DATA SCIENCE - SEGUNDA PREENTREGA

CODER HOUSE

ANÁLISIS DE DATOS EDUCATIVOS DE APROBACIÓN

NICOLÁS CLADERA

ABSTRACT - CONTEXTO COMERCIAL

CODER HOUSE

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

CODER HOUSE

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?

RESUMEN DE METADATA

CODER HOUSE

Fuente de la base de datos: Kaggle

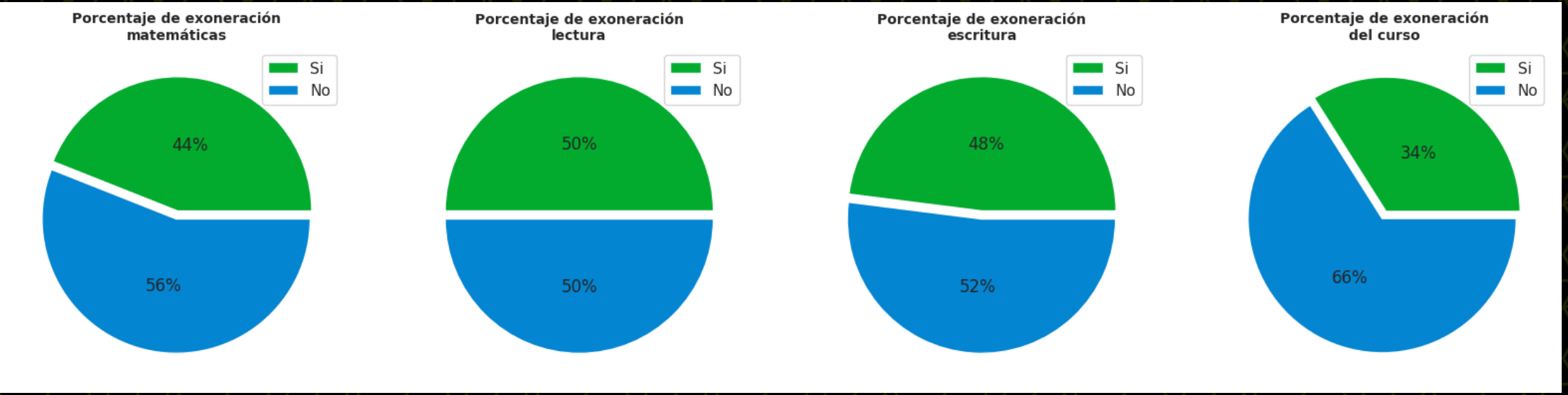
Número de observaciones: 30.000

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.

