %>%

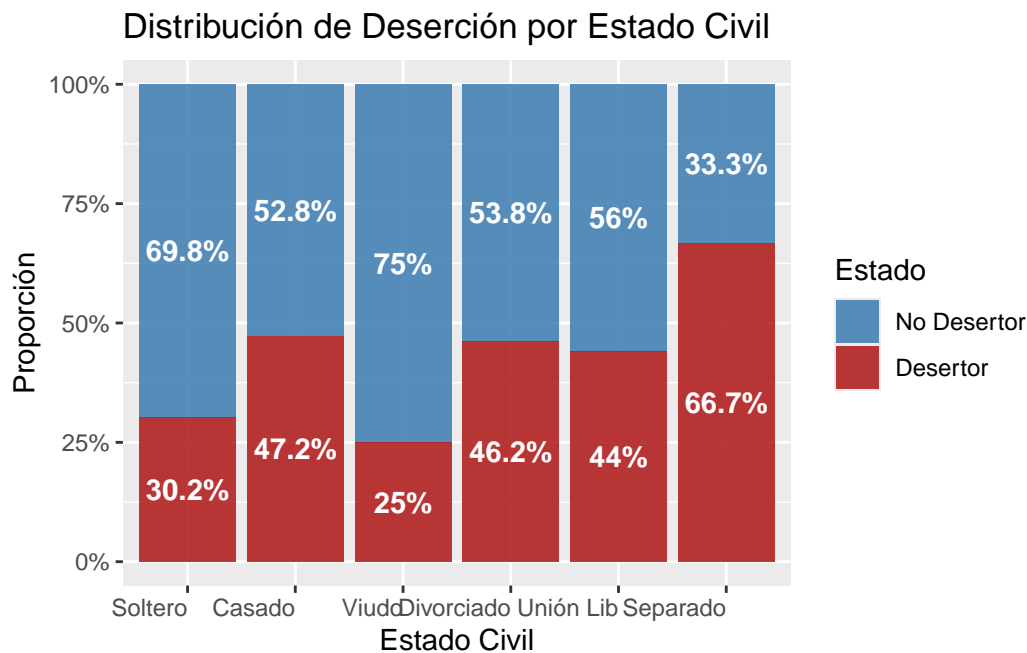
  ggplot(aes(x = Marital.status, y = prop, fill = Target)) +
  geom_bar(stat = "identity", position = "stack", alpha = 0.9) +
  geom_text(
    aes(label = paste0(round(prop * 100, 1), "%")),
    position = position_stack(vjust = 0.5),
    color = "white",
    fontface = "bold",
    size = 4
  ) +
  scale_fill_manual(values = c("steelblue", "firebrick")) +
  scale_y_continuous(labels = scales::percent) +
  labs(

```

```

title = "Distribución de Deserción por Estado Civil",
x = "Estado Civil",
y = "Proporción",
fill = "Estado"
) +
theme(axis.text.x = element_text(angle = 0, hjust = 1))

```



```

datos %>%
  mutate(
    Target = factor(Target,
                     levels = c("Graduate", "Enrolled", "Dropout"),
                     labels = c("No Desertor", "No Desertor", "Desertor")),
    Educational.special.needs = factor(Educational.special.needs, levels = c(0, 1),
                                       labels = c("No", "Sí"))
  ) %>%
  # Agrupar para proporciones
  group_by(Educational.special.needs, Target) %>%
  summarise(n = n(), .groups = "drop") %>%

```

