Análisis Rendimiento Pruebas Saber Pro

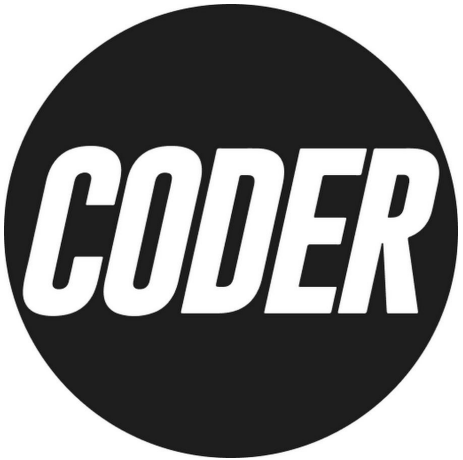
Proyecto Data Science II: Machine Learning para la Ciencia de Datos

Contenido

Examen Saber Pro

CODER

Contenido



Abstract y Audiencia

Para este proyecto del curso Data Science II: Machine Learning para la Ciencia de Datos, se hace uso de un set de datos obtenido de una competencia de Kaggle en la cual se espera generar un modelo predictivo usando datos recopilados de las pruebas saber pro (examen realizado a los estudiantes de educación superior en Colombia para medir el nivel educativo).

Mediante el uso de un set de datos de entrenamiento en el cual se cuenta con una columna que permite conocer el desempeño del estudiante, mediante una clasificación de 4 clases (baja, media baja, media alta y alta), se espera generar un modelo que al ser testeado con un dataset sin esta columna, permita una predicción lo más acertada posible del resultado del estudiante basándose en los factores presentes en las demás columnas (estrato socioeconómico, situación familiar, nivel educativo padres, situación laboral del estudiante, etc).

Metadata

16

Campos

628.896

Registros

Metadata

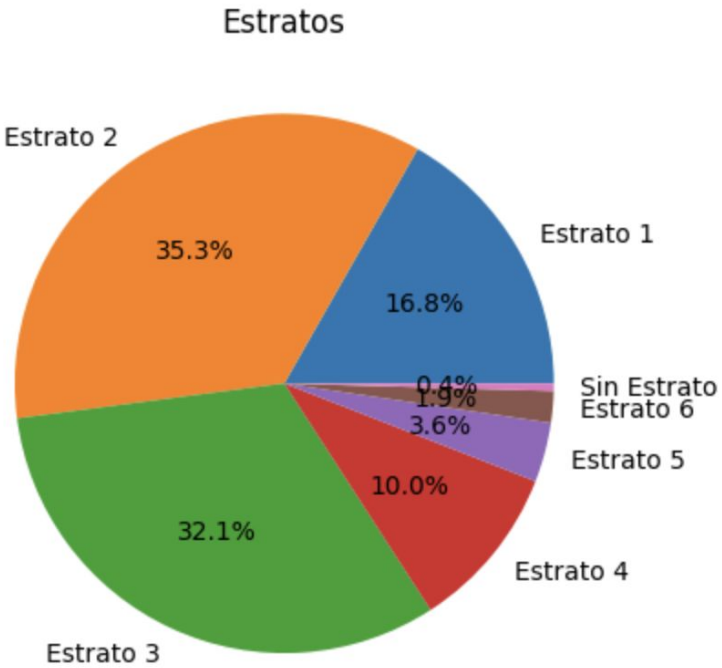
NOMBRE CAMPO	INFORMACIÓN	TIPO VARIABLE
ID	ID ESTUDIANTE	INTEGER
PERIODO	PERIODO EXAMEN	INTEGER
ESTU_PRGM_ACADEMICO	PROGRAMA ACADÉMICO	STRING
ESTU_PRGM_DEPARTAMENTO	DEPARTAMENTO	STRING
ESTU_VALORMATRICULAUNIVERSIDAD	VALOR MATRÍCULA	STRING
ESTU_HORASSEMANATRABAJA	HORAS SEMANALES TRABAJADAS	STRING
FAMI_ESTRATOVIVIENDA	ESTRATO SOCIOECONÓMICO	STRING
FAMI_EDUCACIONPADRE	NIVEL EDUCATIVO PADRE	STRING
FAMI_EDUCACIONMADRE	NIVEL EDUCATIVO MADRE	STRING
ESTU_PAGOMATRICULAPROPIO	¿PAGA MATRÍCULA PROPIA?	STRING
ESTU_PRIVADO_LIBERTAD	¿PRIVADO DE LA LIBERTAD?	STRING
FAMI_TIENEINTERNET	¿TIENE INTERNET?	STRING
FAMI_TIENECOMPUTADOR	¿TIENE COMPUTADOR?	STRING
FAMI_TIENELAVADORA	¿TIENE LAVADORA?	STRING
FAMI_TIENEAUTOMOVIL	¿TIENE AUTOMÓVIL?	STRING
RENDIMIENTO_GLOBAL	RENDIMIENTO	STRING

