Análisis de variables influyentes y predicción de Aprobación/Reprobación en un curso de Cálculo I

Ignacio Saavedra B.

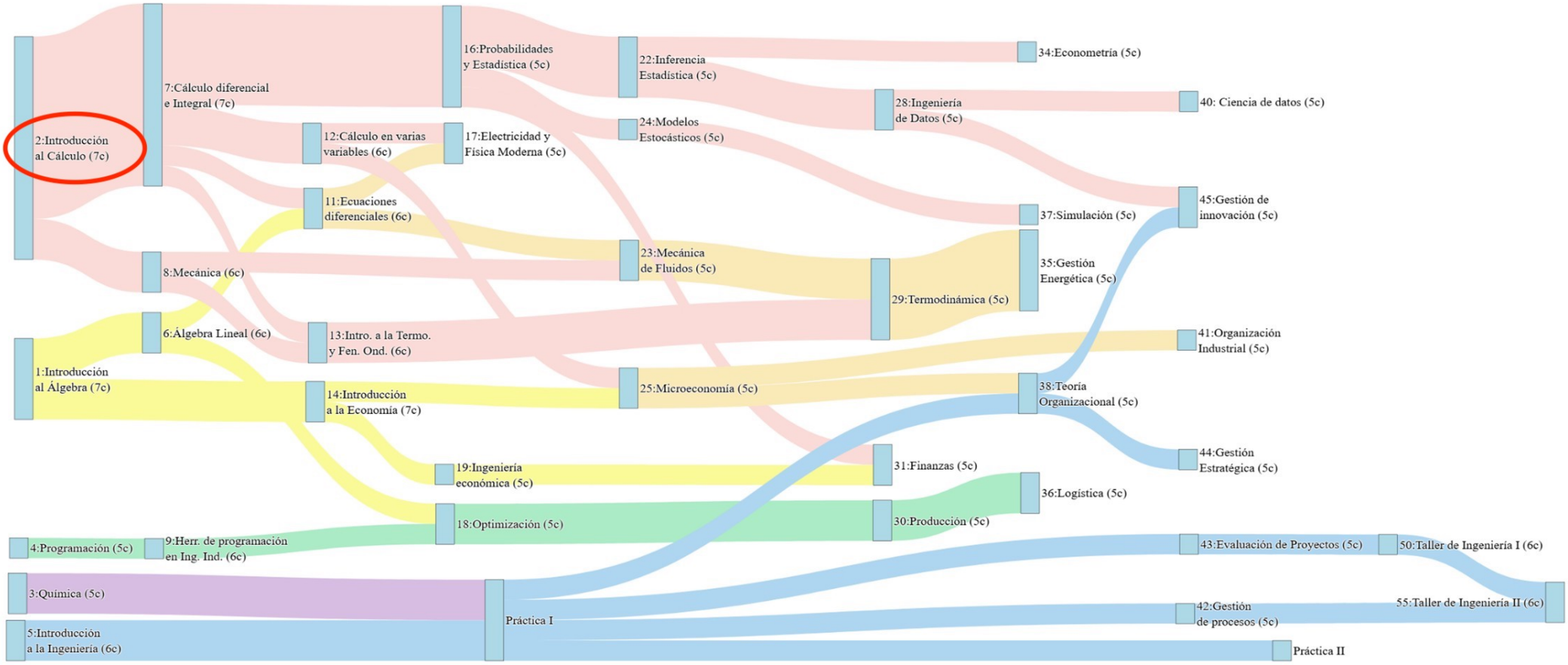
Objetivo

El objetivo del modelo es predecir si un estudiante aprueba o no el curso de Cálculo I. Dada las características del objetivo, se usa el modelo de clasificación Logístico. La predicción debe basarse en variables explicativas que el estudiante posea previamente al inicio del curso. Al ser una asignatura del primer semestre y del primer año, dichas características son en su mayoría, obtenidas de su escolaridad y de pruebas de acceso a la educación superior.

1. Introducción

1.1. Contexto

En las carreras de Ingeniería y, específicamente en el ciclo básico, se ha evidenciado que los cursos de matemáticas son, predominantemente, los cursos más reprobados. Estos cursos inciden directamente en el avance curricular de los estudiantes, en particular, el curso de Cálculo I (o Introducción al Cálculo en la nueva malla) es crítico dada su influencia en otros cursos al ser un *pre-requisito*. En la siguiente página la imagen muestra el diagrama de flujo de los pre-requisitos de los cursos. La aprobación/reprobación de este curso en particular es de mucha importancia ya que incide en el avance curricular, retención/deserción, eliminación, titulación temprana entre otras métricas administrativas e institucionales relevantes dentro del quehacer educativo.


2:Introducción al Cálculo (7c)
7:Cálculo diferencial e Integral (7c)
16:Probabilidades y Estadística (5c)
22:Inferencia Estadística (5c)
34:Econometría (5c)
28:Ingeniería de Datos (5c)
40: Ciencia de datos (5c)
12:Cálculo en varias variables (6c)
17:Electricidad y Física Moderna (5c)
24:Modelos Estocásticos (5c)
11:Ecuaciones diferenciales (6c)
37:Simulación (5c)
45:Gestión de innovación (5c)
8:Mecánica (6c)
23:Mecánica de Fluidos (5c)
29:Termodinámica (5c)
35:Gestión Energética (5c)
6:Álgebra Lineal (6c)
13:Intro. a la Termo. y Fen. Ond. (6c)
41:Organización Industrial (5c)
1:Introducción al Álgebra (7c)
14:Introducción a la Economía (7c)
25:Microeconomía (5c)
38:Teoría Organizacional (5c)
44:Gestión Estratégica (5c)
19:Ingeniería económica (5c)
31:Finanzas (5c)
36:Logística (5c)
18:Optimización (5c)
30:Producción (5c)
4:Programación (5c)
9:Herr. de programación en Ing. Ind. (6c)
43:Evaluación de Proyectos (5c)
50:Taller de Ingeniería I (6c)
3:Química (5c)
Práctica I
42:Gestión de procesos (5c)
55:Taller de Ingeniería II (6c)
5:Introducción a la Ingeniería (6c)
Práctica II

1.2. Objetivo del modelo

Hace 3 años la tasa de reprobación de Cálculo I rondaba el 40 %, hace dos años ha tenido una disminución pasando de una 35 % el 2023 y un 20 % el 2024. El objetivo del modelo es predecir la aprobación/reprobación de los estudiantes del curso de Cálculo I. Con dicha predicción se busca mantener y mejorar los porcentajes de aprobación del curso Cálculo I tomando acciones pedagógicas tempranas con el fin de fortalecer los aprendizajes de estudiantes susceptibles, según el modelo, de reprobación.

2. Descripción del Conjunto de Datos

2.1. Fuentes de datos

Las fuentes de datos son dos:

1. **Admisión:** Esta fuente es una planilla de Excel obtenidas de la oficina de Admisión con los datos estructurados de los estudiantes tales como: SEXO, TIPO COLEGIO, COMUNA COLEGIO, PUNTAJES PAES, PUNTAJES RANKING, PUNTAJE NEM.

2. **ICB (INSTITUTO DE CIENCIAS BÁSICAS):** Esta fuente es una planilla (Google Sheet) de Google Drive la cual tiene los datos estructurados: PUNTAJE DIAGNÓSTICO MATEMÁTICAS y **NOTAS DE LOS ESTUDIANTES DE CÁLCULO I** y otros cursos de matemáticas del primer ciclo.

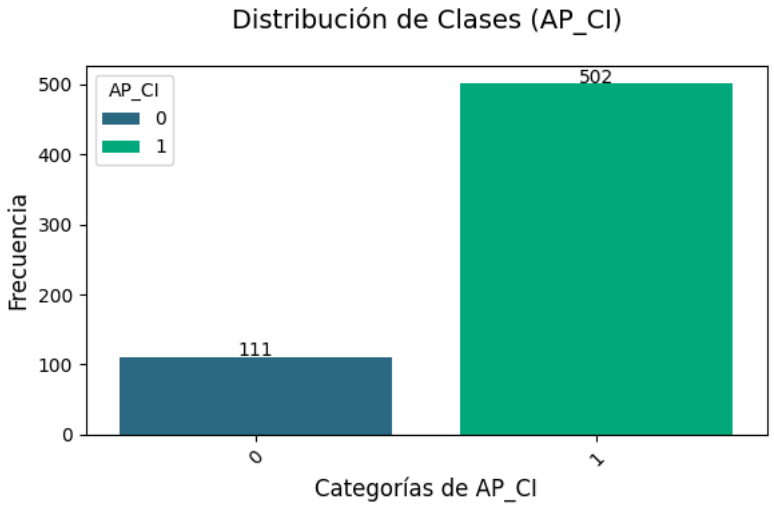
2.2. Distribución de los datos

El número de estudiantes de los cuales tenemos las variables explicativas y objetivo es 613.

Nuestras variables explicativas, en principio son 10, a saber, SEXO, TIPO COLEGIO, COMUNA COLEGIO, PUNTAJE NEM, PUNTAJE RANKING, PUNTAJE LENGUAJE, PAES M1, PAES M2, PUNTAJE CIENCIAS O HISTORIA, DIAGNÓSTICO MATEMÁTICAS.

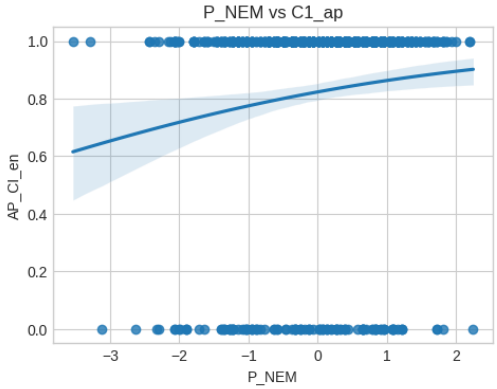
Nuestra variable objetivo es: APRUEBA/REPRUEBA CÁLCULO I, la cual está codificada en 0 si es menor que 4 o 1 si es mayor que 4.

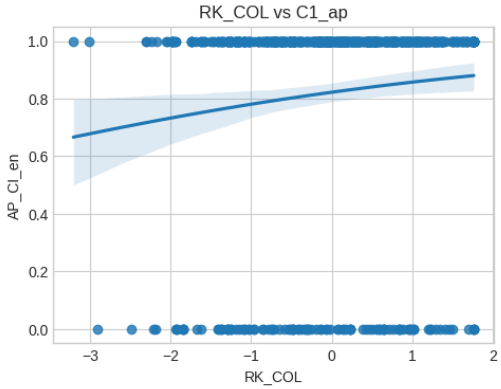
La siguiente gráfica muestra la distribución de la variable objetivo:

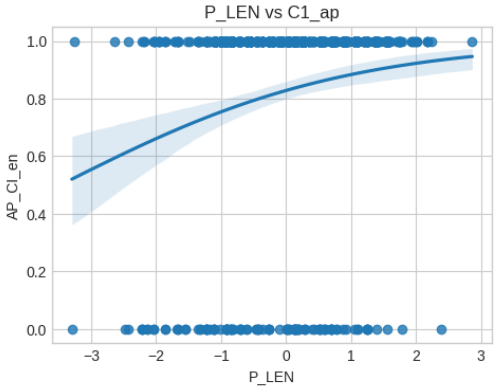


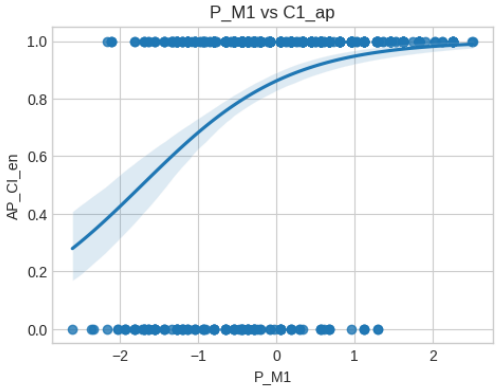
La gráfica nos muestra que la variable objetivo esta distribuida en un 20 % a estudiantes que reprobaron y 80 % que aprobaron el curso.

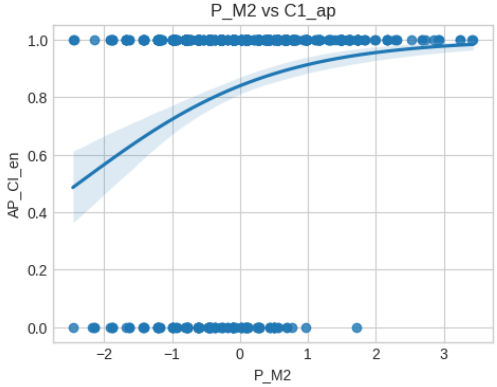
Por último, la siguiente gráfica muestra la relación entre las variables explicativas continuas y la variable objetivo.

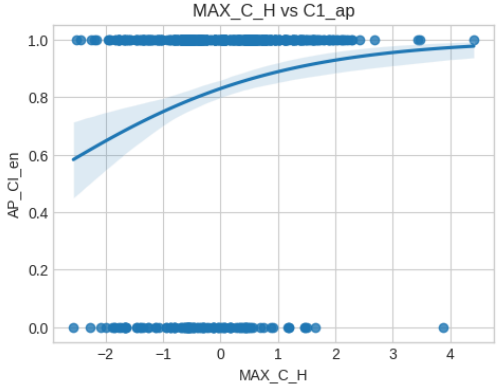

P_NEM vs C1_ap
AP_CI_en
P_NEM

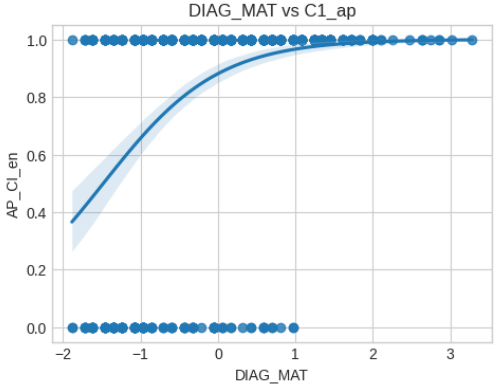

RK_COL vs C1_ap
AP_CI_en
RK_COL

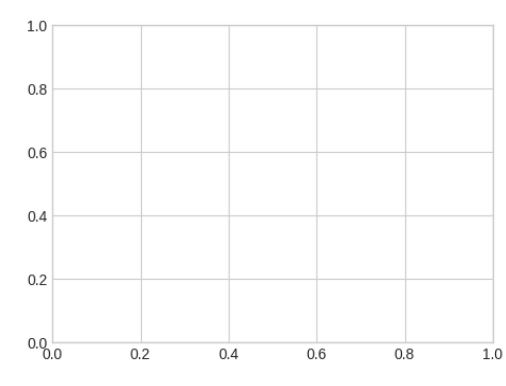

P_LEN vs C1_ap
AP_CI_en
P_LEN


P_M1 vs C1_ap
AP_CI_en
P_M1


P_M2 vs C1_ap
AP_CI_en
P_M2

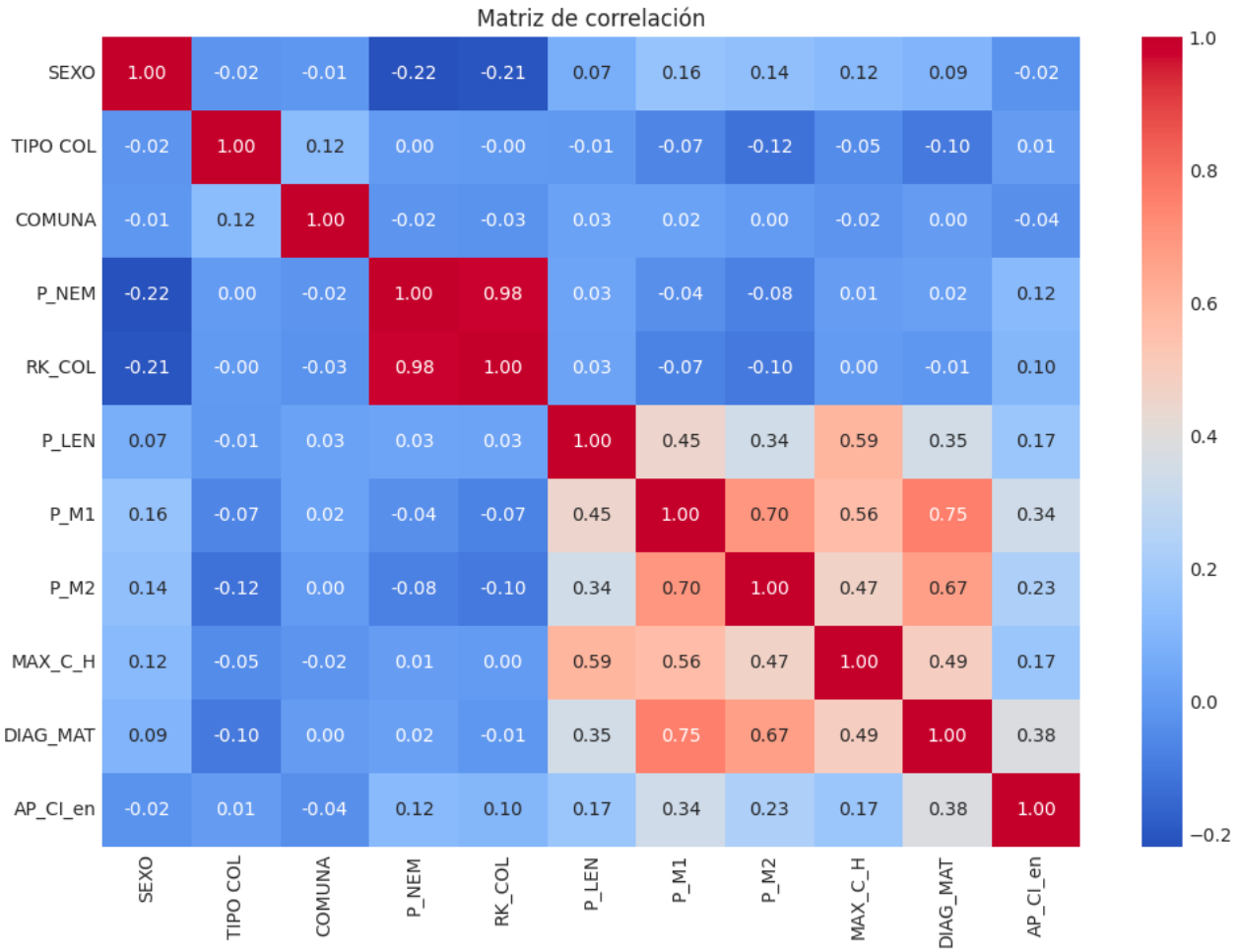

MAX_C_H vs C1_ap
AP_CI_en
MAX_C_H


DIAG_MAT vs C1_ap
AP_CI_en
DIAG_MAT



En la gráfica anterior y en la siguiente gráfica de la matriz de correlación, se observa que, en principio, las variables que más correlación tienen con la variable objetivo (AP_CI_en) son:

- DIAG_MAT = Diagnóstico Matemáticas.
- P_M1 = Puntaje PAES M1.
- P_M2 = Puntaje PAES M2.
- P_LEN = Puntaje PAES LENGUAJE.
- MAX_C_H = Puntaje PAES CIENCIAS O HISTORIA.
- P_NEM = Puntaje NEM.
- RK_COL = Puntaje RANKING COLEGIO.



2.3. Pre-procesamiento

Teniendo en cuenta que el dataset, no tenía datos faltantes u outliers, dado que ya venía pre-procesado, se hizo lo siguiente:

- Codificación de los datos con "etiquetas" usando Label Encoder.
- Normalización/estandarización de los datos dada su naturaleza distinta.
- En principio, se utilizaron todas las características descritas anteriormente.

3. Métodos y Modelos Utilizados

3.1. Modelo

En este punto se compararon 7 modelos de clasificación: Regresión Logística - Random Forest - SVM - KNN - Naive Bayes - XGBoost - Red Neuronal.

Se decidió usar el Modelo Logístico. Junto con el modelo Random Forest, Red Neuronal y XGBoost tienen las mejores métricas y muy parecidas. Sin embargo, dada las caraterísticas (enfoque paramétrico) del modelo Logístico, es que me inclino por él. Esto me permite hacer algunas pruebas estadísticas como el p-value, lo cual me ayuda a reducir el número de variables explicativas y quedarme con las que tienen mayor significancia para mi variable objetivo.

Aquí, al elegir el modelo de Regresión Logística se hicieron algunas pruebas estadísticas, en particular la del p-value que se resumen en la siguiente tabla:

Logit Regression Results

Dep. Variable:	AP_CI_en	No. Observations:	490
Model:	Logit	Df Residuals:	479
Method:	MLE	Df Model:	10
Date:	Fri, 06 Jun 2025	Pseudo R-squ.:	0.2138
Time:	21:27:05	Log-Likelihood:	-178.96
converged:	True	LL-Null:	-227.61
Covariance Type:	nonrobust	LLR p-value:	1.881e-16

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.1264	0.184	11.578	0.000	1.766	2.486
SEXO	-0.0076	0.139	-0.054	0.957	-0.281	0.266
TIPO COL	0.0196	0.145	0.135	0.892	-0.265	0.304
COMUNA	-0.2213	0.138	-1.606	0.108	-0.491	0.049
P_NEM	0.7532	0.619	1.216	0.224	-0.460	1.967
RK_COL	-0.5143	0.604	-0.851	0.395	-1.699	0.670
P_LEN	0.1804	0.169	1.065	0.287	-0.152	0.513
P_M1	0.6668	0.245	2.726	0.006	0.187	1.146
P_M2	-0.0752	0.207	-0.364	0.716	-0.480	0.330
MAX_C_H	-0.4989	0.197	-2.535	0.011	-0.885	-0.113
DIAG_MAT	1.0312	0.245	4.213	0.000	0.552	1.511

Consideramos sólo las que obtuvieron un p-value menor al 0,05 ya que son las que más explicarían nuestra variable objetivo.

	Variable	Coeficiente	Odds_Ratio	P_value	VIF
DIAG_MAT	DIAG_MAT	0.993101	2.699594	0.000025	2.591343
P_M1	P_M1	0.653156	1.921596	0.006413	3.116168
MAX_C_H	MAX_C_H	-0.458738	0.632081	0.011243	1.891279

Estas son, en orden de importancia (Odds Ratio): DIAG_MAT (Diagnóstico Matemáticas), P_M1 (Puntaje Paes_M1) y MAX_C_H (Puntaje máximo(Ciencias, historia)).

Esto permite reducir el número de variables explicativas para la variable objetivo y por consecuente hacer el modelo mucho más simple.

3.2. Parámetros

No consideré una configuración particular en los hiperpárametros en el modelo de regresión logística.

3.3. Partición de datos

En la partición de los datos se usó la configuración estándar

- Partición 80 % para entrenamiento y 20 % para testeo.
- Método de partición aleatorio.

4. Evaluación del Rendimiento del Modelo

Las métricas del modelo de Regresión Logística están resumidas en la siguiente tabla:

📊 Métricas del Modelo (Ordenadas por Valor):

	Métrica	Valor
0	Sensibilidad (Recall)	0.989796
1	F1-Score	0.894009
2	Precisión (Precision)	0.815126
3	Exactitud (Accuracy)	0.813008
4	AUC-ROC	0.757143

5. Interpretación de Resultados

De los resultados obtenidos podemos decir, en general, que teniendo en cuenta que se redujeron las variables a 3, haciendo el modelo muy sencillo y considerando una muestra, en principio desbalanceada, se obtuvieron métricas con valores aceptables. Específicamente,

- **Exactitud (Accuracy):** Se obtuvo un 81 % lo cual es aceptable.
- **Precisión (Precision):** Se obtuvo un 81 % lo cual es aceptable.
- **Sencibilidad (Recall):** Se obtuvo un 99 % lo cual es Bueno.
- **F1-Score:** Se obtuvo un 89 % lo cual es Bueno.
- **AUC-ROC:** Se obtuvo un 75 % lo cual es Aceptable.

Si bien, nos gustarían valores cercanos al 90 % en Exactitud y Precisión, es esperable que estos valores no sean los mejores dada la distribución de nuestra muestra y variable objetivo.

6. Conclusiones

Si bien el modelo Logístico con sólo 3 variables explicativas tiene un desempeño Aceptable en la mayoría de las métricas, es probable que haciendo ajustes en el set de datos como por ejemplo un balanceo de la muestra y variable objetivo podrían mejorar dichas métricas.

Por otra parte, habrá que considerar otros modelos con menos variables explicativas y contrarrestarlas con lo ya obtenido, la limitante está en la complejidad de entender como funcionan los modelos como Red Neuronal o Random Forest que tienen menos acceso a coeficientes o a pruebas estadísticas como el p-value y encontrar pruebas análogas en cada caso.

Por último, es importante considerar que para ser un primer modelo en un dataset real y con un análisis estadístico superficial, las métricas y predicciones son bastante buenas. En efecto, en la prueba de tres estudiantes nos arrojó las siguientes predicciones

Ejemplo	Predicción	Valor real
Estudiante_1	0.968020	1
Estudiante_2	0.532662	0
Estudiante_3	0.959375	1

En este sentido, se prodría considerar en una segunda versión del modelo lo siguiente:

- Balancear la muestra.
- Agregar otras variables explicativas que permitan mejorar los resultados predictorios.
- Realizar transformaciones a las variables explicativas dada la posible no linealidad de estas con la variable objetivo.
- Subir la probabilidad de clasificación "Aprobado", digamos entre 60 % - 80 %