



Desempeño en Matemática



Presentado por:

Adriana Lopez Guerra
Daniela Montoya
Alejandro Ezequiel Martínez
Cristian Andrés Abán

Profesor y tutor:

David Bustos
Rogelio Tobar De La Cruz

Aprender 2018

Data Science





01

Portada

02.

Indice

04.

Tabla de Versionados

05.

Nuestro Equipo

06.

Origen de los datos, objetivos y metas

10.

Diccionario de Variables

14.

Análisis Exploratorio de datos
(EDA)

15.

Análisis Exploratorio de datos (EDA) -
Análisis Univariado

18.

Análisis Exploratorio de datos (EDA) -
Análisis Bivariado

20.

Análisis Exploratorio de datos (EDA) -
Multivariado Bivariado

W
O
D
Z



21.

Análisis Exploratorio de datos (EDA) -
Tratamiento de Valores Nulos

24.

Algoritmos Elegidos

26.

Iteraciones de Optimización

30.

Validación

31.

Futura Lineas

32.

Conclusiones

W
E
D
Z

Tabla de Versionados

VERSIÓN	REFERENCIA
1.0	<p>Primera Entrega se trabajó en:</p> <ul style="list-style-type: none">• Objetivos• Presentación del equipo• Data Set (Origen y criterios)• Data Wrangling y EDA (Analisis Univariado y Bivaridado)
2.0	<p>Segunda Entrega se trabajó en:</p> <ul style="list-style-type: none">• Selección de la variable target• Selección de algoritmos a utilizar.
3.0	<p>Tercera Entrega se trabajó en:</p> <ul style="list-style-type: none">• Se implementaron las mejoras de los modelos• Se encontraron los mejores criterios• Se hizo la validación.
4.0	<p>Entrega final.</p> <ul style="list-style-type: none">• Se siguieron los lineamientos dados por el tutor.

Nuestro Equipo



**Adriana Gabriela
Lopez Guerra**



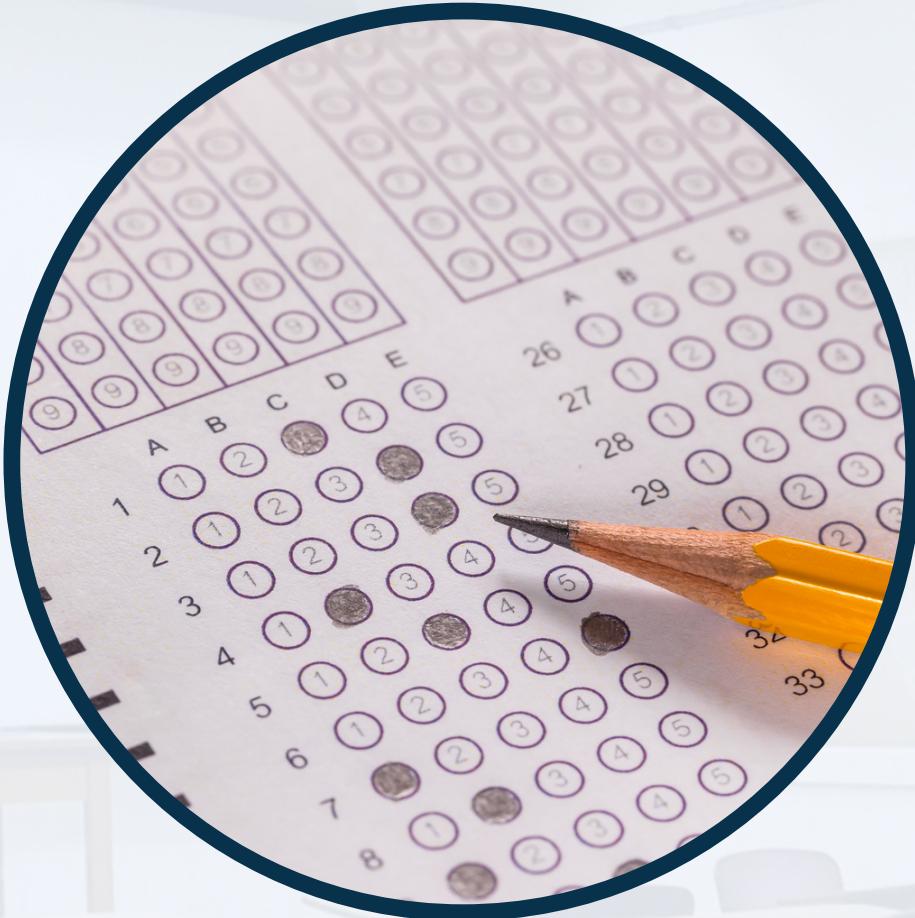
Daniela Montoya



**Alejandro Ezequiel
Martínez**



**Cristian Andrés
Abán M.**



¿Qué es Aprender?

Las evaluaciones Aprender son dispositivos nacionales para medir el nivel de aprendizajes de los estudiantes.

Estas evaluaciones son elaboradas por la Secretaría de Evaluación Educativa del Ministerio de Educación de la Nación Argentina.

Para la confección de las evaluaciones se cuenta con un grupo de profesionales en Educación de todo el país.

El Censo Aprender en 2018 involucró 19645 escuelas y 573939 estudiantes de 6º año del nivel primario.