```

group_by(Educational.special.needs) %>%
mutate(prop = n / sum(n)) %>%

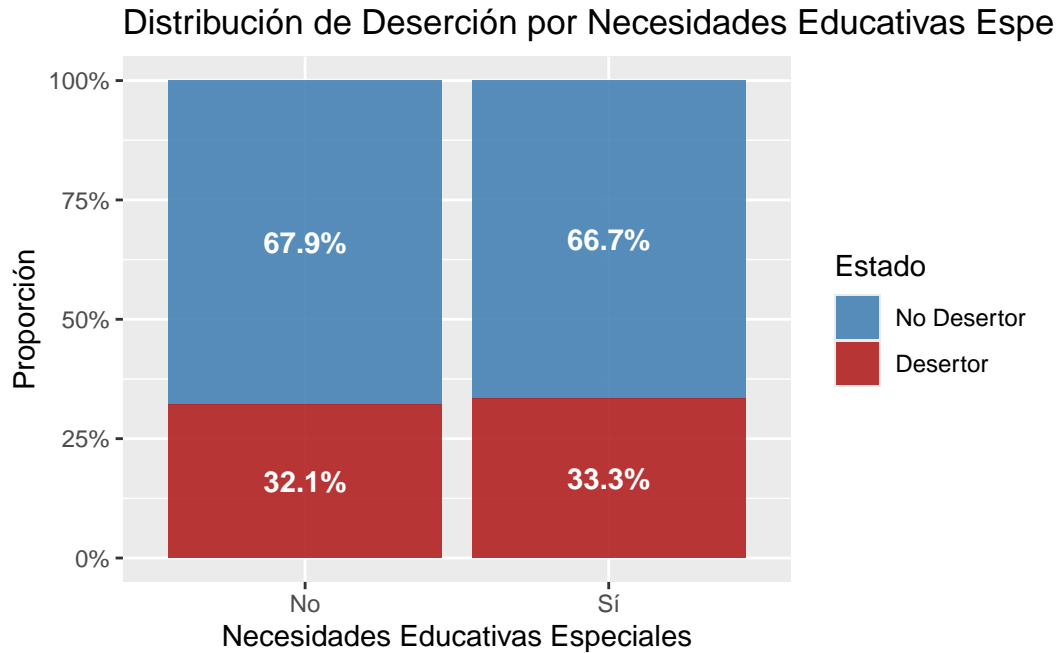
ggplot(aes(x = Educational.special.needs, y = prop, fill = Target)) +
geom_bar(stat = "identity", position = "stack", alpha = 0.9) +
geom_text(
  aes(label = paste0(round(prop * 100, 1), "%")),
  position = position_stack(vjust = 0.5),
  color = "white",
  fontface = "bold",
  size = 4
) +

scale_fill_manual(values = c("steelblue", "firebrick")) +

scale_y_continuous(labels = scales::percent) +

labs(
  title = "Distribución de Deserción por Necesidades Educativas Especiales",
  x = "Necesidades Educativas Especiales",
  y = "Proporción",
  fill = "Estado"
)

```

```

datos %>%
  mutate(
    Target = factor(Target,
                     levels = c("Graduate", "Enrolled", "Dropout"),
                     labels = c("No Desertor", "No Desertor", "Desertor")),
    Debtor = factor(Debtor, levels = c(0, 1), labels = c("No Deudor", "Deudor"))
  ) %>%
  # Agrupar para proporciones
  group_by(Debtor, Target) %>%
  summarise(n = n(), .groups = "drop") %>%
  group_by(Debtor) %>%
  mutate(prop = n / sum(n)) %>%

  ggplot(aes(x = Debtor, y = prop, fill = Target)) +
  geom_bar(stat = "identity", position = "stack", alpha = 0.9) +
  geom_text(
    aes(label = paste0(round(prop * 100, 1), "%")),

```

```

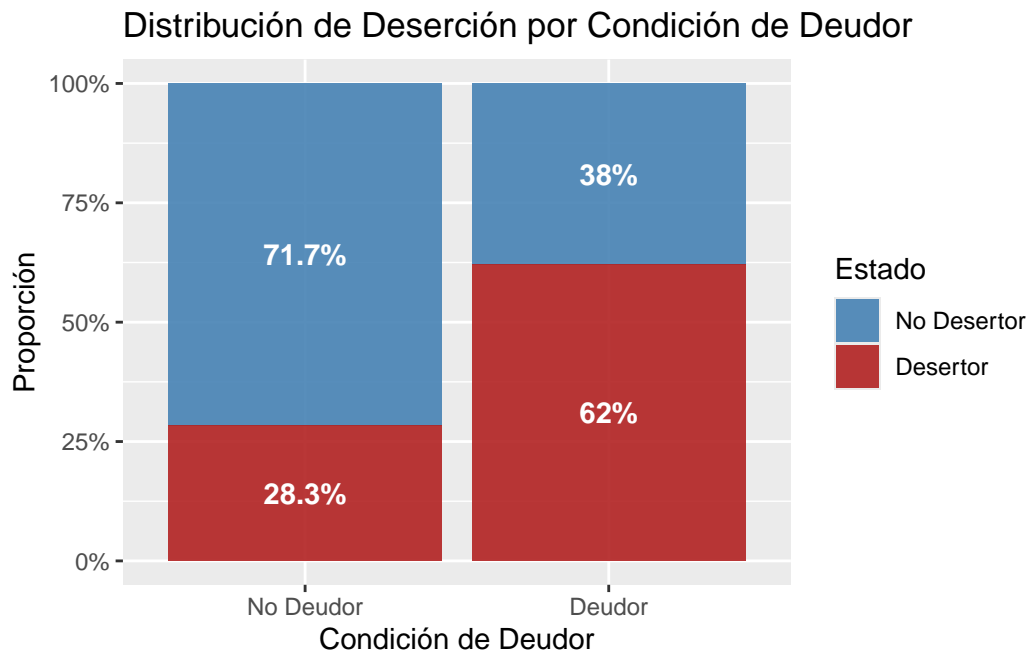
position = position_stack(vjust = 0.5),
color = "white",
fontface = "bold",
size = 4
) +