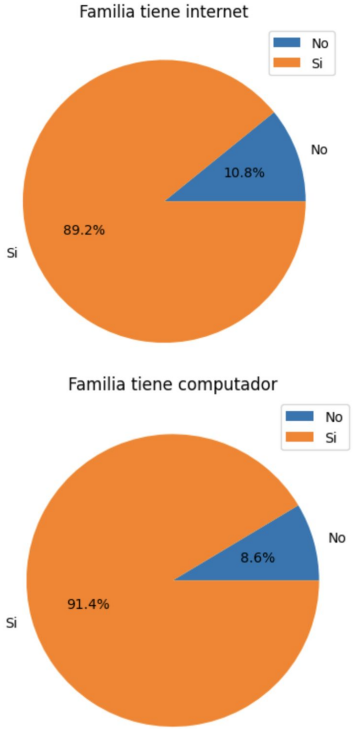
Pregunta e Hipótesis

¿Que factores socio-económicos están afectando el resultado en las pruebas de estado de los estudiantes colombianos de educación superior?

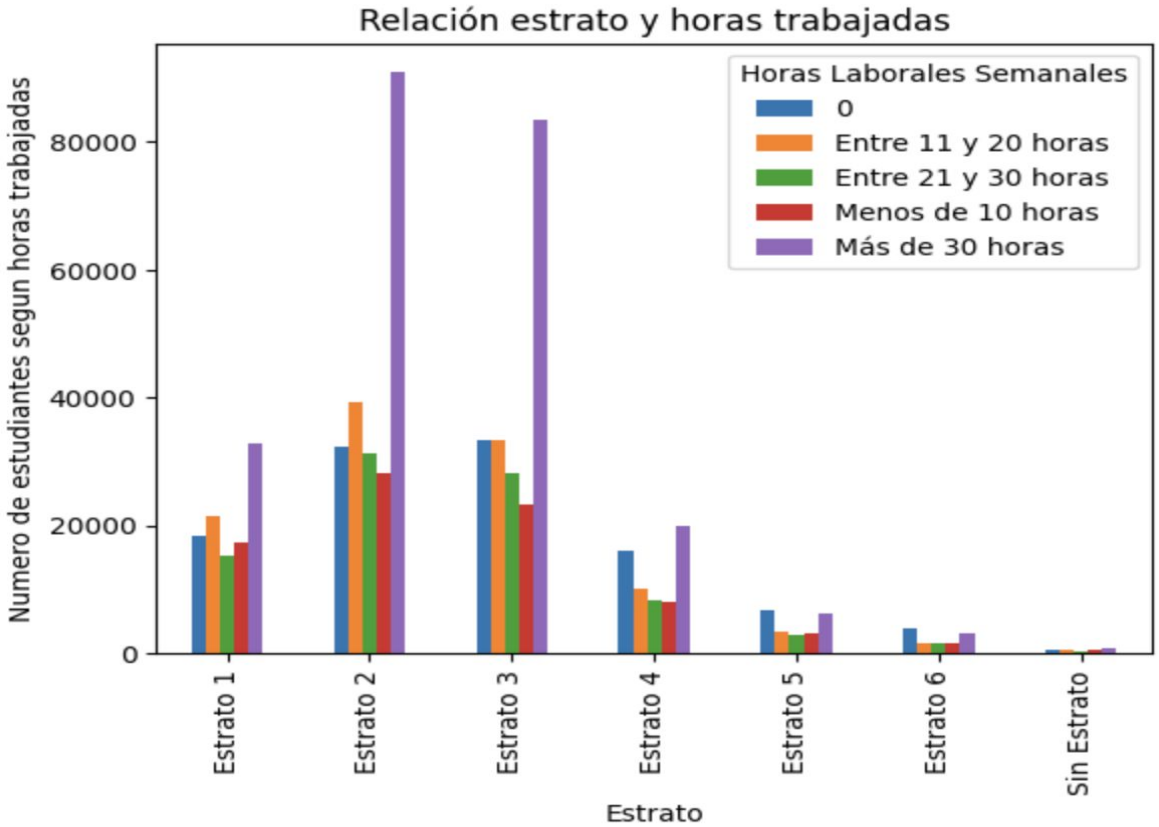
- Distribución de los estratos sociales en los estudiantes que presentan la prueba
- Posible relación entre factores socioeconómicos presentes en el dataset
- ¿Existe una influencia de factores específicos como el periodo del examen, estrato o tenencia de bienes materiales, en el rendimiento general de los estudiantes?