El Problema

Las estadísticas del año 2018 demuestran un comportamiento alarmante en estudiantes de primaria en la Argentina, ya que más de la mitad de los alumnos tiene problemas para resolver operaciones matemáticas sencillas. El panorama no mejora con otras asignaturas, pero por una ventaja bastante amplia, matemática es la materia que lleva la delantera en cuanto a deficiencia en aprendizaje y enseñanza.



Objetivo

Tomando en cuenta las estadísticas nacionales en el sector educativo argentino y sabiendo que existe una crisis en la enseñanza de la matemática en el país, hemos decidido analizar los datos relacionados con el desempeño de los estudiantes de 6.º grado en el año 2018 en las áreas de Lengua y Matemática para entender la tendencia de estancamiento en el área de matemática específicamente.

La finalidad de este proyecto es entender las variables que pueden afectar el desempeño académico de un alumno y de la misma forma analizarlas con el fin de que estas nos permitan predecir el resultado académico de alumnos futuros.



Origen de los Datos

El Data Set seleccionado contiene los datos nacionales de Argentina, sobre las evaluaciones Aprender del año 2018, las mismas fueron realizadas en todo el territorio nacional.

Las evaluaciones analizadas son las correspondientes al 6 año de la primaria, y nos enfocaremos en los resultados del desempeño en el área de matemática fundamentalmente.

El Data Set se obtuvo del sitio oficial:

(<https://www.argentina.gob.ar/educacion/aprender2018>)

Regiones



Encuestados



Datos recolectados



Área de interés





DataSet

El Data Set cuenta con 123 variables, de las cuales 113 son categóricas y 10 son lineales.

Dentro de las variables categóricas encontramos dos posibles valores que nos indican la falta de respuesta o la respuesta múltiple, para el primer caso es él -6 y para el segundo él -9.

Nuestro "Target" será "mdesemp", esta variable cuenta con 27.479 datos NaN, debido a que esta cantidad representa menos del 0,05% eliminaremos estos registros.

Debido a que el Data Set proviene de un Censo, existe la variable "cod_provincia", esta variable representa con un valor numérico el lugar donde reside el alumno encuestado. Estas categorías serán modificadas, originalmente el Data Set tiene 24 posibilidades de respuestas, las cuales hemos modificado a 5, de esta manera dividiremos todo el país en 5 regiones agrupando varias provincias.

$$\text{AY} = \frac{FL}{T+Y} = \frac{(50)(14)}{(5 \times 10^3)(10^4)} = 1.4 \times 10^{-3}$$
$$L = 4 \text{ MM}; F = 60 \text{ N}; AL = 8 \times 10^3 \text{ MM}$$
$$= \frac{FL}{YAL} = \frac{(60)(14)}{(7 \times 10^3)(8 \times 10^3)} = 1.14 \times 10^{-4}$$
$$AL = 5 \times 10^{-6} \text{ m}; L = 2 \text{ MM}; A = 2 \times 10^{-7}$$
$$YA \left(\frac{AL}{L} \right) = (2 \times 10^{-7}) (2 \times 10^{-6}) \left(\frac{5 \times 10^{-6}}{2} \right) = 114$$
$$= 5 \text{ MM}; P = 6000 \text{ N}; A = 8 \times 10^{-5} \text{ m}^2$$

cada alumbrado : $F = \frac{P}{A} = 3000 \text{ N}$
del alumbrado :

VARIABLES

Se tomaron 19 variables para el análisis de la problemática, el criterio de elección de las mismas fue establecido por los autores del estudio. Las variables fueron renombradas para un mejor entendimiento y un manejo más eficiente de la información. Las variables seleccionadas fueron:

NOMBRE ORIGINAL	QUE REPRESENTA	NOMBRE FINAL
ap1	¿Cuántos años tenés?	edad
ap2	Sexo	sexo
ap3a	¿En qué país naciste?	nacionalidad
ap4	¿Con cuantas personas vives?	personas_vives
ap7a	¿Cuáles de estas cosas hay en el lugar donde vivís? Conexión a Internet	casa_internet
ap7c	¿Cuáles de estas cosas hay en el lugar donde vivís? Computadora	casa_compu
ap9	¿Cuál es el máximo nivel educativo de tu mamá?	mama_niveledu
ap10	¿Cuál es el máximo nivel educativo de tu papá?	papa_niveledu



VARIABLES

NOMBRE ORIGINAL	QUE REPRESENTA	NOMBRE FINAL
ap16	¿Fuiste a jardín de infantes?	jardin_inf
ap17	¿Repetiste de grado alguna vez?	repetiste
ap24	Cuando están trabajando en clase... ¿Tus maestros/as te vuelven a explicar si no entendés?	reexplic_doc
ap26a	Pensado en el último mes, ¿hiciste alguna de estas actividades en tu tiempo libre, fuera del horario escolar? Hice deporte	hacen_deporte
ap26b	Pensado en el último mes, ¿hiciste alguna de estas actividades en tu tiempo libre, fuera del horario escolar? Leí un libro	leen_libros
ap27a	¿Tenés celular propio?	celular_propio
ap27b	¿Tenés celular propio con internet?	celular_internet



VARIABLES

NOMBRE ORIGINAL	QUE REPRESENTA	NOMBRE FINAL
ap31	¿Buscás información o conversás sobre estos temas en internet?	bus_inf
ap35	¿Hacés trabajos en grupo con estudiantes de otros grados?	grupo_otros
ap36a	¿En tu escuela pasan estas cosas? Los compañeros de grados más avanzados ayudan a los más pequeños	ayudan_otros
ap41b	¿Cómo fue resolver las pruebas y este cuestionario? Matemática	interp_mate
comapañeros	¿Te llevás bien con tus compañeros y compañeras?	companeros
cod_provincia	Número de jurisdicción	cod_provincia
sector	Sector de gestión	sector
ambito	Ámbito	ambito