scale_fill_manual(values = c("steelblue", "firebrick")) +

scale_y_continuous(labels = scales::percent) +

labs(
  title = "Distribución de Deserción por Condición de Deudor",
  x = "Condición de Deudor",
  y = "Proporción",
  fill = "Estado"
)

```



```

datos %>%
  mutate(
    Target = factor(Target,
                     levels = c("Graduate", "Enrolled", "Dropout"),
                     labels = c("No Desertor", "No Desertor", "Desertor")),
    Scholarship.holder = factor(Scholarship.holder, levels = c(0, 1),
                                labels = c("No Becario", "Becario"))
  ) %>%
  # Agrupar para proporciones
  group_by(Scholarship.holder, Target) %>%
  summarise(n = n(), .groups = "drop") %>%
  group_by(Scholarship.holder) %>%
  mutate(prop = n / sum(n)) %>%

  ggplot(aes(x = Scholarship.holder, y = prop, fill = Target)) +
  geom_bar(stat = "identity", position = "stack", alpha = 0.9) +
  geom_text(
    aes(label = paste0(round(prop * 100, 1), "%")),
    position = position_stack(vjust = 0.5),
    color = "white",
    fontface = "bold",
    size = 4
  ) +

  scale_fill_manual(values = c("steelblue", "firebrick")) +

  scale_y_continuous(labels = scales::percent) +

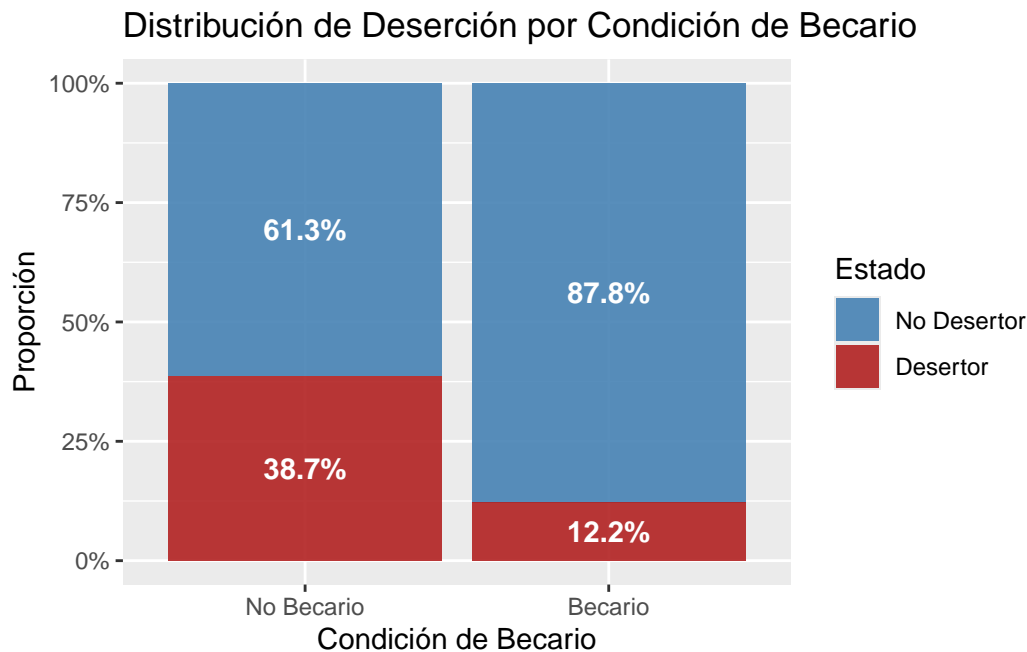
  labs(
    title = "Distribución de Deserción por Condición de Becario",

```