Análisis e Insights

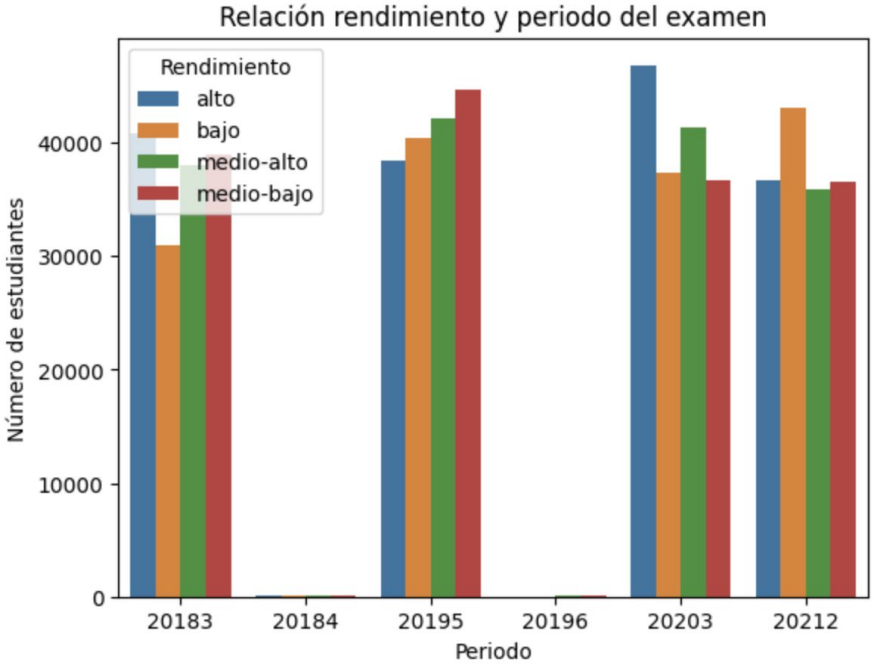


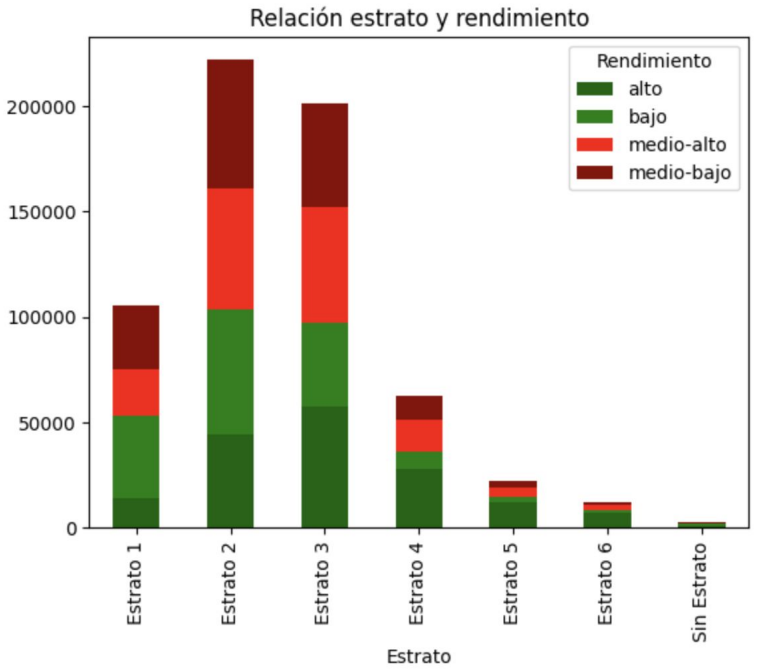


- Distribución estratos socioeconómicos (predominantemente estratos 2 y 3)
- Relación clara entre tenencia de computador e internet por parte de los evaluados



Predominancia de 0 horas en los estratos más altos (5 y 6) y de más de 30 horas laborales semanales en los demás estratos

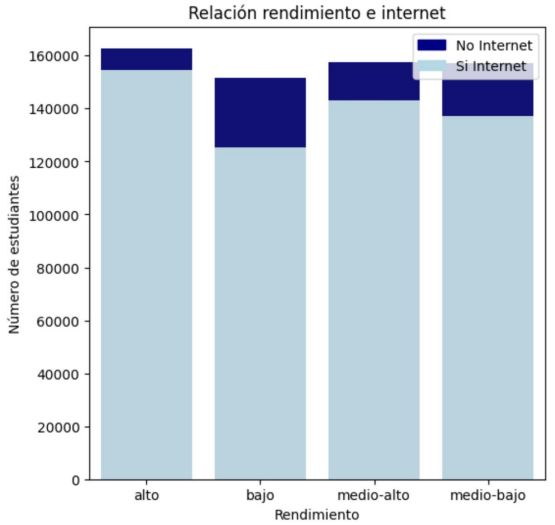


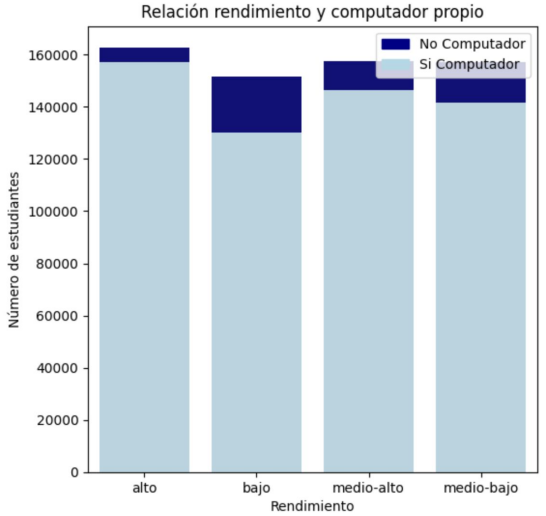


- No existe tendencia clara entre los diferentes periodos presentes en el dataset
- Relación directamente proporcional entre el estrato y el nivel de rendimiento en la prueba

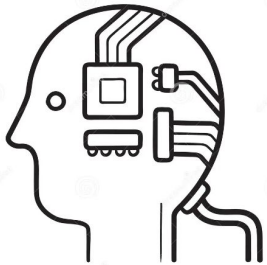
- Comportamiento directamente relacionado entre el rendimiento y la existencia de una conexión a internet y computador propio

- La mayoría de los estudiantes que presentan la prueba obtienen un desempeño alto (aunque las cantidades son cercanas entre las 4 clasificaciones)





Preprocesamiento

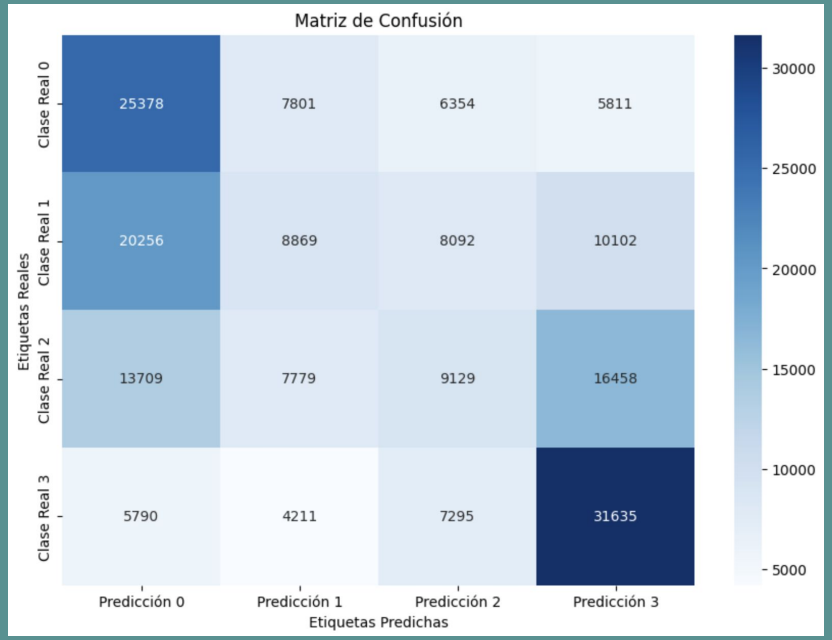


El principal reto se debe a la falta de variables cuantitativas, todas las variables de interés son cualitativas y requieren procesamiento

- Variables SI / NO → One Hot Encoding (OHE)
- Variables Categóricas Ordinales → Ordinal Encoder
- Variables Categóricas Nominales Complejas → Ordinal Encoder (tras análisis de impacto)

Modelos

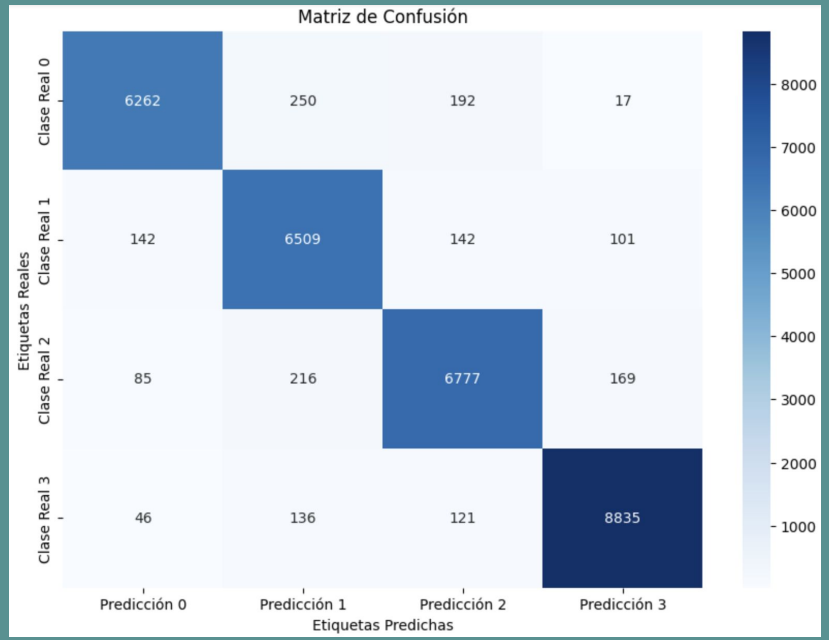
Regresión Logística



39.75%

Accuracy

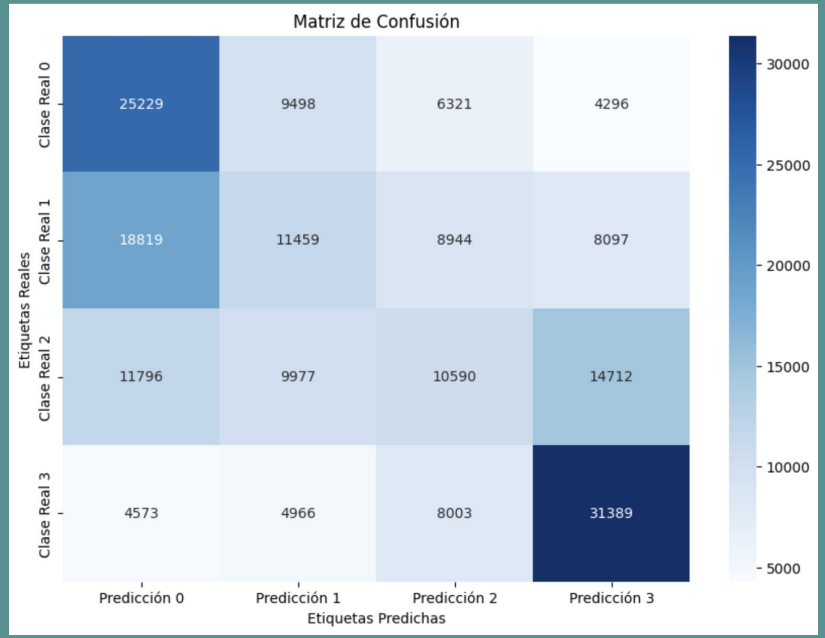
SVM



94.61%

Accuracy

SVM



41.69%

Accuracy

Conclusiones y Recomendaciones

- Se pueden evidenciar resultados similares en la regresión y el random forest, con un accuracy mayor para el segundo, permitiendo concluir que cuenta con mayor capacidad predictiva. Por otro lado el SVM se evidencia sobreentrenado, pues el resultado es demasiado alto para una clasificación de 4 categorías.

- Podemos apreciar que para el dataset con el feature selection (con los 5 features escogidos por el RFE) tenemos una precisión del 37.82% y para el dataset completo una precisión del 39.76%, lo cual permite concluir que aunque estamos reduciendo la dimensión del dataset a menos de la mitad de columnas originales, obtenemos un resultado similar al del set completo.

- En las matrices de confusión apreciamos que tienen comportamientos parecidos, con una muy buena detección de las clase en los extremos (0='bajo' y 3='alto') y un desempeño no tan bueno con las clases intermedias. También apreciamos que la principal diferencia entre ambos modelos se presenta en la detección de bajos reales.

Conclusiones y Recomendaciones

- **Aunque el modelo puede parecer insuficiente a primera vista debido a un accuracy, recall y precision inferiores a un 50%, hay que tener en cuenta que en este caso tenemos 4 clasificaciones posibles, lo cual deja como límite inferior un 25% de accuracy, que sería la probabilidad de adivinar sin ningún análisis de por medio.**

- A futuro es necesario evaluar un modelo de predicción de mayor complejidad, para lo cual sería necesario tal vez escalar el modelo para reducir el número de datos en el entrenamiento. Si esto no diera mejores resultados se debe reevaluar el encoding generado.