Alcoholismo en secundaria y rendimiento académico

Este proyecto se basa en un conjunto de datos que proporciona una visión detallada de los factores socioeducativos que influyen en el comportamiento de los estudiantes, con un enfoque particular en el consumo de alcohol y el rendimiento académico. Los datos, recopilados a través de una encuesta realizada entre estudiantes de secundaria, ofrecen una oportunidad única para explorar la interacción entre el comportamiento social, el consumo de alcohol y el rendimiento académico.

Autores

Diego Liébana
Sergio Vergara

Citation

P. Cortez e A. Silva. Usando a Mineração de Dados para Prever o Desempenho do Aluno do Ensino Médio. Em A. Brito e J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, abril de 2008, EUROSIS, ISBN 978-9077381-39-7.

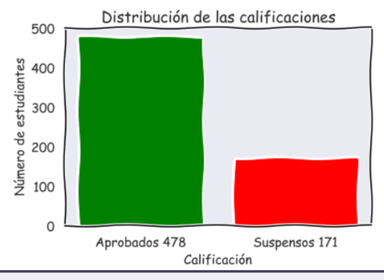
https://xkcd.com/

Dataset

	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	...	freetime	goout	Dalc	Walc	health	absences	G1	G2	Media_G1_G2	Calificacion
0	F	18	U	GT3	A	4	4	at_home	teacher	course	...	3	4	1	1	3	4	0	11	5.5	suspenso
1	F	17	U	GT3	T	1	1	at_home	other	course	...	3	3	1	1	3	2	9	11	10.0	aprobado
2	F	15	U	LE3	T	1	1	at_home	other	other	...	3	2	2	3	3	6	12	13	12.5	aprobado
3	F	15	U	GT3	T	4	2	health	services	home	...	2	2	1	1	5	0	14	14	14.0	aprobado
4	F	16	U	GT3	T	3	3	other	other	home	...	3	2	1	2	5	0	11	13	12.0	aprobado
..	..	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	..	..	...	...
644	F	19	R	GT3	T	2	3	services	other	course	...	4	2	1	2	5	4	10	11	10.5	aprobado
645	F	18	U	LE3	T	3	1	teacher	services	course	...	3	4	1	1	1	4	15	15	15.0	aprobado
646	F	18	U	GT3	T	1	1	other	other	course	...	1	1	1	1	5	6	11	12	11.5	aprobado
647	M	17	U	LE3	T	3	1	services	services	course	...	4	5	3	4	2	6	10	10	10.0	aprobado
648	M	18	R	LE3	T	3	2	services	other	course	...	4	1	3	4	5	4	10	11	10.5	aprobado

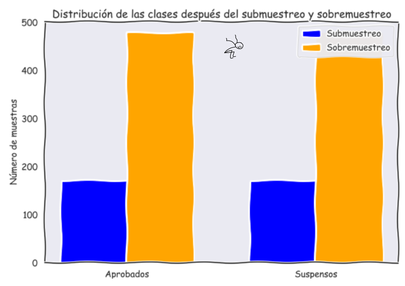
Los atributos objetivo son G1 y G2, que representan las calificaciones de cada semestre.
Se ha creado un 'Media_G1_G2', representa la media de ambos semestres y 'Calificacion' que divide a los estudiantes en : "aprobados" y "suspensos". Este atributo será el objetivo en los modelos de clasificación.

La mayoría se clasifican como "**aprobados**", lleva a un problema de desbalanceo. Para solucionarlo, se ha aplicado "**Random Under Sampling**", técnica que ofrece el mejor resultados.

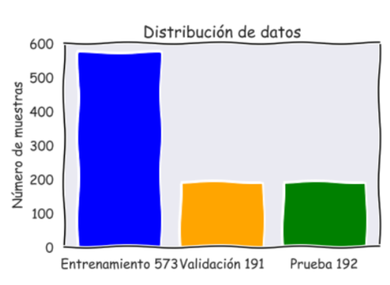


Como resultado, se disponen de tres conjuntos de datos:

- Conjunto de datos **original**
- Conjunto de datos con **submuestreo**
- Conjunto de datos con **sobremuestreo**



train_test_split (
60% Train,
20% Validation,
20% Test)