```

x = "Condición de Becario",
y = "Proporción",
fill = "Estado"
)

```



```

datos %>%
  mutate(
    Target = factor(Target,
                     levels = c("Graduate", "Enrolled", "Dropout"),
                     labels = c("No Desertor", "No Desertor", "Desertor")),
    International = factor(International, levels = c(0, 1),
                           labels = c("Nacional", "Internacional"))
  ) %>%
  # Agrupar para proporciones
  group_by(International, Target) %>%
  summarise(n = n(), .groups = "drop") %>%
  group_by(International) %>%
  mutate(prop = n / sum(n)) %>%

```

```

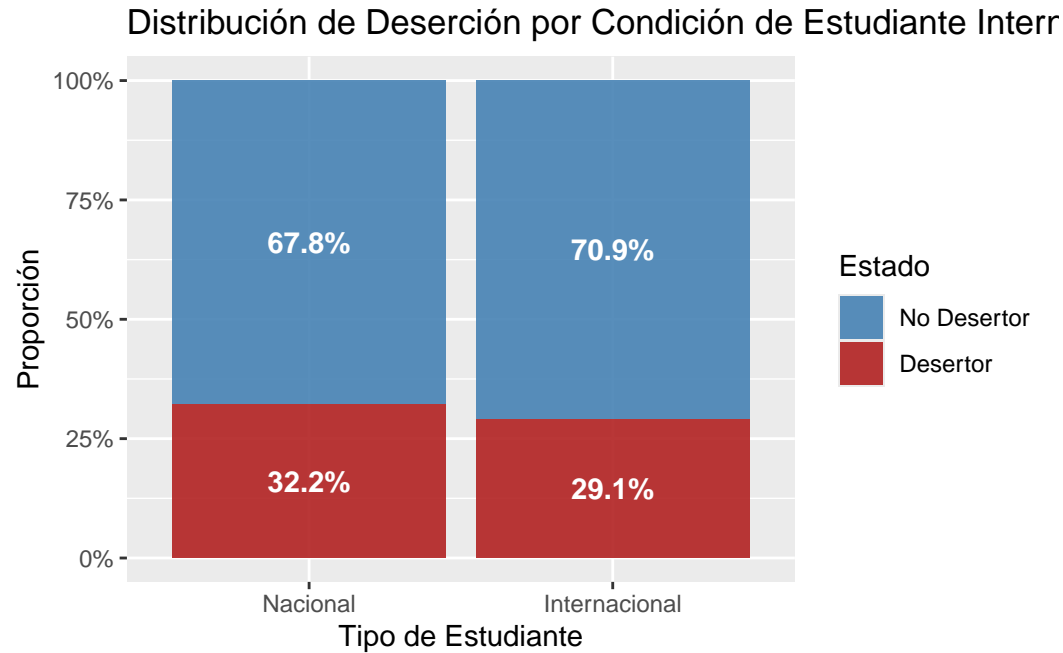
ggplot(aes(x = International, y = prop, fill = Target)) +
geom_bar(stat = "identity", position = "stack", alpha = 0.9) +
geom_text(
  aes(label = paste0(round(prop * 100, 1), "%")),
  position = position_stack(vjust = 0.5),
  color = "white",
  fontface = "bold",
  size = 4
) +

scale_fill_manual(values = c("steelblue", "firebrick")) +

scale_y_continuous(labels = scales::percent) +

labs(
  title = "Distribución de Deserción por Condición de Estudiante Internacional",
  x = "Tipo de Estudiante",
  y = "Proporción",
  fill = "Estado"
)

```



Este apartado es para los tipos de variables