VARIABLES

NOMBRE ORIGINAL	QUE REPRESENTA	NOMBRE FINAL
lpuntaje	Puntaje de Lengua	lpuntaje
mdesemp	Desempeño en Matemática	target
ldesemp	Desempeño en Lengua	ldesemp
mpuntaje	Puntaje de Matemática	mpuntanje
isociao	Indice socioeconómico del alumno	isociao

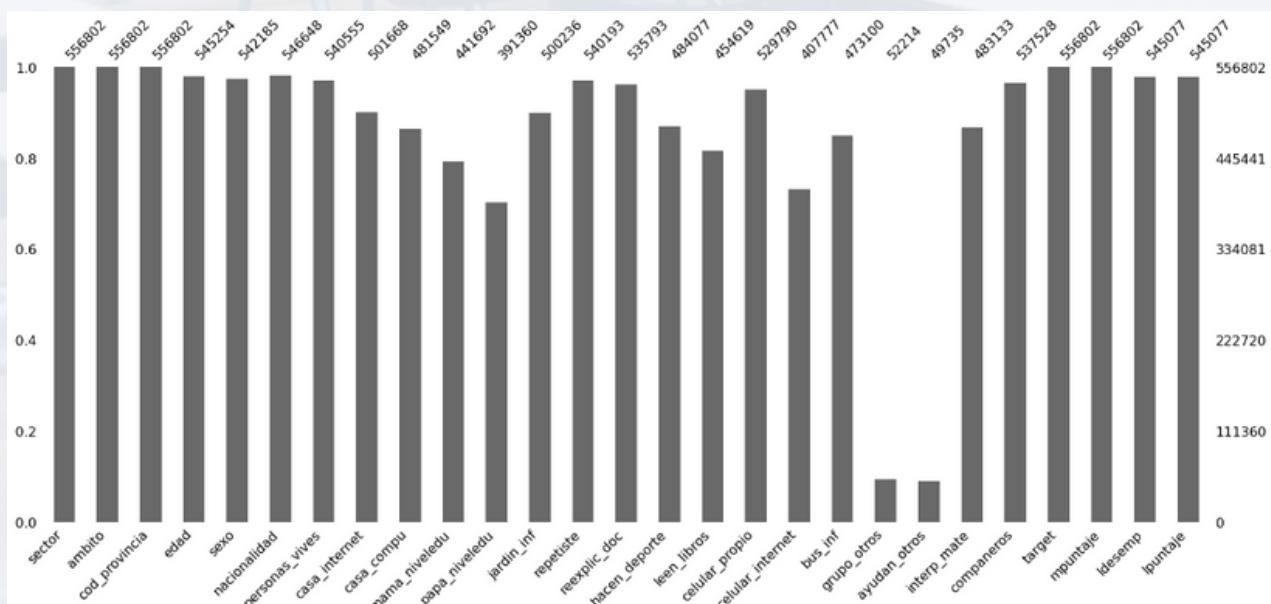
La variable a predecir será "mpuntaje", esta será la variable target del proyecto,



Análisis Exploratorio de Datos (EDA)

Aplicamos el Análisis Exploratorio de Datos (EDA) con las 27 variables que fueron elegidas para este proyecto, implementamos el uso de gráficos y visualizaciones para explorar, entender y aprender sobre los datos.

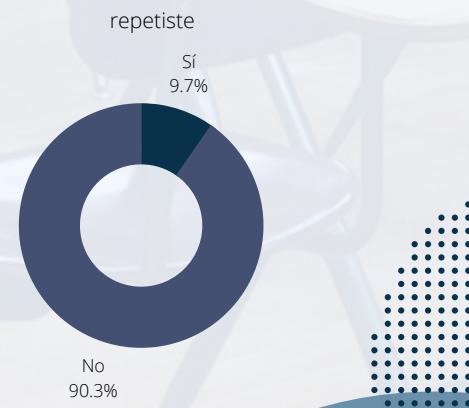
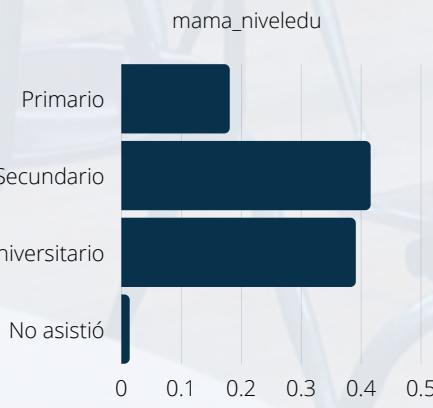
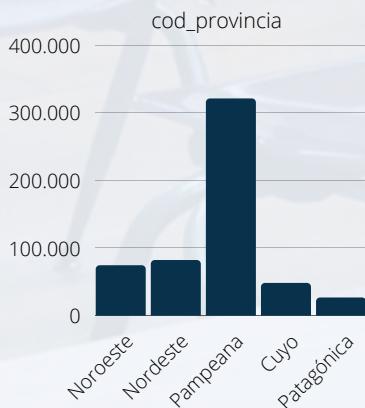
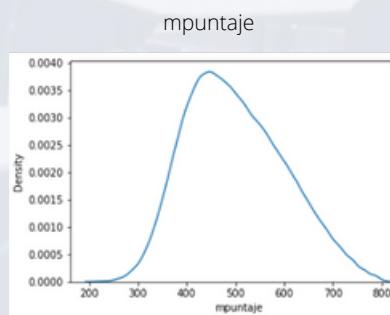
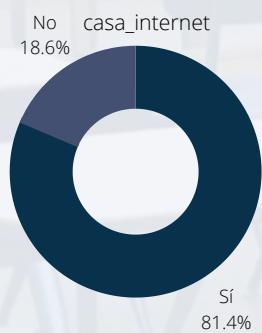
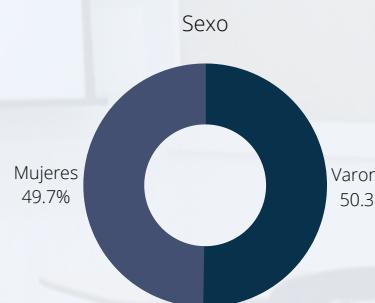
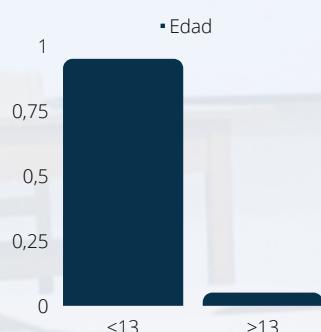
El primer paso de este análisis es ver en qué estado se encuentra la data, ver qué cantidad de nulos tiene cada variable, que tipo de dato es cada variable y de establecer una estrategia para la limpieza del dataset.



Análisis exploratorio de datos (EDA)

Analisis Univariado

Se evaluó individualmente cada una de las variables y para algunas de ellas se realizaron modificaciones a fines de optimizar el análisis.





Análisis exploratorio de datos (EDA)

Análisis Univariado

- La variable "edad", nos indica la edad de los alumnos. El censo se realizó a los alumnos de 6 grados de la primaria, por lo que el 94% de los alumnos tienen menos de 13 años.
- El 5% de los alumnos tiene más de 13 años, esto abarca a aquellos que repitieron el grado escolar como a aquellos alumnos que cumplen los años posterior al mes de julio, lo cual hace que tengan por su fecha de nacimiento un año más.
- El censo se hizo a la misma cantidad de mujeres que de varones.
- El censo se efectuó en Argentina, por lo que más del 97 % de los alumnos son argentinos, y solamente un 3% de los alumnos es extranjero.
- El 66,25% de los estudiantes que respondieron el censo viven con más de 3 personas en su casa.
- El 81.38% de los estudiantes que respondieron el censo tienen internet en su casa.
- El 77.51% de los estudiantes que respondieron tienen una computadora en su casa.
- El 41.53% de los estudiantes que respondieron su mamá tiene un nivel de educación secundaria.
- El 42.95% de los estudiantes que respondieron su papá tiene un nivel de educación secundaria.
- El 98.52% de los estudiantes que respondieron asistieron al jardín de infantes.
- El 90.33% de los estudiantes que respondieron no repitieron de grado.
- El 73.57% de los estudiantes dicen que el docente no vuelve a repetir la explicación si no entendieron el problema.



Análisis exploratorio de datos (EDA)

Análisis Univariado

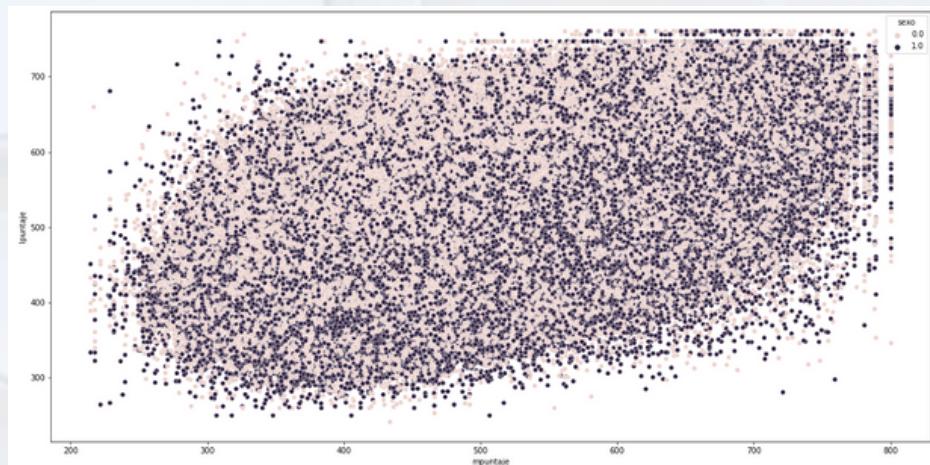
- El 81.95% de los estudiantes que contestaron hacen deporte.
- El 61.3% de los alumnos leen libros en su tiempo libre.
- El 77.7% de los estudiantes poseen celular.
- El 91.28% tiene señal de internet en su celular.
- Nos indica cuantos alumnos realizan trabajos grupales con alumnos de otros cursos.
- Para una gran parte de los alumnos, consideran que aquellos más avanzados ayudan a sus compañeros a realizar tareas.
- Para el 69.6% de los alumnos, fue fácil interpretar los datos de la evaluación matemática.
- La gran mayoría de los encuestados (58.2%) son alumnos de la región pampeana (Buenos Aires, Córdoba, La Pampa y Santa Fe).
- El 69.2% de los estudiantes asisten a escuelas estatales.
- El 88.9% de los estudiantes vive en lugares urbanos.
- Hay más cantidad de resultados de lengua entre 400- 600. Esta distribución de los datos nos muestra que existen pocos valores extremos entre los alumnos encuestados.
- Hay un mayor porcentaje de alumnos aprobó el examen de Lengua.
- Hay una distribución simétrica unimodal similar a una distribución normal para la variable de puntaje en matemática. La mayor cantidad de valores se encuentran entre 400 -600
- La mayoría de los alumnos aprobaron el examen de matemáticas.
- La mayoría de los alumnos posee un índice socioeconómico medio.



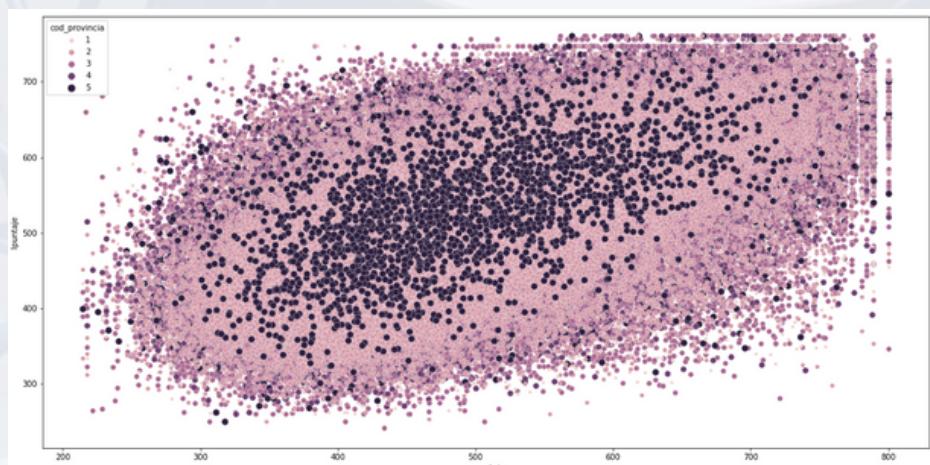
Análisis exploratorio de datos (EDA)

Análisis Bivariado

Comparación del puntaje en lengua y matemáticas respecto al sexo



Comparación del puntaje en lengua y matemáticas respecto a las provincias.

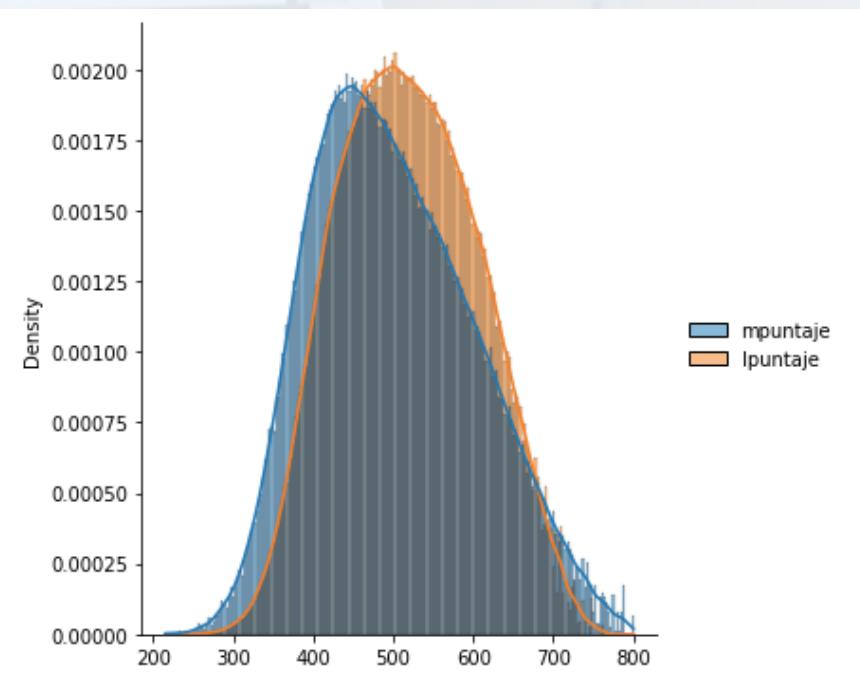




Análisis exploratorio de datos (EDA)

Análisis Bivariado

Comparación del puntaje en lengua y matemáticas respecto a la densidad

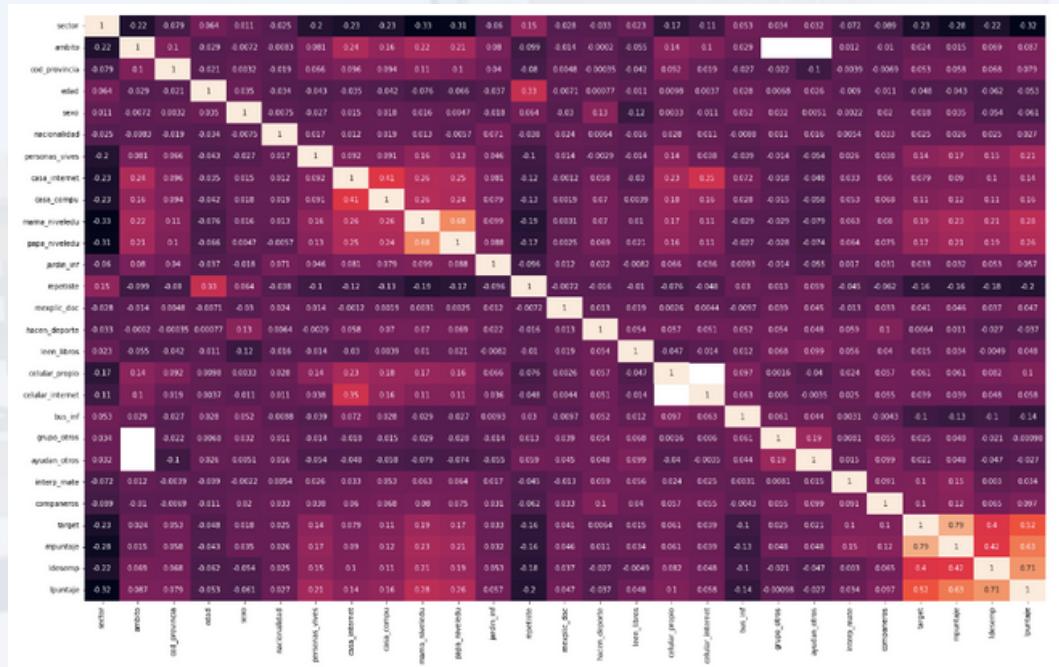




Análisis exploratorio de datos (EDA)

Análisis Multivariado

Mapa de calor para la correlación entre variables



A través del profile del dataset podemos analizar que hay algunas variables que parecen tener mayor correlación entre sí:

- Las variables que hacen referencia al nivel educativo de los padres con un 68%.
- Las variables que hacen referencia a tener internet en casa y tener una computadora en casa con 41%.
- La variable de edad con la variable que representa si un alumno ha repetido.



Análisis exploratorio de datos (EDA)

Tratamiento de Valores Nulos

Como se observó al principio, algunas variables del data set presentaban valores nulos, para poder continuar con el análisis de los datos y poder usar el data set en los procesos futuros se implementó la imputación sensible y la imputación predictiva.

Imputación Sensible

- La variable **edad** tenía un 2% de valores nulos, para no modificar los porcentajes de las diferentes edades, decidimos asignar valores de las categorías "11" y "12" de forma equitativa.
- La variable **nacionalidad** tenía menos de 2% de valores nulos, para este caso, ya que la encuesta se realizó en Argentina, se decidió asignar como nacionalidad argentina a todos los valores nulos de esta categoría, porque la probabilidad de ser extranjero es muy pequeña.
- La variable **sexo** tenía un poco más de 2% de valores nulos y porque la cantidad de niños y niñas es casi la misma, se decidió asignar de forma equitativa la misma cantidad de femenino y masculino.
- La variable **casa_internet** se decidió tratar con valores de 70/30, ya que estos son los porcentajes de respuesta obtenidos.



Análisis exploratorio de datos (EDA)

Imputación Predictiva

Se procedió a utilizar la función "fancyimpute" y se utilizó el método MICE, este realiza una regresión múltiple sobre los datos de la muestra y toma las medias de los mismos.

sector	0	leen_libros	27199
ambito	0	celular_propio	27199
cod_provincia	0	celular_internet	27199
edad	0	bus_inf	27199
sexo	0	grupo Otros	27199
nacionalidad	0	ayudan_Otros	27199
personas_vives	27199	interp_mate	27199
casa_internet	0	companeros	27199
casa_compu	27199	target	0
mama_niveledu	115110	mpuntaje	0
papa_niveledu	165442	ldesemp	11725
jardin_inf	27199	lpuntaje	11725
repetiste	27199	dtype:	int64

- Se eliminaron los nulos que se repetían en varias variables a la vez.
- Se unificaron las variables **mama_niveledu** y **papa_niveledu**, se tomó el valor más alto entre ambas variables para la creación de la variable **familia_edu**, esto permitió la reducción en valores nulos.
- Se implementaron dummies para la variable **cod_provincia**.



Análisis exploratorio de datos (EDA)

Resultado Final del Data Set

Como resultado final tenemos un data set de 30 variables, con un total de 427 943 registros, este es el data set que será usado para el entrenamiento de algoritmos de clasificación.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 427943 entries, 0 to 556799
Data columns (total 30 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   sector            427943 non-null   int64  
 1   ambito             427943 non-null   int64  
 2   edad               427943 non-null   float64 
 3   sexo               427943 non-null   float64 
 4   nacionalidad       427943 non-null   float64 
 5   personas_vives     427943 non-null   float64 
 6   casa_internet      427943 non-null   float64 
 7   casa_compu          427943 non-null   float64 
 8   jardin_inf          427943 non-null   float64 
 9   repetiste           427943 non-null   float64 
 10  reexplic_doc        427943 non-null   float64 
 11  hacen_deporte       427943 non-null   float64 
 12  leen_libros          427943 non-null   float64 
 13  celular_propio      427943 non-null   float64 
 14  celular_internet    427943 non-null   float64 
 15  bus_inf              427943 non-null   float64 
 16  grupo_otros          427943 non-null   float64 
 17  ayudan_otros          427943 non-null   float64 
 18  interp_mate           427943 non-null   float64 
 19  edu_0.0                427943 non-null   uint8  
 20  edu_1.0                427943 non-null   uint8  
 21  edu_2.0                427943 non-null   uint8  
 22  edu_3.0                427943 non-null   uint8  
 23  provincia_1             427943 non-null   uint8  
 24  provincia_2             427943 non-null   uint8  
 25  provincia_3             427943 non-null   uint8  
 26  provincia_4             427943 non-null   uint8  
 27  provincia_5             427943 non-null   uint8  
 28  target                  427943 non-null   float64 
 29  ldesemp                 427943 non-null   float64 

dtypes: float64(19), int64(2), uint8(9)
memory usage: 75.5 MB
```



Algoritmos Elegidos

Los algoritmos elegidos para este proyecto fueron:

- Random Forest
- Regresión Logística
- SGD Classification
- Ridge Classifier
- k-Nearest-Neighbor Classifier
- Bagging classifier
- LightGBM Classifier
- XGBClassifier

Se procede a la creación de una tabla donde se guardan todos los resultados de los modelos entrenados.

```
cols = ['Case', 'SGD', 'Ridge', 'KNN', 'Bagging', 'RndForest', 'LogReg', 'LGB', 'XGBR']

resul = pd.DataFrame(columns=cols)
resul.set_index("Case", inplace=True)
resul.loc['Accuracy'] = [0,0,0,0,0,0,0,0]
resul.loc['F1'] = [0,0,0,0,0,0,0,0]
resul.loc['Precision'] = [0,0,0,0,0,0,0,0]
resul.loc['Recall'] = [0,0,0,0,0,0,0,0]

models = [sgd,ridge,knn,bag,rf,lr,lgg,xgb]

col = 0
for model in models:
    model.fit(X_train_f,y_train_f)
    resul.iloc[0,col] = model.score(X_test_f,y_test_f)
    y_pred = model.predict(X_test_f)
    resul.iloc[1,col] = f1_score(y_test_f, y_pred)
    resul.iloc[2,col] = precision_score(y_test_f, y_pred)
    resul.iloc[3,col] = recall_score(y_test_f, y_pred)
    col += 1

resul.head()
```



Algoritmos Elegidos

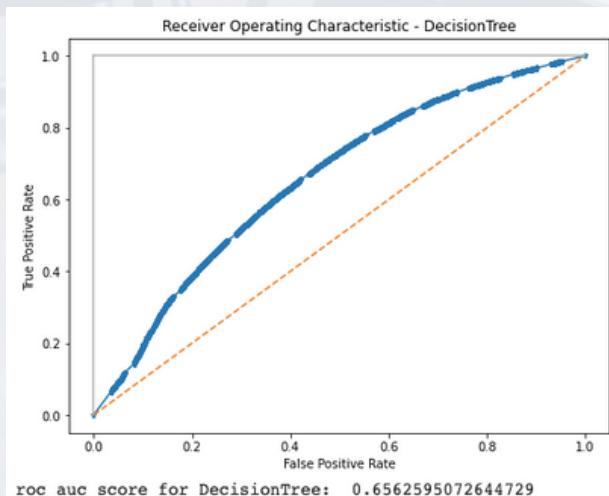
Los resultados obtenidos fueron distribuidos en una tabla que nos permite hacer una comparación eficiente con respecto al aspecto que se decida ponderar como más útil.

	SGD	Ridge	KNN	Bagging	RndForest	LogReg	LGB	XGBR
Case								
Accuracy	0.681827	0.684369	0.630752	0.64359	0.647406	0.685375	0.686968	0.687173
F1	0.779317	0.777763	0.710821	0.72398	0.727667	0.77701	0.773017	0.775253
Precision	0.67144	0.677525	0.675503	0.681183	0.68304	0.680198	0.688641	0.6857
Recall	0.928495	0.912812	0.750035	0.772517	0.778532	0.905954	0.880957	0.891712

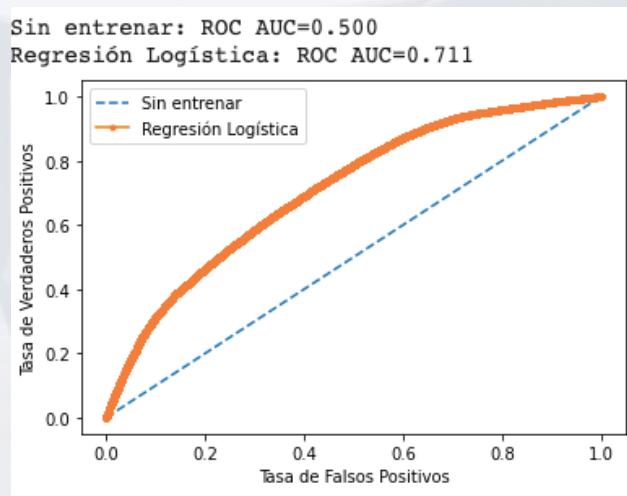
Curvas ROC

La curva ROC (Receiver Operating Characteristic) se utiliza para evaluar el rendimiento de los algoritmos de clasificación binaria, es decir, entre dos clases o categorías (1 o 0, Verdadero o Falso, etc.).

RandomForestClassifier



LogisticRegression





Iteraciones de Optimización

Para la optimización tomamos una fracción del data set, representada por el 1%, ya que el tamaño del mismo hace muy complicada la ejecución.

```
df_sample=dfinal.sample(frac=0.01,random_state=24)

#Separamos en X e y la X mayuscula por que es matriz y es vector
X = df_sample.drop("target", axis=1) #Elimino de mi dataset la variable a predecir
y = df_sample.target #Defino el Target
#Me quedo con 33% para test y 67% para train
X_train_s, X_test_s, y_train_s, y_test_s = train_test_split(X,y,train_size=0.67,test_size=0.33,random_state=42)
#Aquí elijo train y test para la muestra aleatoria
```

Hyperparameter Tuning

La optimización o ajuste de hiperparametros es el proceso de elegir el conjunto de parámetros más óptimos para un algoritmo de aprendizaje, esto permite la mejora del modelo.

Optimización de Random Forest

Utilizamos GridSearchCV para optimizar el Random Forest

```
n_estimators=[int(x) for x in np.linspace(start=10, stop=80, num=10)]
max_features=['auto','sqrt']
max_depth=[2,4,8]
min_samples_split=[2,5,7]
min_samples_leaf=[1,2,4]
bootstrap=[True,False]
random_state=[42]
```



Iteraciones de Optimización

Optimización de RandomForest

```
param_grid={'n_estimators' : n_estimators,
'max_features' : max_features,
'max_depth': max_depth,
'min_samples_split' : min_samples_split,
'min_samples_leaf' : min_samples_leaf,
'bootstrap': bootstrap,
'random_state':random_state
}'
```

```
rf_Model=RandomForestClassifier()
rf_Grid=GridSearchCV(estimator=rf_Model,param_grid=param_grid,cv=5,verbose=2,n_jobs=4)
rf_Grid.fit(X_train_s,y_train_s)
```

```
Fitting 5 folds for each of 1080 candidates, totalling 5400 fits
GridSearchCV(cv=5, estimator=RandomForestClassifier(), n_jobs=4,
            param_grid={'bootstrap': [True, False], 'max_depth': [2, 4, 8],
                        'max_features': ['auto', 'sqrt'],
                        'min_samples_leaf': [1, 2, 4],
                        'min_samples_split': [2, 5, 7],
                        'n_estimators': [10, 17, 25, 33, 41, 48, 56, 64, 72,
                                         80],
                        'random_state': [42]}),
            verbose=2)
```



Iteraciones de Optimización

Optimización de Random Forest

Mejores parámetros:

- bootstrap: 'False'
- max_depth: 4
- max_features: 'auto'
- min_samples_leaf: 4
- min_samples_split: 2
- n_estimators: 10
- random_state: 42

Resultado optimizado de Ramdon Forest en del DataFrame completo.

	precision	recall	f1-score	support
0.0	0.72	0.30	0.42	55774
1.0	0.67	0.92	0.78	85448
accuracy			0.68	141222
macro avg	0.70	0.61	0.60	141222
weighted avg	0.69	0.68	0.64	141222



Iteraciones de Optimización

Optimización de la Regresión Logística

Utilizamos GridSearchCV para optimizar la Regresión Logística

```
param_grid_lr=[  
    {'penalty' : ['l2'],  
     'C': np.logspace(-4,4,20),  
     'solver': ['lbfgs','newton-cg','liblinear','sag','saga'],  
     'max_iter':[100,1000,2500,5000],  
     'random_state':[24,42,123,0]  
}  
]  
  
clf_lr=GridSearchCV(logModel,param_grid=param_grid_lr,cv=5,verbose=True,n_jobs=-1)  
best_clf_lr=clf_lr.fit(X_train_s,y_train_s)  
  
Fitting 5 folds for each of 1600 candidates, totalling 8000 fits
```

Mejores parámetros:

- C: 1.623776739188721
- max_iter: 100
- penalty: l2
- random_state: 24
- solver: liblinear



Iteraciones de Optimización

Optimización de la Regresión Logística

Resultado optimizado de la Regresión Logística en del DataFrame completo.

	precision	recall	f1-score	support
0.0	0.71	0.35	0.47	55774
1.0	0.68	0.91	0.78	85448
accuracy			0.69	141222
macro avg	0.69	0.63	0.62	141222
weighted avg	0.69	0.69	0.65	141222



Validación

Validación simple

Se aplicó la validación del modelo obtenido a través de Random Forest utilizando hiperparámetro de "gini" y max_depth=4.

```
from sklearn.model_selection import cross_val_score
from numpy import mean, std

# Crear modelo
model = RandomForestClassifier(random_state=1,n_estimators=30, criterion="gini",max_depth=4)

# Evaluar el modelo
scores = cross_val_score(model,X_s,y_s,scoring='accuracy')

scores

array([0.67172897, 0.69042056, 0.66471963, 0.67406542, 0.67368421])

# Reportar el performance
print('Accuracy: %.3f (%.3f)' % (mean(scores), std(scores)))

Accuracy: 0.675 (0.008)
```

Obtenemos un accuracy de 0.675, o lo que equivale a una dispersión de 0,008; lo que nos indica que el modelo generado debería poder ser capaz de predecir los resultados.

Futuras Líneas

Para la mejora de este proyecto creemos que podría ser bastante interesante analizar otros años con la finalidad de ver cómo se comportan las variables predictoras a través del tiempo, de igual forma se podría hacer la misma evaluación para la predicción del desempeño de los alumnos en el área de lenguaje.

Otro posible análisis sería ver el desempeño de cada región ya que cada uno utiliza metodologías pedagógicas diferentes.

