DATA SCIENCE – PROYECTO FINAL

EXONERACIÓN EDUCATIVA

NICOLÁS CLADERA

ÍNDICE

1.

PRESENTACIÓN

Nicolás Cladera Rissolini
26 años
Uruguay

Formación Profesional:

- Licenciatura en Economía (UdelaR)
- Data Analytics (CODER)
- Tableau (CODER)
- English Proficiency (Alianza)

Experiencia Laboral:

- Data Analyst – DINOT (Dirección Nacional de Ordenamiento Territorial)
- Analista Comercial – Cooperativa Acac

2.

ABSTRACT

ABSTRACT - CONTEXTO COMERCIAL

En el siguiente trabajo se analiza un dataset con información sobre las calificaciones de exámenes de secundaria de alrededor de 30.000 alumnos, así como ciertas características socioeconómicas de los mismos, como sexo, etnicidad, estudios de los padres, y otras variables que se explicarán más con detalle en la sección correspondiente. Se intenta explicar si dichas características pueden tener influencia o no en la probabilidad de un alumno para aprobar los exámenes (obtener 50 puntos o más) o exonerar las asignaturas (obtener 70 o más).

Con la educación siendo un importante indicador para el éxito económico y social de las naciones, y teniendo en cuenta que varios países de Latinoamérica están experimentando reformas educativas en este momento, o tienen planes de hacerlo en el corto plazo, parece oportuno identificar áreas fuera de lo que es estrictamente educativo que puedan estar alterando los resultados educativos. Además, un modelo que pueda estimar un resultado educativo promedio de un grupo de estudiantes en un centro educativo puede ayudar a encarar posibles problemas antes de que se hagan evidentes como resultado de exámenes.

PREGUNTAS A RESPONDER

Con este análisis se intentan responder preguntas como:

- ¿Son estadísticamente significativas las diferencias de sexo o etnicidad para explicar las diferencias en los resultados educativos de los alumnos?
- ¿Son estadísticamente significativas las diferencias en contextos familiares como el nivel educativo de los padres o el nivel socioeconómico para explicar las diferencias en los resultados educativos de los alumnos?
- ¿Es posible predecir si un alumno tiene mayor probabilidad de aprobar las pruebas o exonerar las asignaturas (o el curso en general) basado en variables diferentes a la cantidad de horas de estudio del alumno?

3.

DATA ACQUISITION

RESUMEN DE METADATA

Fuente de la base de datos: Kaggle

Número de observaciones: 30.000

Descripción de las variables:

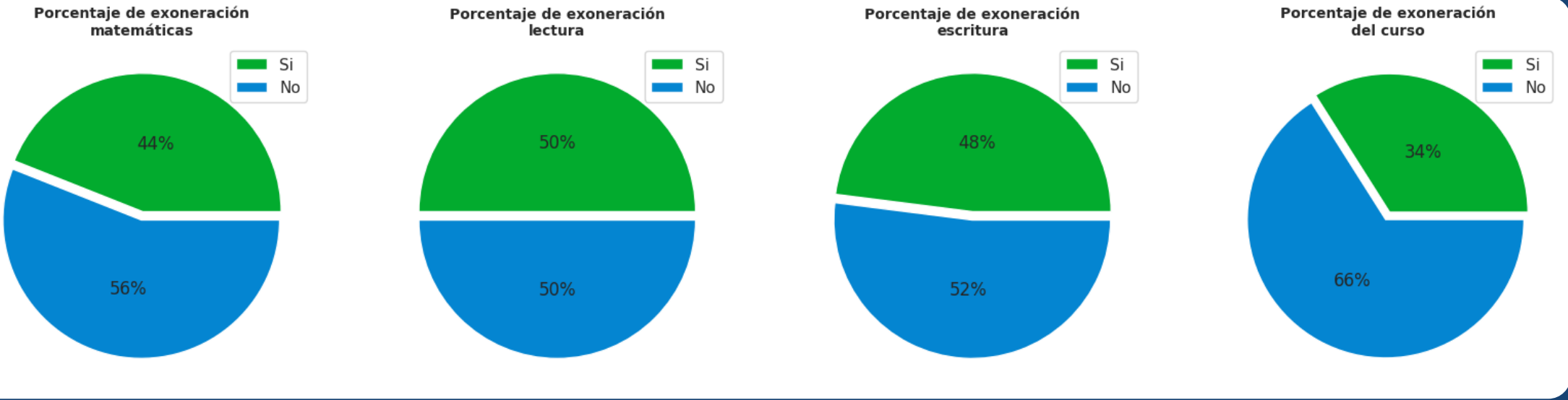
- ID = Identificador del estudiante
- Sexo = Sexo del estudiante (masculino o femenino)
- Etnia = Etnia del estudiante, dividido en 5 grupos (Blanco, Afroamericano, Hispano, Asiático, Nativo Americano)
- Educ_padres = Nivel educativo de los padres del estudiante, toma los valores de liceo incompleto, liceo completo, terciarios incompeltos, estudios técnicos y diploma universitario.
- Tipo_almuerzo = Indica si el estudiante consume el almuerzo estándar o el reducido/gratis.
- Curso_prep = Indica si el estudiante completó o no el curso de preparación para los exámenes.
- Estado_civil_padres = Indica si los padres están casados, divorciados, viudos o si son padres solteros.
- Deporte = Indica si el estudiante practica deportes, varía entre nunca, a veces y regularmente.
- Primer_hijo = Indica "Si" si el estudiante fue el primer hijo de la familia, "No" de lo contrario.
- Hermanos = Indica el número de hermanos del estudiante.
- Transporte = Indica si el estudiante va a la institución educativa en el bus escolar o en transporte privado.
- Estudio_semanal = Indica las horas semanales de estudio que dedica en promedio el estudiante. Varia entre <5, 5-10 y >10.
- Score_mat = Indica el puntaje en la prueba de matemáticas (sobre 100)
- Score_lect = Indica el puntaje en la prueba de lectura (sobre 100)
- Score_esc = Indica el puntaje en la prueba de escritura (sobre 100)

Se agregan también 3 variables dummies que valen 1 si el estudiante aprobó las pruebas de cada asignatura y otras 3 que valen 1 si el estudiante exoneró los cursos.

4.

EDA Y FEATURE ENGINEERING

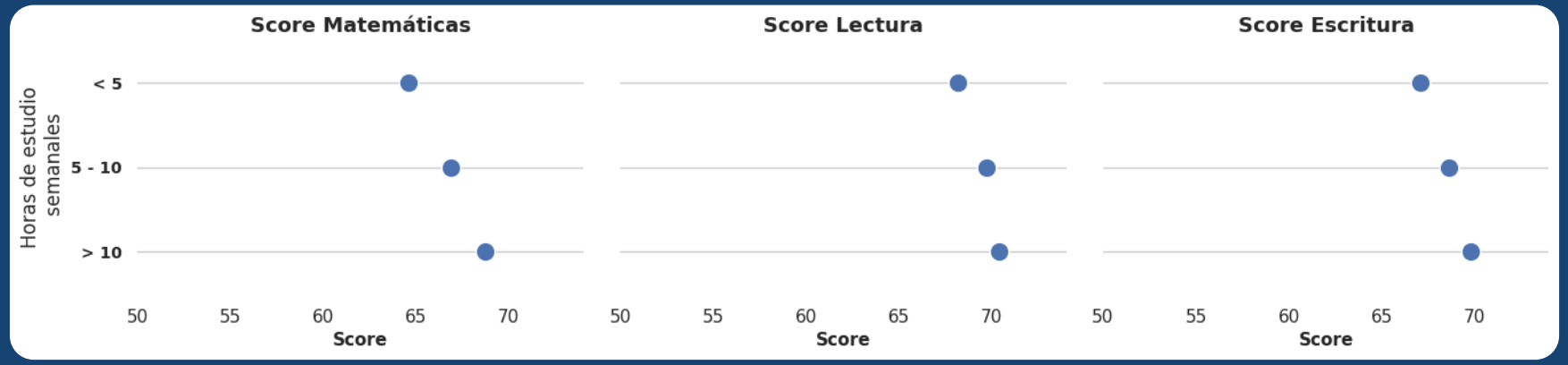
ANÁLISIS UNIVARIADO Y BIVARIADO

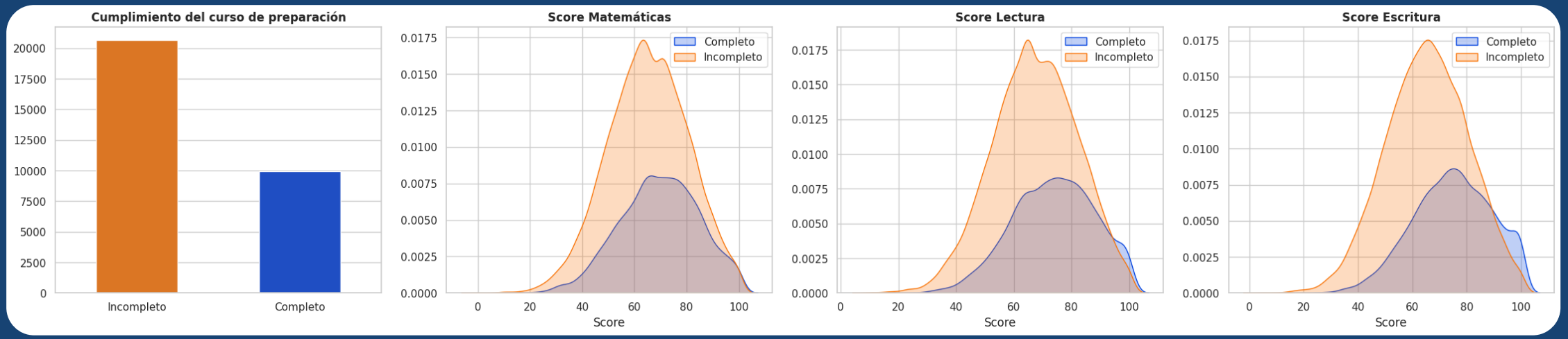


Los alumnos no consiguen más de un 50% de exoneración en ninguna de las asignaturas, siendo el peor matemáticas, con un 44% de exoneración.

Se puede ver la necesidad de crear un modelo de este tipo, ya que los porcentajes de exoneración del curso en el centro educativo de prueba son relativamente bajos, de un 34%.

ANÁLISIS UNIVARIADO Y BIVARIADO

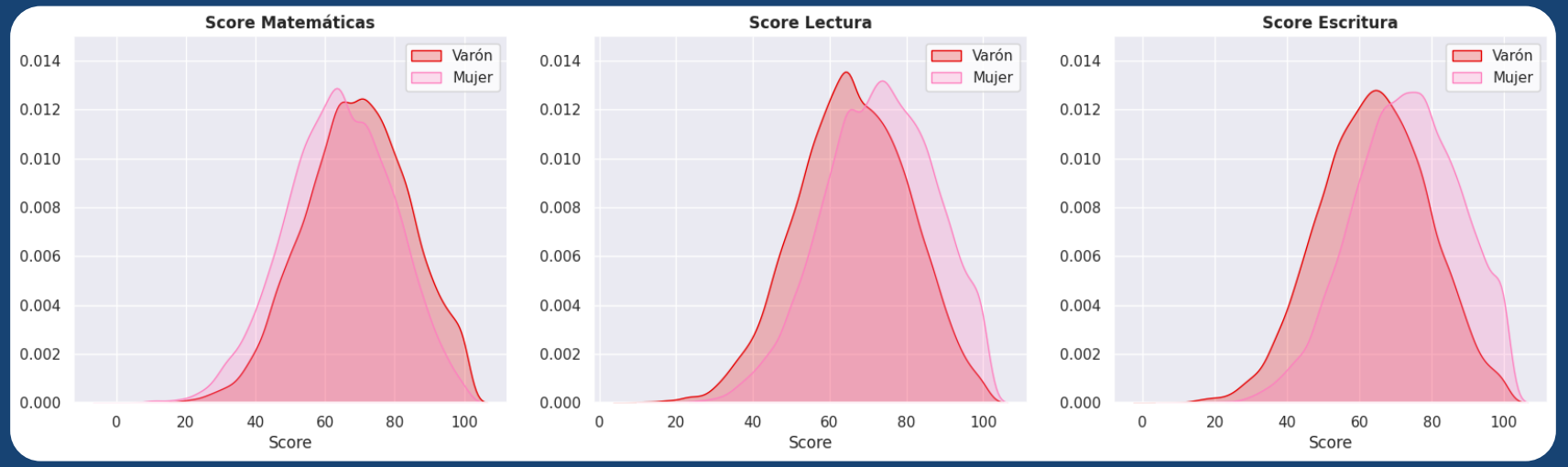


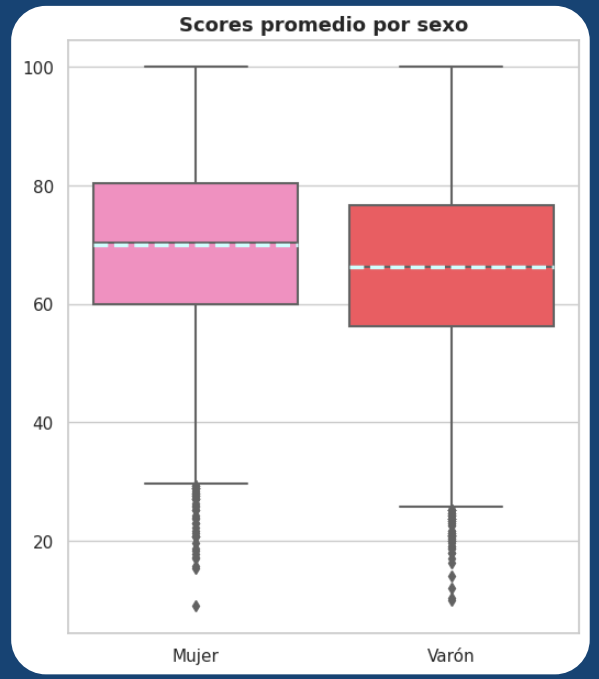


Como se mencionó anteriormente la idea es agregar variables que no estén relacionadas con el estudio al modelo, pero no se puede dejar afuera las mismas, ya que hay una clara tendencia a una mejora en las calificaciones cuando se estudia una mayor cantidad de horas.

Además los alumnos que terminan el curso de preparación tienden a tener un rendimiento marginalmente superior.

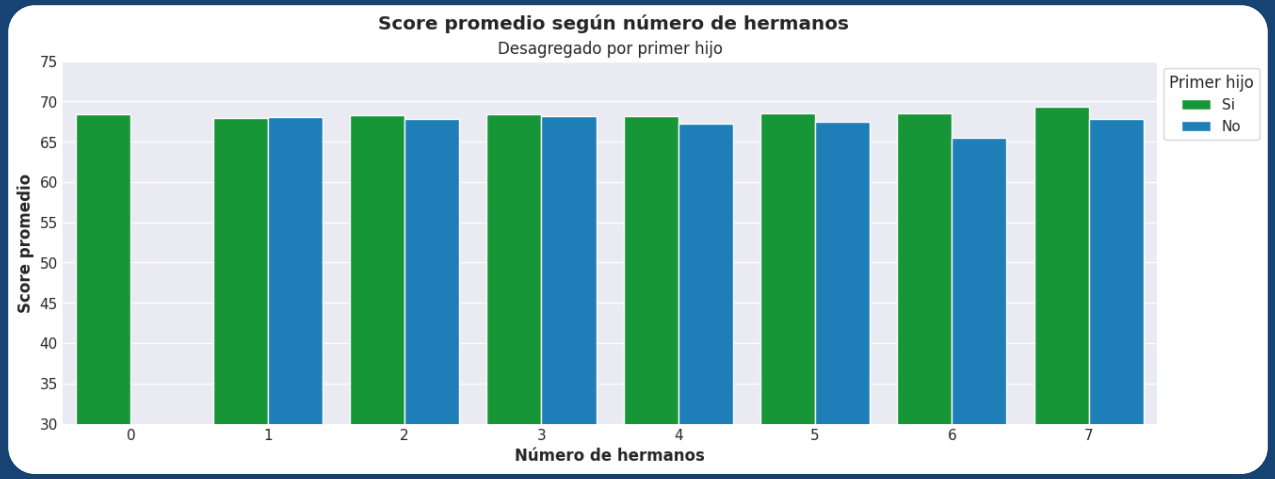
ANÁLISIS UNIVARIADO Y BIVARIADO

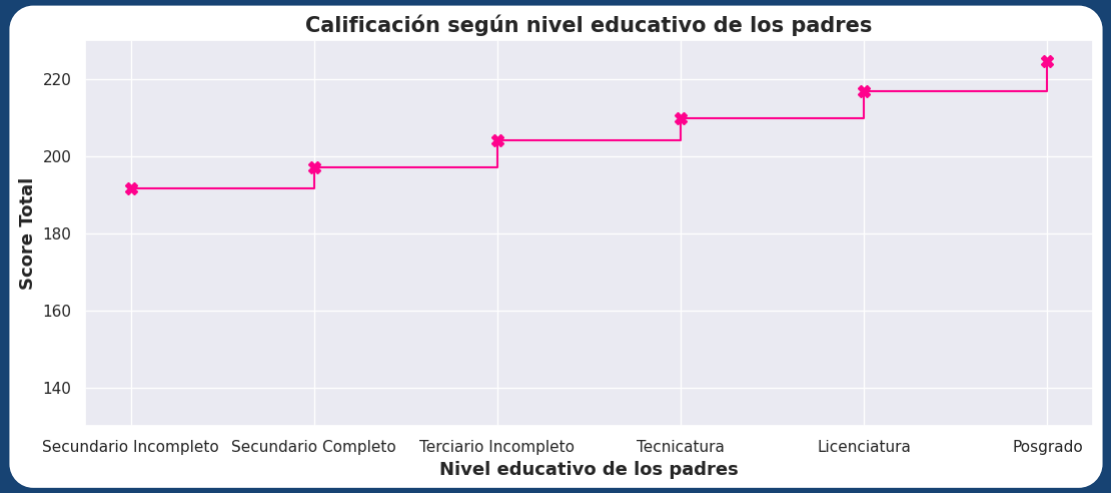




Se puede ver que las mujeres tienen en promedio menos problemas con la exoneración, exceptuando en matemáticas, donde los varones tienen un desempeño relativamente mejor. Esto es algo que se repite en general también entre las étnias, las muejeres se desempeñan en general mejor que los varones.

ANÁLISIS UNIVARIADO Y BIVARIADO

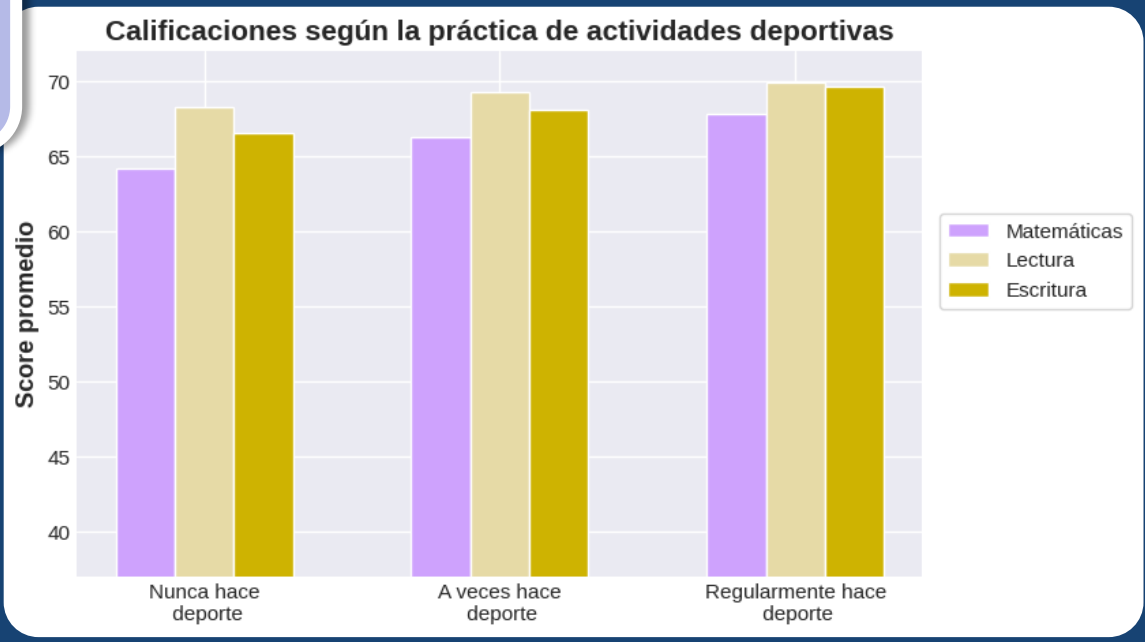


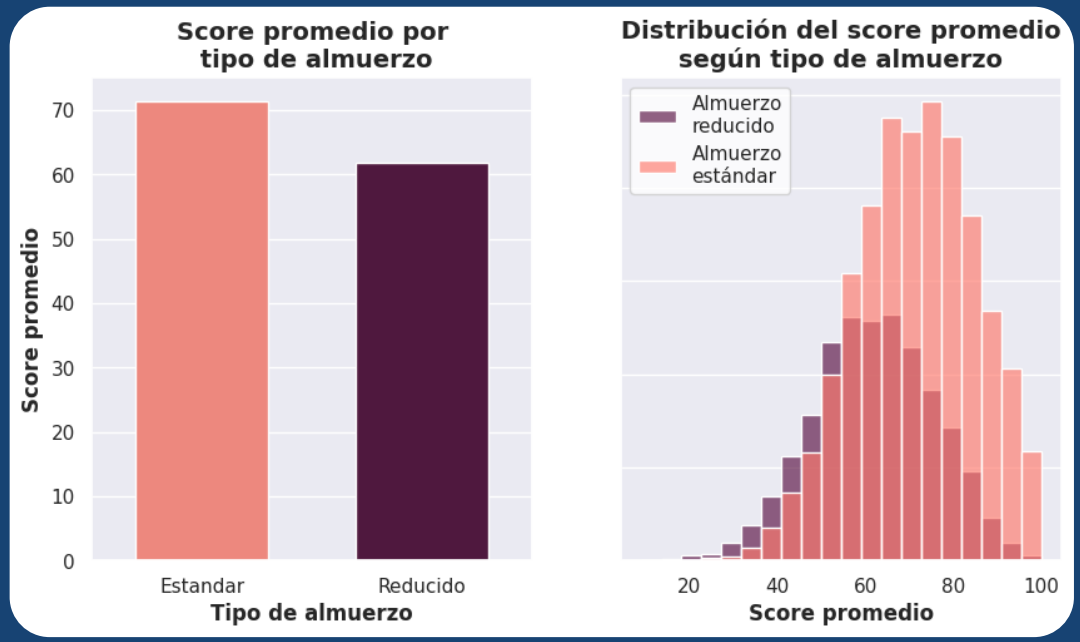


Yendo a la situación familiar no se aprecian mayores diferencias entre el número de hermnanos para el primer hijo, pero si hay una tendencia marginal a la baja en la calificación a medida que el número de hermanos aumenta y no se es el primer hijo.

Se puede concluir también que los alumnos cuyos padres tienen un nivel educativo más elevado tienden a tener mejores resultados educativos.

ANÁLISIS UNIVARIADO Y BIVARIADO

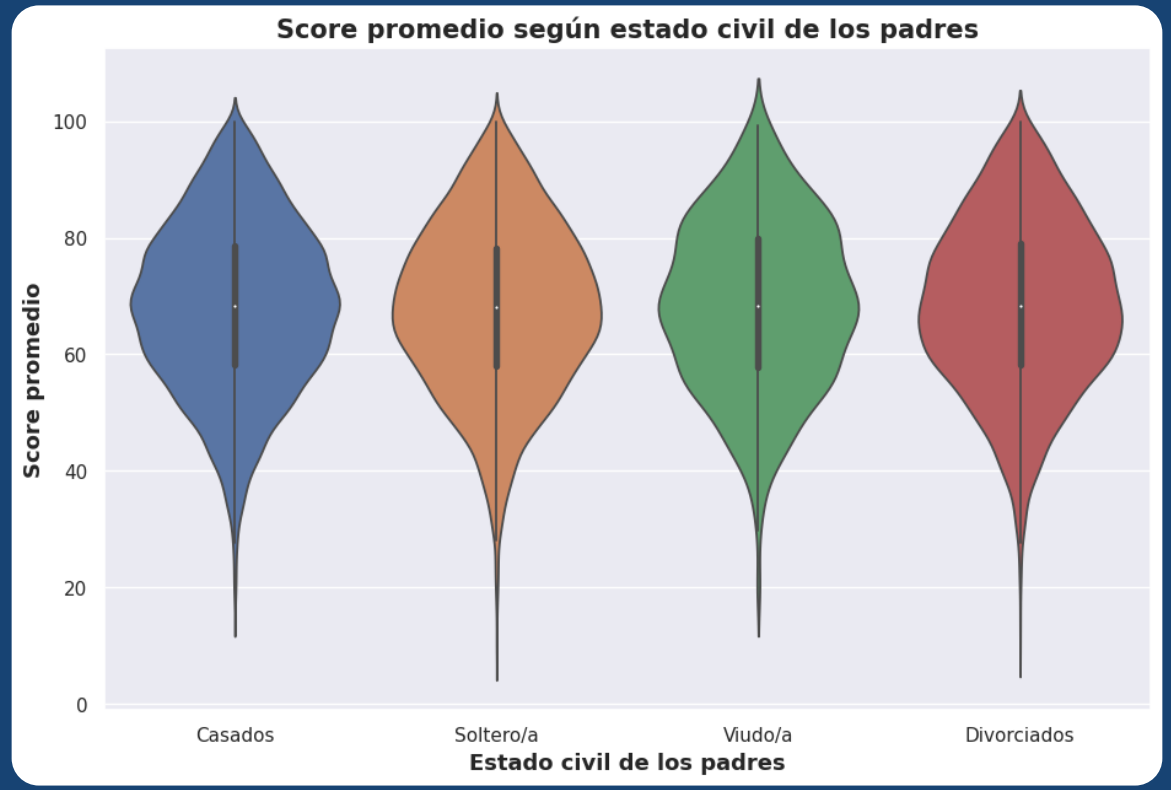




Otra variable a considerar pueden ser las actividades deportivas, donde se desprende de los datos que los alumnos que realizan actividades deportivas de forma más regular obtienen mejores resultados.

En la segunda gráfica se distingue el promedio entre los alumnos que consumen el almuerzo estándar y el almuerzo reducido. Podemos concluir que una peor alimentación diaria podría tener un efecto sobre las calificaciones obtenidas por los alumnos. Esta variable es un proxy al nivel socioeconómico del alumno.

ANÁLISIS UNIVARIADO Y BIVARIADO



No hay mayores diferencias en las calificaciones según el estado civil de los padres, probablemente no sea una buena variable para realizar predicciones.

Como se puede comprobar hay variables no directamente relacionadas al estudio que parecen afectar el desempeño educativo, por lo tanto es importante tener un modelo que permita identificar cuáles son exactamente esas variables y cuánto afectan a los resultados educativos, para poder elaborar políticas que ataquen dichos problemas.

SELECCIÓN DE VARIABLES

Luego de limpiar las variables de outliers y valores nulos se realiza un proceso de backward elimination para determinar, dentro de un modelo explicando la exoneración del curso, cuáles variables no serían relevantes.

El resultado es que las variables que indican el estado civil de los padres, el medio de transporte y la cantidad de hermanos van a ser descartadas para la realización de los modelos.

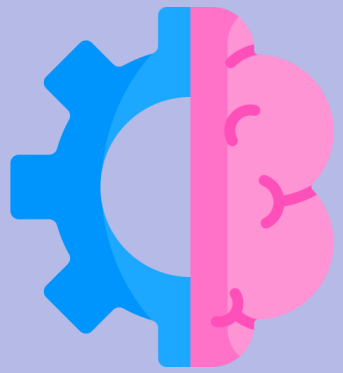
```
scoreX = score_by.drop(columns = 'ex_curso')
X = scoreX

scoreY = score_by.ex_curso
Y = scoreY
```

```
def backward_elimination(X, Y,significance_level = 0.05):
    features = X.columns.tolist()
    while(len(features)>0):
        features_with_constant = smod.add_constant(X[features])
        p_values = smod.OLS(Y, features_with_constant).fit().pvalues[1:]
        max_p_value = p_values.max()
        if(max_p_value >= significance_level):
            excluded_feature = p_values.idxmax()
            features.remove(excluded_feature)
        else:
            break
    return features

backward_elimination(X,Y)
```

5.



MODELOS DE MACHINE LEARNING

SUMMARY

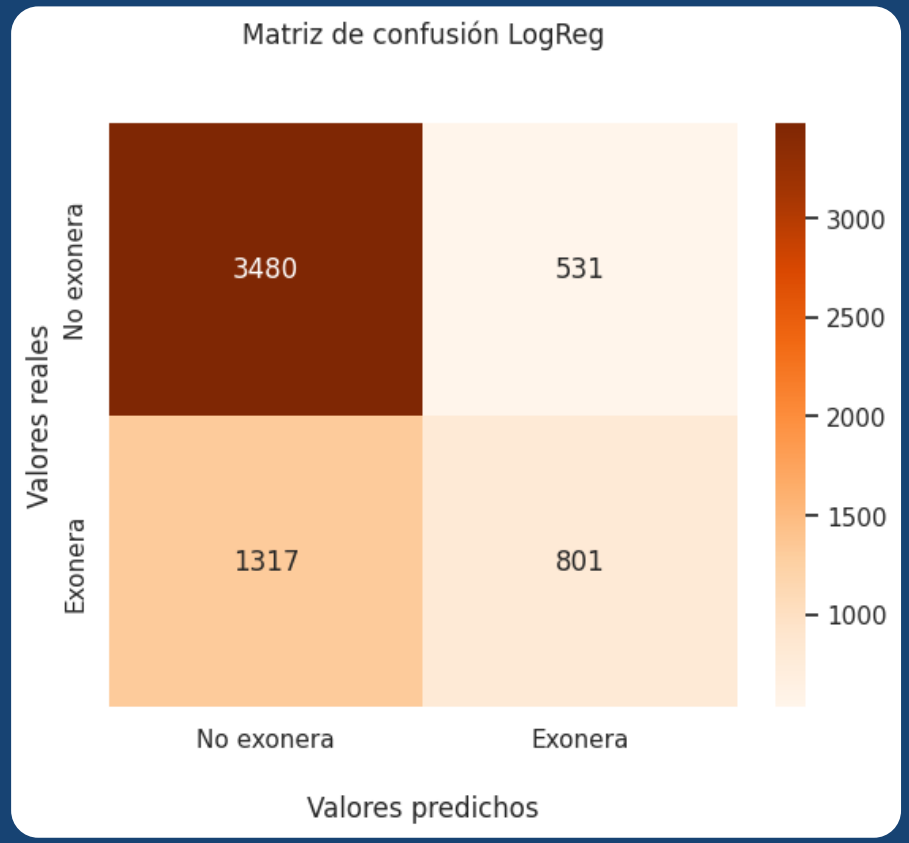
Métricas				
Modelo	Accuarcy	Precisión	Recall	F1 score
LogReg	0,70	0,73	0,87	0,79
XGB	0,69	0,73	0,83	0,78

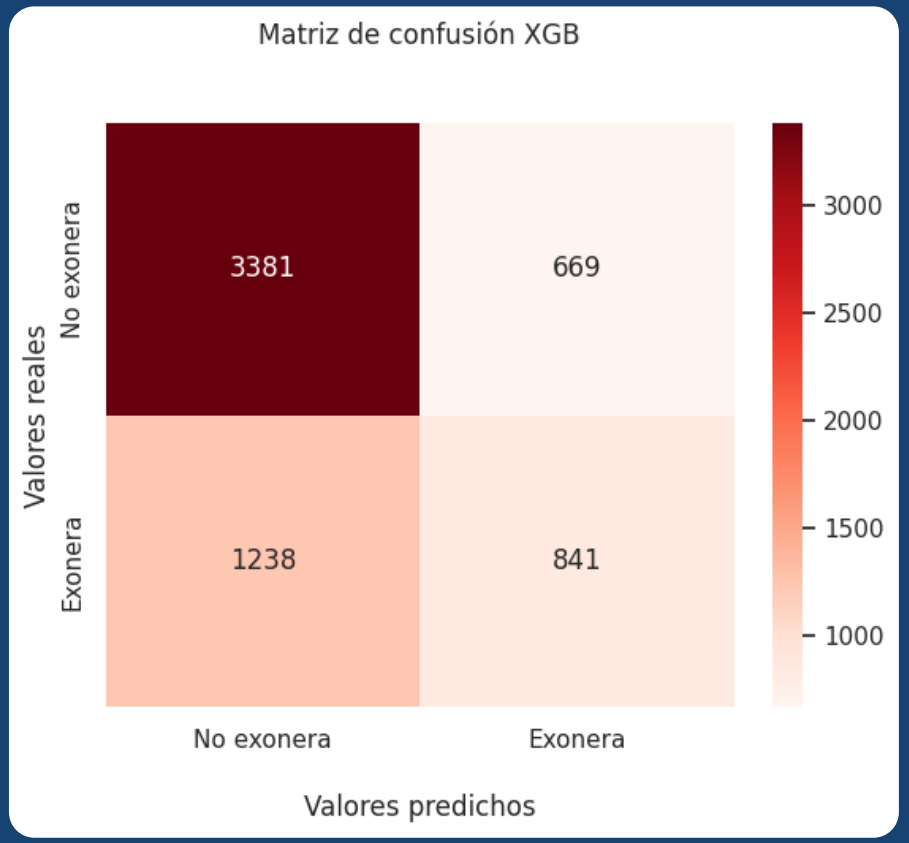
Se realizaron dos modelos de Machine Learning, un modelo de regresión logística (LogReg) y un modelo XGBoost (XGB), para ambos se dividieron los datos en un 80% para entrenamiento y 20% para testing.

Las métricas (Precisión, Recall y F1 score) se centran en los resultados para la variable ex_curso = 0, ya que lo más importante es poder identificar correctamente los alumnos que pueden llegar a tener problemas para la exoneración. Para los objetivos del trabajo es un peor error el clasificar a un alumno que no va a exonerar como uno que si lo haría, que el caso contrario.

Aunque el modelo de regresión logística es ligeramente mejor que el XGBoost, la diferencia no es suficiente como para descartar este último, por lo que se hará hypertuning con ambos.

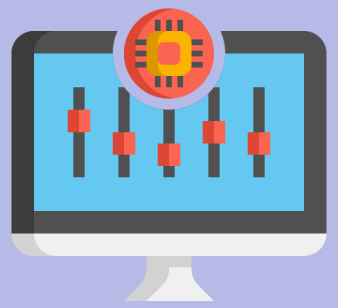
SUMMARY





Se presentan las matrices de confusión para ambos modelos como complemento de la tabla resumen de métricas.

6.



HYPERTUNING

HYPERTUNING DE PARÁMETROS

```
params_grid = {
        'solver': ['lbfgs', 'liblinear', 'newton-cg', 'newton-cholesky', 'sag', 'saga'],
        'class_weight': [{0:2.0, 1:1.0}, {0:1.2, 1:1.0}, 'balanced'],
        'penalty': ['l1', 'l2', 'elasticnet', None],
        'max_iter': [100, 150, 200]
        }
```

Métricas				
Modelo	Accuarcy	Precisión	Recall	F1 score
LogReg	0,70	0,73	0,87	0,79
LogReg Tuned	0,70	0,71	0,91	0,80

Se realiza hypertuning de parámetros para ambos modelos por el método de Halving Grid Search, y se comparan los resultados con las mismas métricas que se presentaron con los modelos iniciales.

Para el modelo LogReg se ve una mejora de 0,04 en el recall, 0,01 en el F1 score, y una disminución de 0,02 en la precisión, manteniendo el accuarcy al mismo nivel.

HYPERTUNING DE PARÁMETROS

```
params_grid_xgb = {
        'min_child_weight': [1, 5, 10],
        'gamma': [0.5, 1, 1.5, 2, 5],
        'subsample': [0.6, 0.8, 1.0],
        'colsample_bytree': [0.6, 0.8, 1.0],
        'max_depth': [3, 5, 6, 7],
        }
```

Métricas				
Modelo	Accuarcy	Precisión	Recall	F1 score
XGB	0,69	0,73	0,83	0,78
XGB Tuned	0,71	0,73	0,88	0,80

Para el modelo XGB se ve un aumento de 0,02 en el accuarcy, la precisión se mantiene, aumenta el recall en 0,05 y el F1 score en 0,02.

7.

CONCLUSIÓN

CONCLUSIONES

El dataset posee variables extra curriculares que pueden ser parcialmente determinantes de la no exoneración del curso. La variable objetivo (ex_curso) presenta una distribución desbalanceada, donde el 66% de los individuos de la muestra no exoneran el curso.

Mediante el EDA y feature selection se eligen las variables que se van a utilizar para los modelos, con el objetivo de determinar si efectivamente esas variables extra curriculares pueden ayudar a explicar la variable objetivo.

Luego de realizar un modelo de regresión logística y un XGBoost, se llega a la conclusión de que el segundo es el modelo preferido, con un accuarcy de 71%, una precisión del 73%, un recall de 88% y un F1 score de 80%.