```
datos_num <- datos %>%
  select(
    Previous.qualification..grade.,
    Application.order,
    Admission.grade,
    Age.at.enrollment,
    Curricular.units.1st.sem..credited.,
    Curricular.units.1st.sem..enrolled.,
    Curricular.units.1st.sem..approved.,
    Curricular.units.1st.sem..evaluations.,
    Curricular.units.1st.sem..grade.,
    Curricular.units.1st.sem..without.evaluations.,
    Curricular.units.2nd.sem..credited.,
    Curricular.units.2nd.sem..enrolled.,
    Curricular.units.2nd.sem..evaluations.,
    Curricular.units.2nd.sem..approved.,
```

```

Curricular.units.2nd.sem..grade.,
Curricular.units.2nd.sem..without.evaluations.,
)

tabla_num <- data.frame(
  Variables_Numericas = names(datos_num)
)

tabla_num

```

```

Variables_Numericas
1 Previous.qualification..grade.
2 Application.order
3 Admission.grade
4 Age.at.enrollment
5 Curricular.units.1st.sem..credited.
6 Curricular.units.1st.sem..enrolled.
7 Curricular.units.1st.sem..approved.
8 Curricular.units.1st.sem..evaluations.
9 Curricular.units.1st.sem..grade.
10 Curricular.units.1st.sem..without.evaluations.
11 Curricular.units.2nd.sem..credited.
12 Curricular.units.2nd.sem..enrolled.
13 Curricular.units.2nd.sem..evaluations.
14 Curricular.units.2nd.sem..approved.
15 Curricular.units.2nd.sem..grade.
16 Curricular.units.2nd.sem..without.evaluations.

```

```

datos_cat <- datos %>%
  select(
    Marital.status,
    Application.mode,
    Course,
    Daytime.evening.attendance.,
    Previous.qualification,
    Nacionality,
    Mother.s.qualification,
    Father.s.qualification,
    Mother.s.occupation,
    Father.s.occupation,
    Displaced,
    Educational.special.needs,
    Debtor,
    Tuition.fees.up.to.date,
    Gender,
    Scholarship.holder,
    International,
    Target
  )

tabla_cat <- data.frame(
  Variables_Categoricas = names(datos_cat)
)

tabla_cat

```

```

      Variables_Categoricas
1      Marital.status

```



```

2           Application.mode
3           Course
4 Daytime.evening.attendance.
5       Previous.qualification
6           Nacionality
7       Mother.s.qualification
8       Father.s.qualification
9       Mother.s.occupation
10      Father.s.occupation
11           Displaced
12 Educational.special.needs
13           Debtor
14      Tuition.fees.up.to.date
15           Gender
16      Scholarship.holder
17           International
18           Target

```

Aplicando Regresión Logística

Nota: Se utilizaron más técnicas para verificar la veracidad de la información obtenida, de esta manera, se presentan los resultados

```

### Pasamos los casos de Desertor y Matriculado/Graduado a 1,0 respectivamente
datos$Target_bin <- ifelse(datos$Target == "Dropout", 1, 0)
datos$Target_bin <- factor(datos$Target_bin, levels = c(0, 1))

modelo <- glm(Target_bin ~ Age.at.enrollment +
               Curricular.units.1st.sem..enrolled. +
               Curricular.units.1st.sem..approved. +

```

```

Curricular.units.2nd.sem..enrolled. +
Curricular.units.2nd.sem..approved. +
Marital.status +
Course +
Displaced +
Educational.special.needs +
Debtor +
Tuition.fees.up.to.date +
Gender +
Scholarship.holder +
International +
Mother.s.qualification +
Father.s.qualification +
Mother.s.occupation +
Father.s.occupation,
data = datos,
family = binomial)

```

```
summary(modelo)
```

Call:

```

glm(formula = Target_bin ~ Age.at.enrollment + Curricular.units.1st.sem..enrolled. +
  Curricular.units.1st.sem..approved. + Curricular.units.2nd.sem..enrolled. +
  Curricular.units.2nd.sem..approved. + Marital.status + Course +
  Displaced + Educational.special.needs + Debtor + Tuition.fees.up.to.date +
  Gender + Scholarship.holder + International + Mother.s.qualification +
  Father.s.qualification + Mother.s.occupation + Father.s.occupation,
  family = binomial, data = datos)

```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	3.281e-01	3.061e-01	1.072	0.28376	
Age.at.enrollment	5.174e-02	8.160e-03	6.340	2.30e-10	***
Curricular.units.1st.sem..enrolled.	1.938e-01	7.154e-02	2.709	0.00675	**
Curricular.units.1st.sem..approved.	-2.236e-01	3.999e-02	-5.591	2.26e-08	***
Curricular.units.2nd.sem..enrolled.	4.547e-01	7.994e-02	5.688	1.28e-08	***
Curricular.units.2nd.sem..approved.	-6.501e-01	3.752e-02	-17.325	< 2e-16	***
Marital.status	-1.547e-01	9.662e-02	-1.601	0.10930	
Course	-7.115e-05	2.369e-05	-3.004	0.00267	**
Displaced	2.428e-01	1.066e-01	2.276	0.02283	*
Educational.special.needs	3.750e-01	4.074e-01	0.920	0.35738	
Debtor	4.756e-01	1.602e-01	2.969	0.00299	**
Tuition.fees.up.to.date	-2.300e+00	1.744e-01	-13.189	< 2e-16	***
Gender	3.119e-01	1.017e-01	3.068	0.00215	**
Scholarship.holder	-5.670e-01	1.351e-01	-4.197	2.70e-05	***
International	-6.478e-01	3.343e-01	-1.938	0.05264	.
Mother.s.qualification	8.506e-03	3.856e-03	2.206	0.02739	*
Father.s.qualification	-3.450e-03	3.808e-03	-0.906	0.36498	
Mother.s.occupation	-1.205e-02	4.667e-03	-2.582	0.00981	**
Father.s.occupation	2.169e-03	4.781e-03	0.454	0.65006	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 5554.5 on 4423 degrees of freedom
 Residual deviance: 2857.5 on 4405 degrees of freedom
 AIC: 2895.5

Number of Fisher Scoring iterations: 5

```
### Convertir coeficientes a Odds Ratios
```

```
exp(coef(modelo))
```

(Intercept)	Age.at.enrollment
1.3883642	1.0530970
Curricular.units.1st.sem..enrolled.	Curricular.units.1st.sem..approved.
1.2138236	0.7996545
Curricular.units.2nd.sem..enrolled.	Curricular.units.2nd.sem..approved.
1.5756897	0.5220126
Marital.status	Course
0.8566540	0.9999289
Displaced	Educational.special.needs
1.2747588	1.4549726
Debtor	Tuition.fees.up.to.date
1.6089201	0.1002784
Gender	Scholarship.holder
1.3660759	0.5672287
International	Mother.s.qualification
0.5231712	1.0085426
Father.s.qualification	Mother.s.occupation
0.9965560	0.9880200
Father.s.occupation	
1.0021713	

```
## intervalos de confianza
```

```
exp(confint(modelo))
```

2.5 % 97.5 %

(Intercept)	0.76298132	2.5346840
Age.at.enrollment	1.03637701	1.0700896
Curricular.units.1st.sem..enrolled.	1.05323464	1.3945712
Curricular.units.1st.sem..approved.	0.73939508	0.8649507
Curricular.units.2nd.sem..enrolled.	1.34938965	1.8466195
Curricular.units.2nd.sem..approved.	0.48453253	0.5613541
Marital.status	0.70771288	1.0335316
Course	0.99988262	0.9999756
Displaced	1.03520944	1.5727865
Educational.special.needs	0.63708361	3.1658270
Debtor	1.17358287	2.1998528
Tuition.fees.up.to.date	0.07083641	0.1404101
Gender	1.11880938	1.6668522
Scholarship.holder	0.43381057	0.7369488
International	0.26680737	0.9898260
Mother.s.qualification	1.00095438	1.0162059
Father.s.qualification	0.98914065	1.0040234
Mother.s.occupation	0.97888650	0.9970416
Father.s.occupation	0.99279309	1.0116109

```
### Evaluación del modelo
```

```
# Nagelkerke
```

```
nagelkerke(modelo)
```

```
$Models
```

```
Model: "glm, Target_bin ~ Age.at.enrollment + Curricular.units.1st.sem..enrolled. + Curr
```

```
Null: "glm, Target_bin ~ 1, binomial, datos"
```

```
$Pseudo.R.squared.for.model.vs.null
```

	Pseudo.R.squared
McFadden	0.485563
Cox and Snell (ML)	0.456457
Nagelkerke (Cragg and Uhler)	0.638330

```
$Likelihood.ratio.test
```

Df.diff	LogLik.diff	Chisq	p.value
-18	-1348.5	2697.1	0