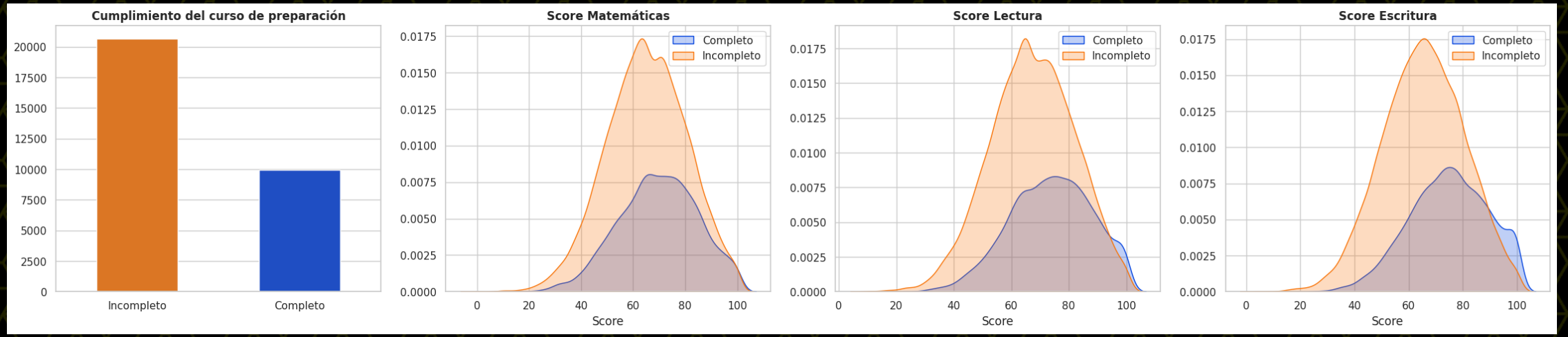
VISUALIZACIONES E INSIGHTS

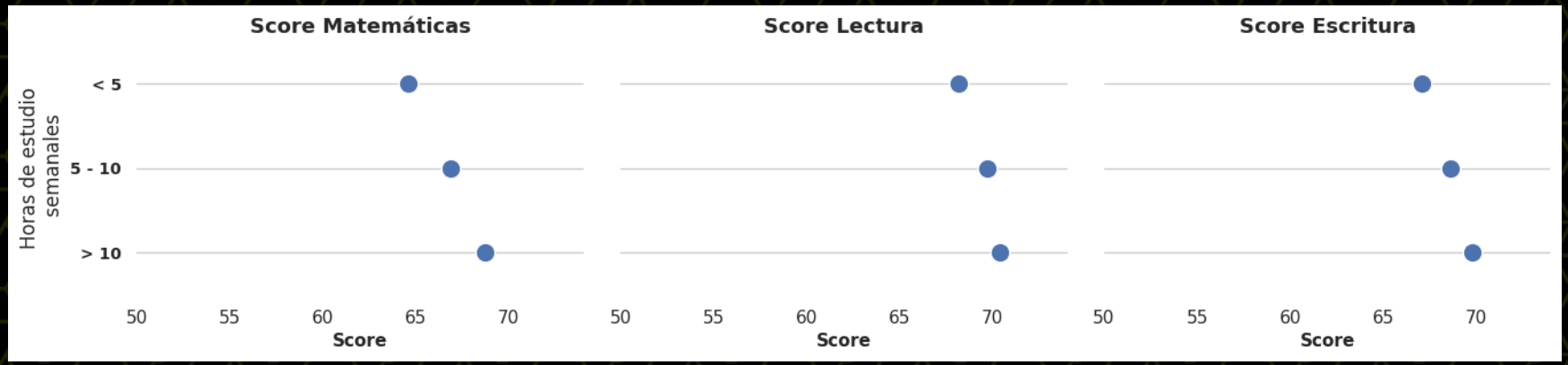


Los alumnos no consiguen más de un 50% de exoneración en ninguna de las asignaturas, siendo el peor matemáticas, con un 44% de exoneración.

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%.

VISUALIZACIONES E INSIGHTS

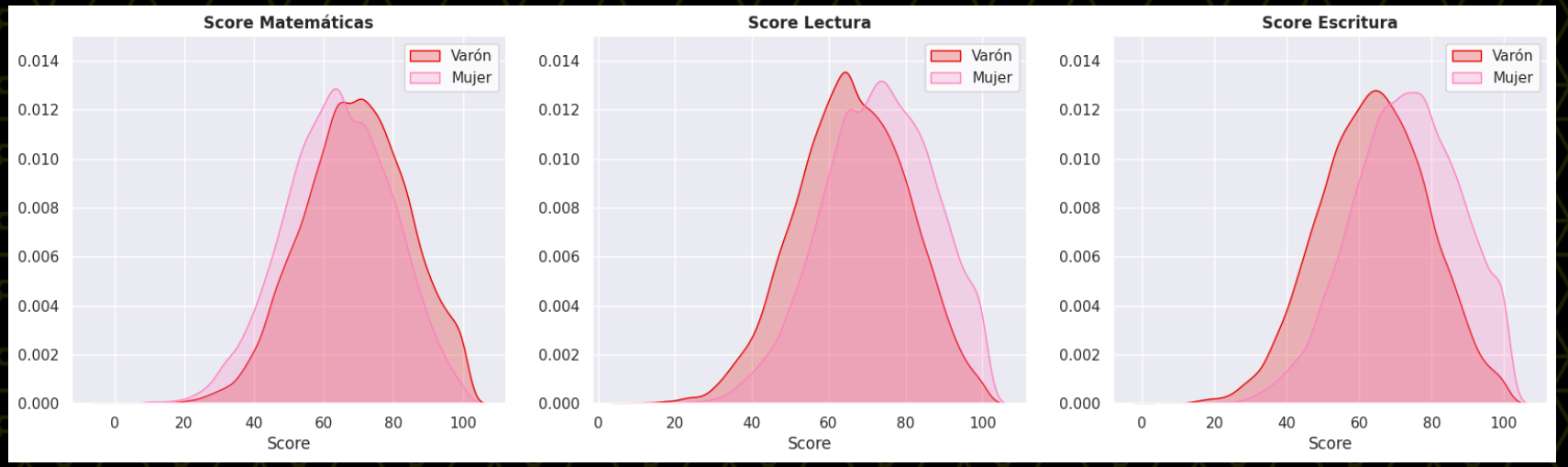


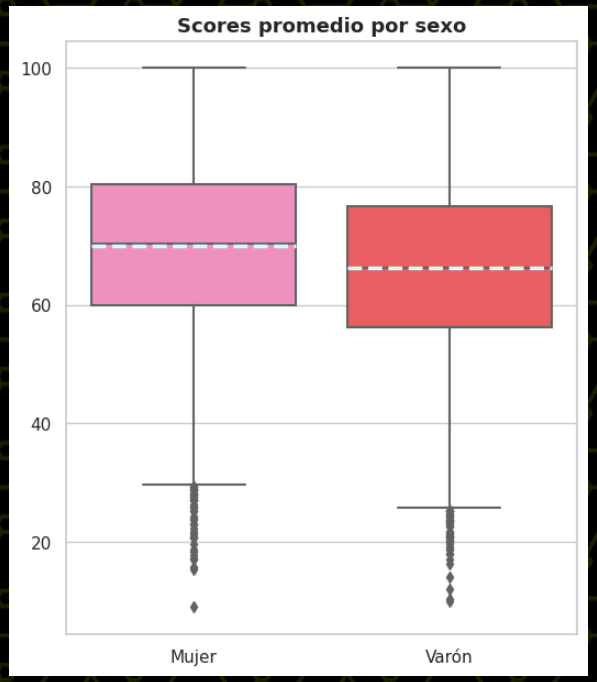


Como se mencionó anteriormente la idea es agregar variables que no estén relacionadas con el estudio al modelo, pero no se puede dejar afuera las mismas, ya que hay una clara tendencia a una mejora en las calificaciones cuando se estudia una mayor cantidad de horas.

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

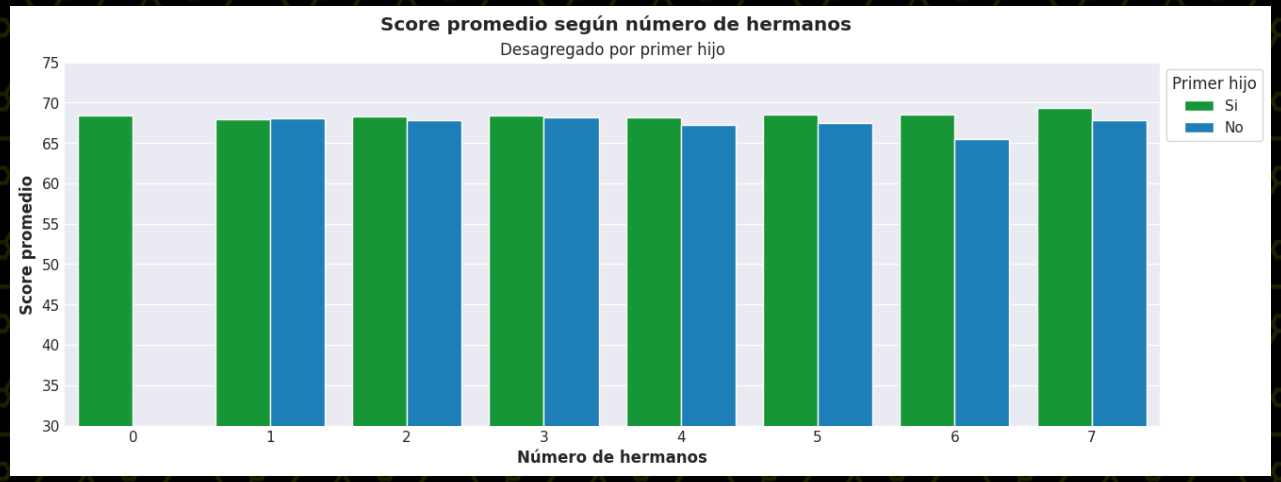
VISUALIZACIONES E INSIGHTS

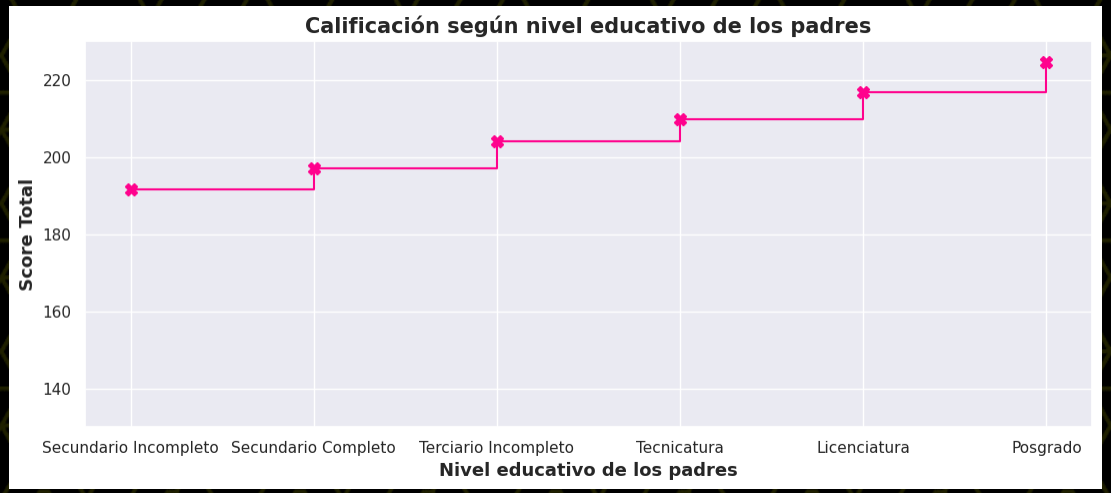




Como una primer variable a estudiar se puede ver que las mujeres tienen en promedio menos problemas con la exoneración, exceptuando en matemáticas, donde los varones tienen un desempeño relativamente mejor. Esto es algo que se repite en general también entre las étnias, las muejeres se desempeñan en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

VISUALIZACIONES E INSIGHTS

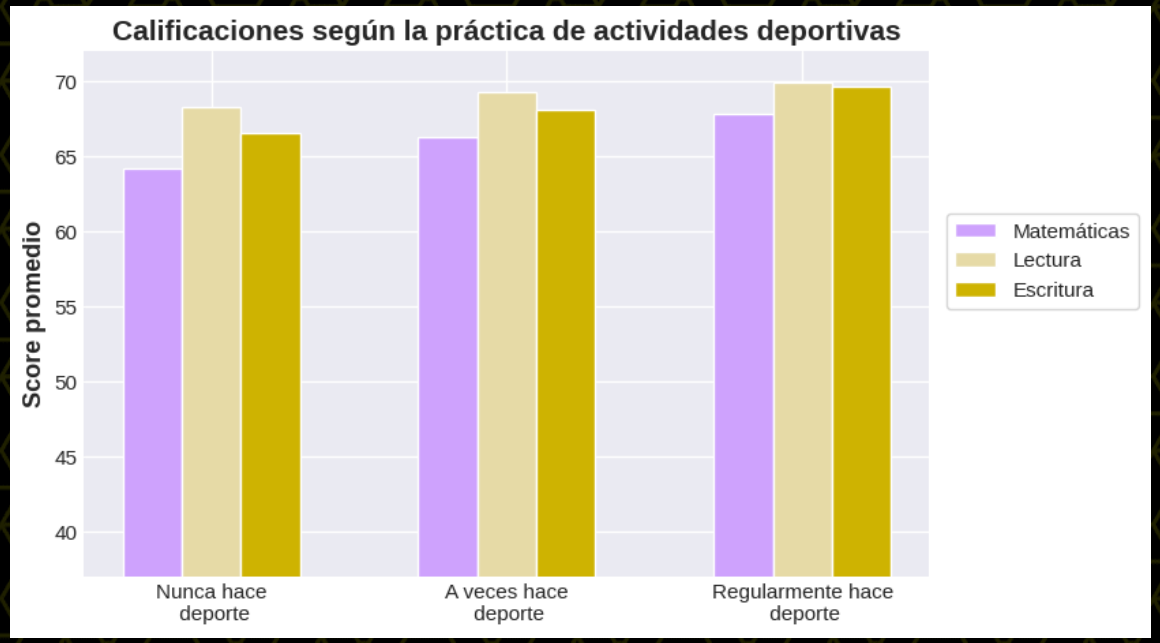


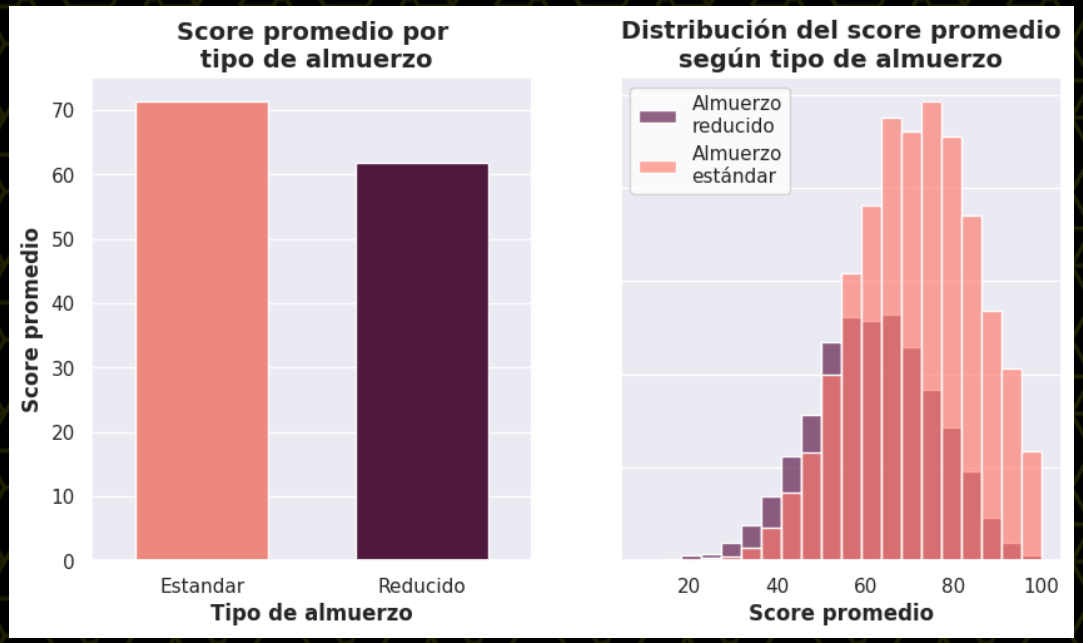


Yendo a la situación familiar no se aprecian mayores diferencias entre el número de hermnanos para el primer hijo, pero si hay una tendencia marginal a la baja en la calificación a medida que el número de hermanos aumenta y no se es el primer hijo.

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

VISUALIZACIONES E INSIGHTS

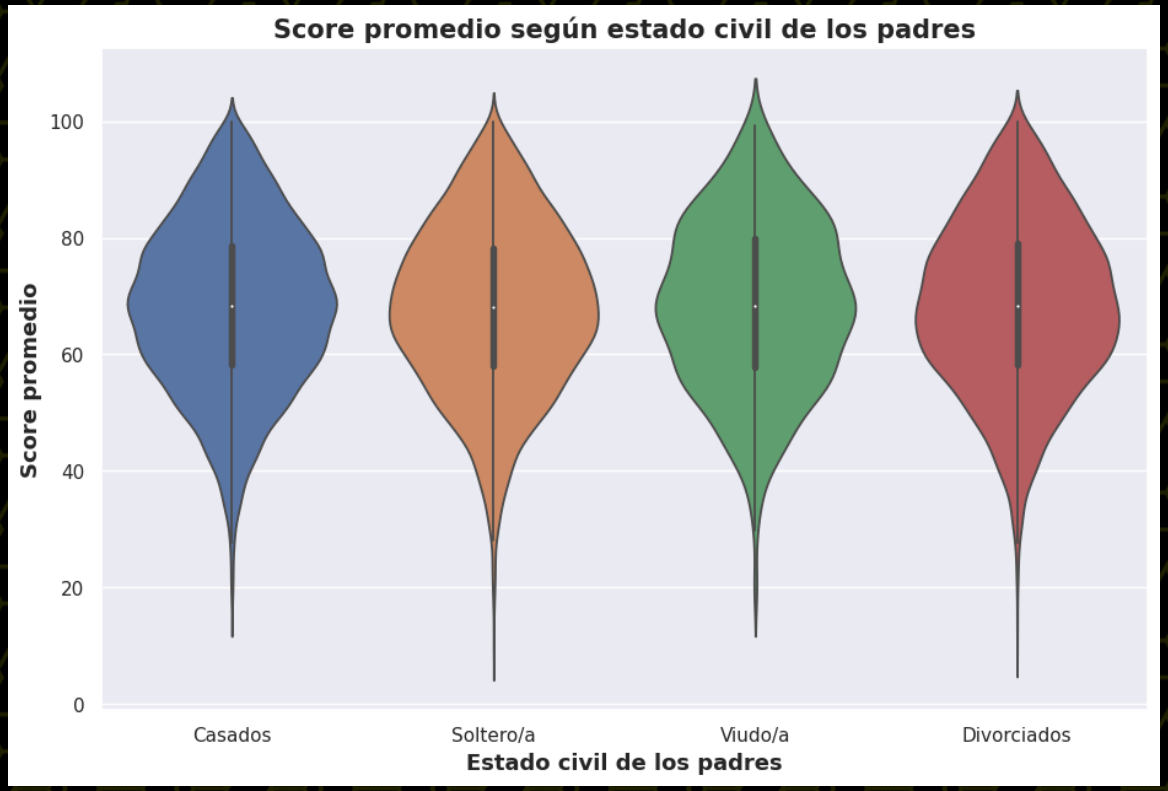




Otra variable a considerar pueden ser las actividades deportivas, donde se desprende de los datos que los alumnos que realizan actividades deportivas de forma más regular obtienen mejores resultados.

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.

VISUALIZACIONES E INSIGHTS



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.