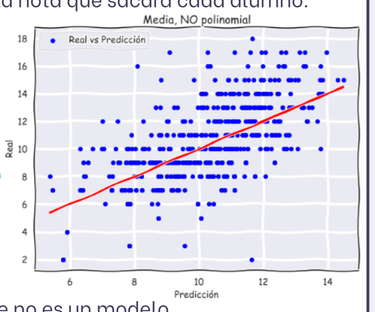


Modelos de Regresión

Regresión Lineal

Se intentan crear modelos para predecir la nota que sacará cada alumno.

```
MSE para G1 =  7.59  ( original )
MSE para G2 =  7.38  ( original )
MSE para Media =  6.9  ( original )

MSE para G1 =  8.1  ( undersampled )
MSE para G2 =  10.0  ( undersampled )
MSE para Media =  8.33  ( undersampled )

MSE para G1 =  6.48  ( oversampled )
MSE para G2 =  8.66  ( oversampled )
MSE para Media =  7.03  ( oversampled )
```

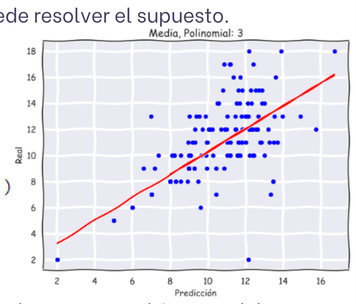


Al tener unos MSE tan altos nos indica que no es un modelo adecuado.

Regresión Polinomial (poli: 3)

Debido al alto valor de las MSE de la regresión lineal se usan polinomiales para ver si de una manera sencilla se puede resolver el supuesto.

```
MSE para G1 =  7.56  ( original )
MSE para G2 =  8.54  ( original )
MSE para Media =  7.17  ( original )

MSE para G1 =  8.52  ( undersampled )
MSE para G2 =  10.85  ( undersampled )
MSE para Media =  8.99  ( undersampled )

MSE para G1 =  4.5  ( oversampled )
MSE para G2 =  5.82  ( oversampled )
MSE para Media =  4.9  ( oversampled )
```



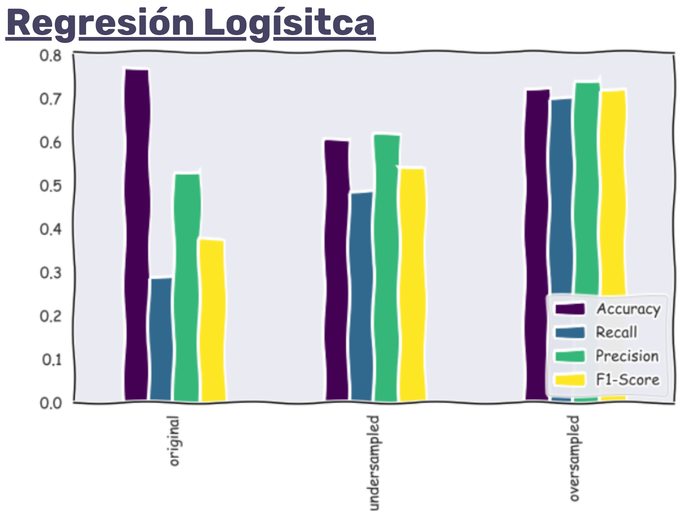
El MSE se reduce, pero no es suficiente. Probaremos un ultimo modelo.

Red Neuronal regresión

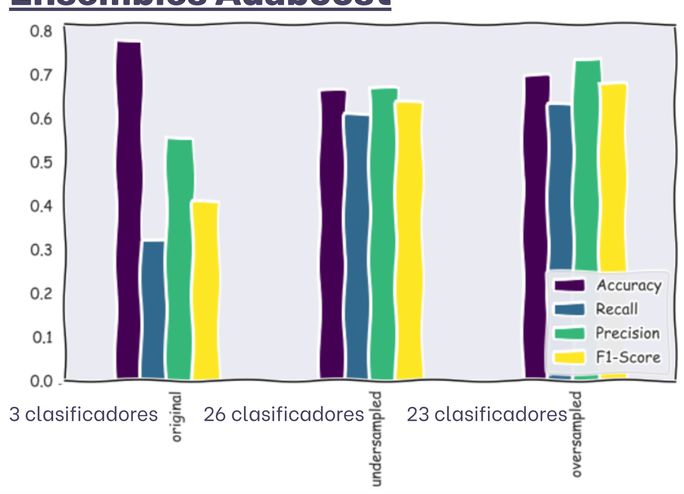
Probamos este ultimo modelo de predicción de notas, el cual ofrece un porcentaje de aciertos del 6%. Este es tan bajo que descartamos definitivamente problemas de regresión de notas.

Modelos de Clasificación

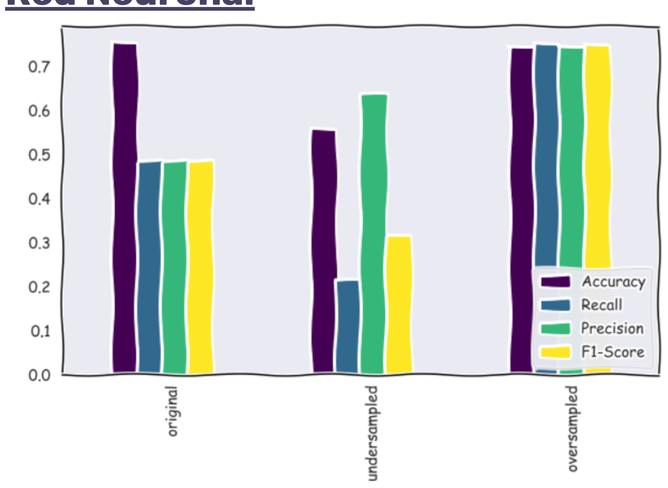
Regresión Logísitca



Ensembles Adaboost



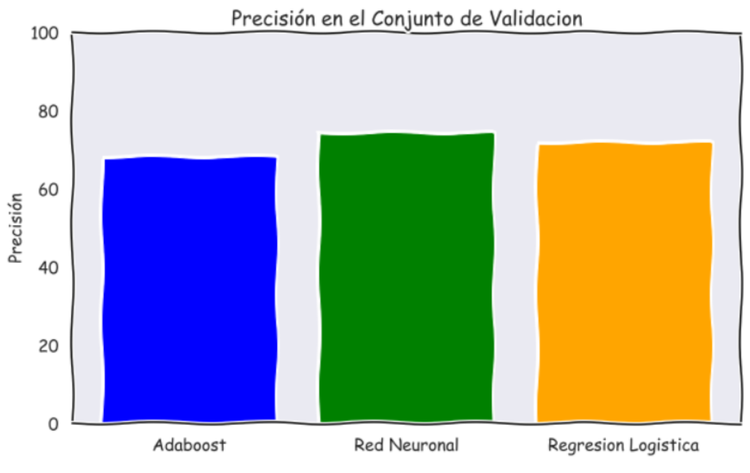
Red Neuronal



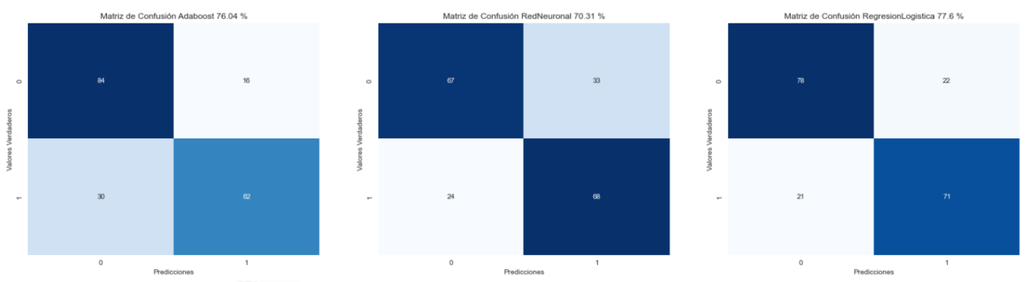
Evaluacion de los modelos

Se comparan las gráficas para cada dataset en cada modelo, observando que en el dataset original muestra, aunque precision elevada, parámetros muy bajos.
Por ello se ha optado por usar los modelos entrenados con Oversampled que ofrecen el mejor resultado en los parametros.
Validamos los modelos para obtener el mejor de ellos para el supuesto.

Nombre	Precisión
Adaboost	68.06
Red Neuronal	74.35
Regresion Logistica	71.73

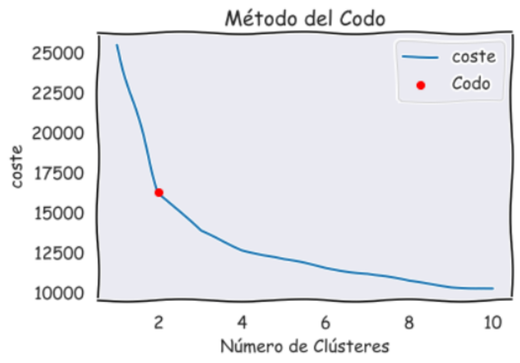


El mejor modelo de clasificacion se concluye en que es la **Red Neuronal** entrenada con el dataset **Oversampled**.
Ofreciendo un rendimiento de **70,31%** en test.



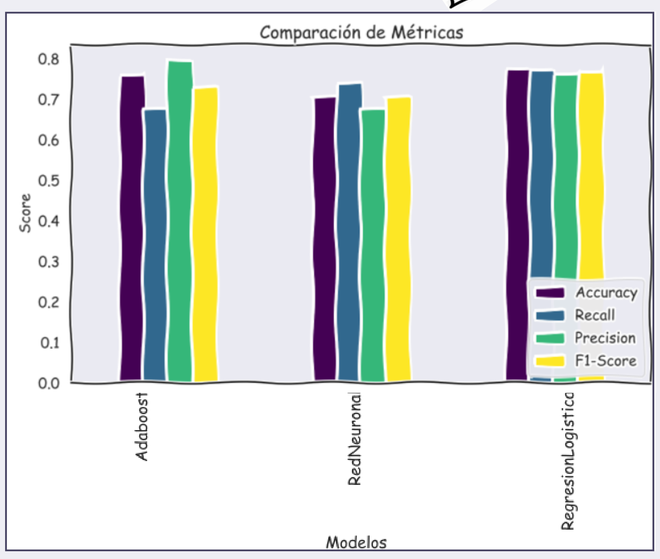
K-mens

Originalmente el dataset tiene los datos sin etiquetar, por lo que es interesante ver como se comporta un modelo de clustering. Por lo que se aplica k-means para estudiar los clusters.




```
Moda para el clúster 0:
   sex  age  address  famsize  Pstatus  Medu  Fedu  traveltime  studytime
0    0   16        1        0        1     4     2           1          2

   schoolsup  ...  Fjob_other  Fjob_services  Fjob_teacher  reason_course
0          0  ...           1              0             0              0

   reason_home  reason_other  reason_reputation  guardian_father  \
0            0             0                  0                0

   guardian_mother  guardian_other
0                1              0

[1 rows x 45 columns]
```

```
Moda para el clúster 1:
   sex  age  address  famsize  Pstatus  Medu  Fedu  traveltime  studytime
0    0   17        1        0        1     2     2           1          2

   schoolsup  ...  Fjob_other  Fjob_services  Fjob_teacher  reason_course
0          0  ...           1              0             0              0

   reason_home  reason_other  reason_reputation  guardian_father  \
0            0             0                  0                0

   guardian_mother  guardian_other
0                1              0
```



Conclusión:

Este estudio resalta la importancia de un adecuado tratamiento y selección de los datos para el entrenamiento de los modelos. Un modelo puede tener una alta precisión, pero si está sesgado hacia una categoría debido a la naturaleza de los datos de entrenamiento, puede no estar funcionando de manera óptima.
En cuanto al consumo de alcohol en los adolescentes, los resultados indican que no es la principal causa de sus bajas calificaciones. En cambio, otros factores, que podrían ser incluso la causa de su consumo de alcohol, parecen tener un impacto más significativo en su rendimiento académico.
Por lo tanto, al abordar el rendimiento académico de los estudiantes, es crucial considerar una variedad de factores y no centrarse únicamente en uno. Este enfoque más holístico permitirá una comprensión más completa de las causas subyacentes del rendimiento académico y puede conducir a intervenciones más efectivas.
En resumen, este análisis subraya la importancia de un enfoque cuidadoso y considerado tanto en el tratamiento de los datos como en la interpretación de los resultados del modelo.