```
$Number.of.observations
```

```
Model: 4424
```

```
Null: 4424
```

```
$Messages
```

```
[1] "Note: For models fit with REML, these statistics are based on refitting with ML"
```

```
$Warnings
```

```
[1] "None"
```

```
# Matriz de confusión
```

```
prob <- predict(modelo, type = "response")
```

```
pred <- ifelse(prob > 0.5, 1, 0)
```

```
pred <- factor(pred, levels = c(0, 1))
```

```
table(Predicción = pred, Real = datos$Target_bin)
```

	Real	
Predicción	0	1

0	2836	379
1	167	1042

```
# Curva ROC y AUC

roc_curve <- roc(datos$Target_bin, prob)
auc(roc_curve)
```

Area under the curve: 0.9154

```
### Revisar multicolinealidad (VIF)
vif(modelo)
```

Age.at.enrollment	Curricular.units.1st.sem..enrolled.
1.753128	14.649118
Curricular.units.1st.sem..approved.	Curricular.units.2nd.sem..enrolled.
5.268771	14.998800
Curricular.units.2nd.sem..approved.	Marital.status
4.075880	1.456451
Course	Displaced
1.727072	1.205242
Educational.special.needs	Debtor
1.006361	1.142268
Tuition.fees.up.to.date	Gender
1.151121	1.078595
Scholarship.holder	International
1.061526	1.034956
Mother.s.qualification	Father.s.qualification
1.567911	1.487045
Mother.s.occupation	Father.s.occupation
5.002125	4.939781

Calidad global del modelo

Nagelkerke = 0.638

Un pseudo- R^2 de 0.63 es excelente para un modelo de regresión logística, ya que logra explicar un 63% de la probabilidad de deserción.

AUC = 0.915

Esto quiere decir que es un modelo altamente predictivo, ya que según su escala, por encima del 0.90 es un modelo sobresaliente

Interpretación de coeficientes importantes (Odds Ratios)

Variables académicas

Son las más relevantes. Dominan completamente el modelo.

Variable	OR	Interpretación
1st.sem.enrolled	1.21	Cada curso <i>inscrito</i> extra implica que aumente la probabilidad de deserción 21%
1st.sem.approved	0.80	Cada curso aprobado extra implica que reduce un 20% la probabilidad
2nd.sem.enrolled	1.57	Los estudiantes “sobrecargados” tienden a salirse
2nd.sem.approved	0.52	Cada aprobación implica que reduzca la probabilidad en casi 50%

Conclusión: La carga académica y el rendimiento explican casi toda la deserción.

Variables financieras

Variable	OR	Interpretación
Tuition.fees.up.to.date	0.10	Estar al día reduce la deserción un 90%
Debtor	1.61	Los deudores hacen que aumente un 61% la probabilidad

Conclusión: Las finanzas son otro bloque crítico.

Variables personales

Variable	OR	Interpretación
Age.at.enrollment	1.05	Más edad → ↑ riesgo pequeño pero significativo
Gender (Hombres)	1.37	Mayor riesgo que las mujeres
Scholarship.holder	0.56	Becarios desertan menos

Variables familiares

Estás variables aportan, pero marginalmente.

Variable	p-valor	OR	Conclusión
Mother.s.qualification	0.027	1.008	Efecto pequeño, pero significativo
Mother.s.occupation	0.009	0.988	Ligera reducción del riesgo

Variables que no aportan

- Toda variable relacionada con el padre
- La nacionalidad

Matriz de confusión

El código presenta lo siguiente:

- Verdaderos negativos: 2836
- Falsos negativos: 379
- Verdaderos positivos: 1042
- Falsos positivos: 167

Métricas aproximadas:

- Efectividad $\approx 88.0\%$
- Sensibilidad $\approx 73\%$
- Especificidad $\approx 94\%$

Así, el modelo detecta muy bien a los que NO desertan; sin embargo, se pierde un porcentaje considerable de desertores (379 casos).

Conclusiones claras

Así entonces, podemos concluir, gracias a la técnica de Regresión Logística, que los factores que afectan más a las deserciones en educación superior son los factores financieros y académicos. Además, se presentan los factores personales y familiares, los cuales generan un aporte en la deserción, pero no de manera tan significativa. De igual manera, existen factores geográficos y personales que no afectan en lo absoluto en la deserción.

Bibliografía: