Introducción

En el ámbito educativo, comprender los factores que contribuyen al éxito o fracaso estudiantil es esencial para mejorar los sistemas educativos y garantizar que cada estudiante tenga la oportunidad de alcanzar su máximo potencial. En este contexto, este estudio se enfoca en investigar y analizar el rendimiento estudiantil, con el objetivo de identificar los factores que influyen en el éxito y el fracaso académico.

Para ello, llevaremos a cabo un análisis exploratorio de los datos (EDA). Luego, limpiaremos esos datos y los prepararemos para la fase de regresión. Por último, crearemos un modelo de regresión y comprobaremos su eficacia.

Los objetivos finales son:

- Conocer las principales variables que influyen en fracaso o éxito estudiantil.
- Desarrollar un modelo de regresión que nos permita predecir si un estudiante se graduará o no con, al menos, un 85% de precisión.

Descripción de los datos

El conjunto de datos que vamos a usar ofrece información sobre estudiantes inscritos en diversos programas de pregrado en una institución educativa. Incluye datos demográficos, socioeconómicos y académicos que nos ayudarán a analizar las causas del abandono y el éxito estudiantil.

El dataset se puede encontrar en: https://github.com/emibarrod/estadistica-viu

```
# Opción para poder visualizar todas las columnas
options(repr.matrix.max.cols=50, repr.matrix.max.rows=100)
# Carga de los datos. Nos aseguramos de que las
# columnas string se carguen como factores
dataset <- read.csv("dataset.csv", stringsAsFactors=TRUE)
```

Vemos que todas las columnas tienen valores numéricos, excepto la columna objetivo:

```
head(dataset, 5)
```

	Marital.status	Application.mode	Application.order	Course	Daytime.evenir
	<int>	<int>	<int>	<int>	
1	1	8	5	2	
2	1	6	1	11	
3	1	1	5	5	
4	1	8	2	15	
5	2	12	1	3	

Aún así, que sean valores numéricos, no significa que sean columnas numéricas. La mayoría, como veremos en un apartado posterior, son columnas categóricas.

El dataset tiene 4424 filas y 35 columnas.

```
dim(dataset)
```

4424 · 35

Un resumen de las características de todas las columnas:

```
summary(dataset)
```

```
 Marital.status  Application.mode Application.order     Course
 Min.   :1.000   Min.   : 1.000   Min.   :0.000     Min.   : 1.000
 1st Qu.:1.000   1st Qu.: 1.000   1st Qu.:1.000     1st Qu.: 6.000
 Median :1.000   Median : 8.000   Median :1.000     Median :10.000
 Mean   :1.179   Mean   : 6.887   Mean   :1.728     Mean   : 9.899
 3rd Qu.:1.000   3rd Qu.:12.000   3rd Qu.:2.000     3rd Qu.:13.000
 Max.   :6.000   Max.   :18.000   Max.   :9.000     Max.   :17.000
 Daytime.evening.attendance Previous.qualification  Nacionality
 Min.   :0.0000             Min.   : 1.000         Min.   : 1.000
 1st Qu.:1.0000             1st Qu.: 1.000         1st Qu.: 1.000
 Median :1.0000             Median : 1.000         Median : 1.000
 Mean   :0.8908             Mean   : 2.531         Mean   : 1.255
 3rd Qu.:1.0000             3rd Qu.: 1.000         3rd Qu.: 1.000
 Max.   :1.0000             Max.   :17.000         Max.   :21.000
 Mother.s.qualification Father.s.qualification Mother.s.occupation
 Min.   : 1.00          Min.   : 1.00          Min.   : 1.000
 1st Qu.: 2.00          1st Qu.: 3.00          1st Qu.: 5.000
 Median :13.00          Median :14.00          Median : 6.000
 Mean   :12.32          Mean   :16.46          Mean   : 7.318
 3rd Qu.:22.00          3rd Qu.:27.00          3rd Qu.:10.000
 Max.   :29.00          Max.   :34.00          Max.   :32.000
 Father.s.occupation   Displaced      Educational.special.needs
 Min.   : 1.000      Min.   :0.0000   Min.   :0.00000
 1st Qu.: 5.000      1st Qu.:0.0000   1st Qu.:0.00000
 Median : 8.000      Median :1.0000   Median :0.00000
 Mean   : 7.819      Mean   :0.5484   Mean   :0.01153
 3rd Qu.:10.000      3rd Qu.:1.0000   3rd Qu.:0.00000
 Max.   :46.000      Max.   :1.0000   Max.   :1.00000
     Debtor       Tuition.fees.up.to.date     Gender
Scholarship.holder
 Min.   :0.0000   Min.   :0.0000          Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:1.0000          1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.0000   Median :1.0000          Median :0.0000   Median :0.0000
 Mean   :0.1137   Mean   :0.8807          Mean   :0.3517   Mean   :0.2484
 3rd Qu.:0.0000   3rd Qu.:1.0000          3rd Qu.:1.0000   3rd Qu.:0.0000
 Max.   :1.0000   Max.   :1.0000          Max.   :1.0000   Max.   :1.0000
 Age.at.enrollment International     Curricular.units.1st.sem..credited.
 Min.   :17.00     Min.   :0.00000   Min.   : 0.00
 1st Qu.:19.00     1st Qu.:0.00000   1st Qu.: 0.00
 Median :20.00     Median :0.00000   Median : 0.00
 Mean   :23.27     Mean   :0.02486   Mean   : 0.71
 3rd Qu.:25.00     3rd Qu.:0.00000   3rd Qu.: 0.00
 Max.   :70.00     Max.   :1.00000   Max.   :20.00
 Curricular.units.1st.sem..enrolled.
Curricular.units.1st.sem..evaluations.
 Min.   : 0.000                     Min.   : 0.000
 1st Qu.: 5.000                     1st Qu.: 6.000
 Median : 6.000                     Median : 8.000
 Mean   : 6.271                     Mean   : 8.299
 3rd Qu.: 7.000                     3rd Qu.:10.000
 Max.   :26.000                     Max.   :45.000
 Curricular.units.1st.sem..approved. Curricular.units.1st.sem..grade.
 Min.   : 0.000                     Min.   : 0.00
 1st Qu.: 3.000                     1st Qu.:11.00
 Median : 5.000                     Median :12.29
 Mean   : 4.707                     Mean   :10.64
 3rd Qu.: 6.000                     3rd Qu.:13.40
 Max.   :26.000                     Max.   :18.88
 Curricular.units.1st.sem..without.evaluations.
 Min.   : 0.0000
 1st Qu.: 0.0000
 Median : 0.0000
 Mean   : 0.1377
 3rd Qu.: 0.0000
 Max.   :12.0000
 Curricular.units.2nd.sem..credited. Curricular.units.2nd.sem..enrolled.
 Min.   : 0.0000                     Min.   : 0.000
 1st Qu.: 0.0000                     1st Qu.: 5.000
 Median : 0.0000                     Median : 6.000
 Mean   : 0.5418                     Mean   : 6.232
 3rd Qu.: 0.0000                     3rd Qu.: 7.000
 Max.   :19.0000                     Max.   :23.000
 Curricular.units.2nd.sem..evaluations.
Curricular.units.2nd.sem..approved.
 Min.   : 0.000                        Min.   : 0.000
 1st Qu.: 6.000                        1st Qu.: 2.000
 Median : 8.000                        Median : 5.000
 Mean   : 8.063                        Mean   : 4.436
 3rd Qu.:10.000                        3rd Qu.: 6.000
 Max.   :33.000                        Max.   :20.000
 Curricular.units.2nd.sem..grade.
 Min.   : 0.00
 1st Qu.:10.75
 Median :12.20
 Mean   :10.23
 3rd Qu.:13.33
 Max.   :18.57
 Curricular.units.2nd.sem..without.evaluations. Unemployment.rate
 Min.   : 0.0000                                Min.   : 7.60
 1st Qu.: 0.0000                                1st Qu.: 9.40
```

```
Median : 0.0000                    Median :11.10
Mean   : 0.1503                    Mean   :11.57
```

Descripción de las columnas

```
Min.   :-0.800   Min.   :-4.060000   Dropout :1421
```

Nombre de la columna	Descripción
Marital status	El estado civil del estudiante. (Categórico)
Application mode	El método de aplicación utilizado por el estudiante. (Categórico)
Application order	El orden en el que el estudiante presentó la solicitud. (Numérico)
Course	El curso realizado por el estudiante. (Categórico)
Daytime/evening attendance	Si el estudiante asiste a clases durante el día o por la noche. (Categórico)
Previous qualification	La calificación obtenida por el estudiante antes de inscribirse en educación superior. (Categórico)
Nacionality	La nacionalidad del estudiante. (Categórico)
Mother's qualification	La calificación de la madre del estudiante. (Categórico)
Father's qualification	La calificación del padre del estudiante. (Categórico)
Mother's occupation	La ocupación de la madre del estudiante. (Categórico)
Father's occupation	La ocupación del padre del estudiante. (Categórico)
Displaced	Si el estudiante es una persona desplazada. (Categórico)
Educational special needs	Si el estudiante tiene alguna necesidad educativa especial. (Categórico)
Debtor	Si el estudiante es deudor. (Categórico)
Tuition fees up to date	Si las cuotas de matrícula del estudiante están al día. (Categórico)
Gender	El género del estudiante. (Categórico)
Scholarship holder	Si el estudiante es beneficiario de una beca. (Categórico)
Age at enrollment	La edad del estudiante al momento de la inscripción. (Numérico)
International	Si el estudiante es un estudiante internacional. (Categórico)
Curricular units 1st sem (credited)	El número de unidades curriculares acreditadas por el estudiante en el primer semestre. (Numérico)
Curricular units 1st sem (enrolled)	El número de unidades curriculares inscritas por el estudiante en el primer semestre. (Numérico)
Curricular units 1st sem (evaluations)	El número de unidades curriculares evaluadas por el estudiante en el primer semestre. (Numérico)
Curricular units 1st sem (approved)	El número de unidades curriculares aprobadas por el estudiante en el primer semestre. (Numérico)

Nuestro objetivo es la columna Target, la cual tiene 3 valores:

```
1 unique(as.vector.factor(dataset$Target))
```

'Dropout' · 'Graduate' · 'Enrolled'

Cuando preparemos los datos, nos aseguraremos de quedarnos solo con las columnas Dropout y Graduate, ya que los alumnos que están aún cursando sus estudios no nos aportan información.

Preparación del dataset

Como hemos comentado anteriormente, no tendremos en cuenta los alumnos que aún están cursando sus estudios.

```
1 dataset_limpio <- dataset[!(dataset$Target=="Enrolled"),]
```

Tras esto el dataset pasa de tener 4424 columnas a 3630 columnas.

```
1 dim(dataset_limpio)
```

3630 · 35

La columna Target es categórica y debemos transformarla a numérica. En este caso será una variable binaria:

```
1 # Label encoding
2 dataset_limpio$Target <- unclass(dataset_limpio$Target)
3 # Transformación a valor numérico
4 dataset_limpio$Target <- as.numeric(dataset_limpio$Target)
5 # Binarizamos al variable
6 dataset_limpio$Target[dataset_limpio$Target == 1] <- 0
7 dataset_limpio$Target[dataset_limpio$Target == 3] <- 1
```

Vemos que ahora la columna Target es binaria:

```
1 head(dataset_limpio["Target"], 3)
```

A data.frame: 3 × 1

	Target
	<dbl>
1	0
2	1
3	0

Calculamos las correlaciones de las distintas columnas con respecto a nuestra columna objetivo:

```
columnas <- c(colnames(dataset)!="Target")
correlaciones <- round(cor(dataset_limpio[,columnas], dataset_limpio$Target), 2)
```

Obtenemos las 10 columnas que más correlación tienen con nuestra columna objetivo para usarlas en el modelo de regresión:

```
head(correlaciones[order(correlaciones,decreasing=TRUE),], 10)
```

```
Curricular.units.2nd.sem..approved.:  0.65 Curricular.units.2nd.sem..grade.:  0.61
Curricular.units.1st.sem..approved.:  0.55 Curricular.units.1st.sem..grade.:  0.52
Tuition.fees.up.to.date:              0.44 Scholarship.holder:                0.31
```

Colinearidad

```
columnas <- c("Curricular.units.2nd.sem..approved.", "Curricular.units.2nd.sem..grade.", "Curricular.units.1st.sem..appro
"Curricular.units.1st.sem..grade.", "Tuition.fees.up.to.date", "Scholarship.holder", "Curricular.units.2nd.sem..enrolled.
"Curricular.units.1st.sem..enrolled.",  "Displaced")
correlaciones <- round(cor(dataset_limpio[,columnas], dataset_limpio[,columnas]), 2)
correlaciones
```

	Curricular.units.2nd.sem..approved.	Curricu
Curricular.units.2nd.sem..approved.	1.00	
Curricular.units.2nd.sem..grade.	0.79	
Curricular.units.1st.sem..approved.	0.92	
Curricular.units.1st.sem..grade.	0.69	
Tuition.fees.up.to.date	0.33	
Scholarship.holder	0.21	
Curricular.units.2nd.sem..enrolled.	0.70	
Curricular.units.1st.sem..enrolled.	0.67	
Displaced	0.08	

Regresión logística

Función de precisión

La función de precisión será simple: el número de predicciones correctas entre el total de predicciones realizadas.

```
check_accuracy <- function(predictions, true_values) {
  count <- 0
  total <- length(true_values)
  for (x in 1:total) {
    if (predictions[x]==true_values[x]) {
      count <- count +1
    }
  }
  round(count/total, 3)
}
```

Creación del modelo

Usaremos las filas desde la 1 hasta la 3000 para entrenar el modelo y de la 3001 a la 3630 para test.

```
modelo <- glm(
  Target ~ Curricular.units.2nd.sem..approved. + Curricular.units.2nd.sem..grade. + Curricular.units.1st.sem..approved. +
Curricular.units.1st.sem..grade. + Tuition.fees.up.to.date + Scholarship.holder + Curricular.units.2nd.sem..enrolled.   +
Curricular.units.1st.sem..enrolled. + Displaced, data = dataset_limpio[1:3000,], family = "binomial")
```

```
summary(modelo)
```

```
Call:
glm(formula = Target ~ Curricular.units.2nd.sem..approved. +
    Curricular.units.2nd.sem..grade. + Curricular.units.1st.sem..approved.
+
    Curricular.units.1st.sem..grade. + Tuition.fees.up.to.date +
    Scholarship.holder + Curricular.units.2nd.sem..enrolled. +
    Curricular.units.1st.sem..enrolled. + Displaced, family = "binomial",
    data = dataset_limpio[1:3000, ])

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -2.88281    0.34976  -8.242  < 2e-16
***
Curricular.units.2nd.sem..approved.  1.03591    0.07010  14.778  < 2e-16
***
Curricular.units.2nd.sem..grade.     0.13751    0.05309   2.590  0.00959
**
Curricular.units.1st.sem..approved.  0.52175    0.07545   6.916 4.66e-12
***
Curricular.units.1st.sem..grade.    -0.05850    0.05209  -1.123  0.26145
Tuition.fees.up.to.date              2.97801    0.28631  10.401  < 2e-16
***
Scholarship.holder                   0.87887    0.16999   5.170 2.34e-07
***
Curricular.units.2nd.sem..enrolled. -0.79620    0.11826  -6.733 1.66e-11
***
Curricular.units.1st.sem..enrolled. -0.47606    0.10831  -4.395 1.11e-05
***
Displaced                           -0.03432    0.14159  -0.242  0.80849
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Predicción

```
test = dataset_limpio[3001:3630, c("Curricular.units.2nd.sem..approved.", "Curricular.units.2nd.sem..grade.", "Curricular
"Curricular.units.1st.sem..grade.", "Tuition.fees.up.to.date", "Scholarship.holder", "Curricular.units.2nd.sem..enrolled.
"Curricular.units.1st.sem..enrolled.",  "Displaced")]
test$Target_prob <- predict(modelo, newdata = test, type = "response")
```

```
test$Target_prediction[test$Target_prob >= 0.5] <- 1
test$Target_prediction[test$Target_prob < 0.5] <- 0
```

```
predictions <- test$Target_prediction
```

```
true_values <- dataset_limpio[3001:3630, "Target"]
```

```
check_accuracy(predictions, true_values)
```

```
0.903
```

Conclusiones

Vemos que con un modelo simple de regresión logística hemos conseguido un 90% de precisión. Siempre se trata a las regresiones como modelos débiles pero, usados en el problema correcto, pueden ser muy pontentes.

Recursos

https://www.mdpi.com/2306-5729/7/11/146
https://stats.oarc.ucla.edu/r/dae/logit-regression/
https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

✓ 0 s completado a las 0:25