

Trabajo Final Integrador para Licenciatura en Sistemas

Modelo de minería de datos aplicado a la detección de propensión a desaprobación en examen aprender en el área de matemáticas

Estudiante:

Tutores:

Violi Pablo Ezequiel

Dr. Pytel Pablo

Lic. Loidi Laura Gabriela

Presentación



Violi Pablo Ezequiel

27 años

Estudiante UNLa

Cohorte 2012

Analista de datos / Data Scientist



Introducción



Objetivo del trabajo:

Generar un modelo de propensión que permita detectar aquellos estudiantes con mayor riesgo de desaprobar el examen de matemáticas de las evaluaciones Aprender, emitidas por el Ministerio de Educación

Fuentes utilizadas:

- Bases de encuesta Aprender 2019
- Bases de Evaluación Matemáticas 2019

Temporalidad:

Trabajo comenzado durante 2020, y finalizado en Diciembre 2021



Marco Teórico: Encuestas Aprender



¿Cuál es el objetivo?

Aprender busca producir evidencia de carácter diagnóstico para el análisis, la reflexión y la toma de decisiones orientadas a garantizar el derecho a la educación.



¿Qué información releva?

Permite conocer el grado de dominio que las y los estudiantes de nivel primario y secundario tienen sobre un recorte específico de contenidos y capacidades cognitivas durante su trayectoria escolar e identificar los factores sociodemográficos y las condiciones en que se enseña y se aprende.



¿Cómo se construye la prueba?

La evaluación es desarrollada por el Ministerio de Educación, a través de la Secretaría de Evaluación e Información Educativa, en acuerdo con el Consejo Federal de Educación y los equipos técnicos jurisdiccionales de evaluación educativa y con la validación técnica realizada por un equipo de lectores críticos e itemistas especializados en las áreas a evaluar.

Fuente: Ministerio de Educación



Marco Teórico: Bases de Datos



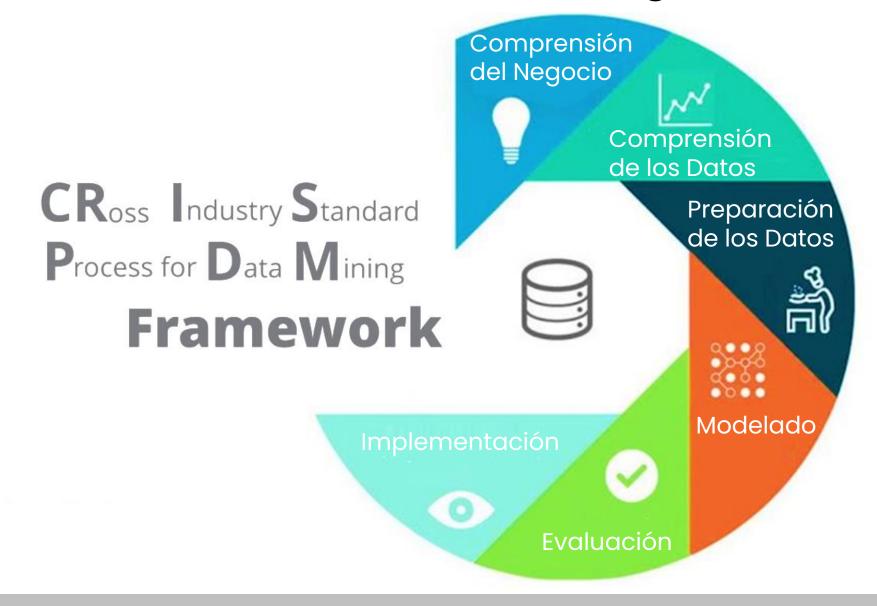


Marco Teórico: Minería de datos





Marco Teórico: Metodología CRISP-DM





Herramientas

Para llevar a cabo el proyecto, se trabaja:

- Sobre la interfaz web Jupyter Lab que permite programar en lenguaje Python 2.7.
- Usando las librerías Numpy, Pandas, ScikitLearn y Matplotlib entre otras

















Departamento de Desarrollo Productivo y Tecnológico



1. Comprensión del Negocio





- ¿Importa el acceso a internet para la performance del estudiante?
- ¿El resultado del examen varía según los ingresos económicos del hogar?
- ¿Importa el nivel de educación de los padres para que el estudiante incorpore correctamente los conceptos de Matemática?

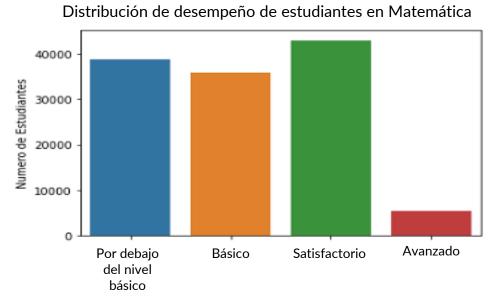
Objetivo de la minería de datos:

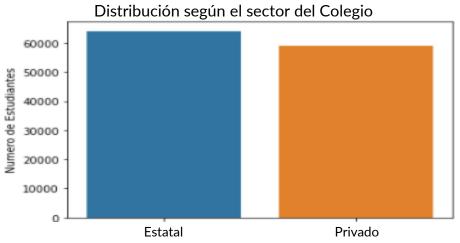
Obtener un modelo que tenga un Accuracy de al menos **0,70** y un área bajo la curva ROC mayor a **0,75**.

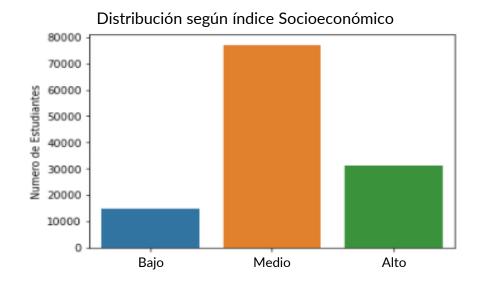


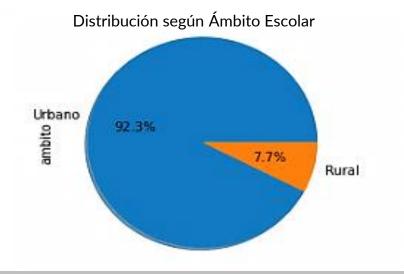


2. Comprensión de los Datos













3. Preparación de los Datos

Durante esta etapa, entre otras tareas, se realizan las siguientes:

- > Se toman los 65 atributos con mayor correlación a la nota del estudiante
- > Se arma una variable binaria en base a la nota del alumno:

Valor Final	Valor Origen	
1	Por debajo del nivel básico	
0	Básico	
	Satisfactorio	
	Avanzado	

> Se separan los datos en dos datasets, uno para el entrenamiento, y otro para validación:

Dataset de Entrenamiento			
Valor de atributo Target	Cantidad		
0	29.086		
1	29.086		

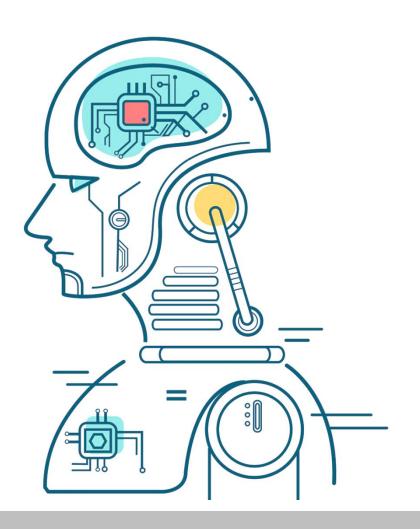
Dataset de Validacion				
Valor de atributo Target	Cantidad			
0	21.122			
1	9.633			





4. Modelado

Algoritmos utilizados durante esta etapa:



- o Regresiones Logísticas
- o Naive Bayes
- o Clasificador KNN
- o Arboles de Decision
- o Random Forest
- o LightGBM
- o XGBoosting





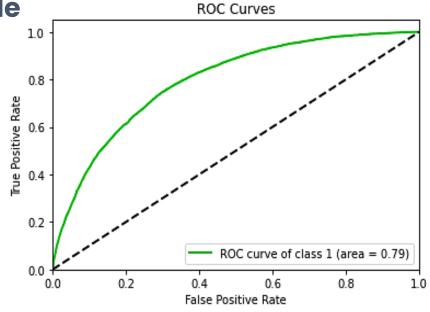
5. Evaluación

El mejor resultado fue obtenido utilizando el algoritmo <u>LightGBM</u>

En los datos de validación se genera un Accuracy de 0,714, y un área bajo la curva ROC de 0,79

Objetivo de la minería de datos:

- ✓ Accuracy mayor a 0,70
- ✓ Área bajo la curva ROC mayor a 0,75



Curva ROC - Modelo LightGBM

Clase	precisión	recall	f1-score
0	0,857	0,699	0,770
1	0,530	0,744	0,619
Promedio	0,693	0,722	0,694

Accuracy:	0.714
-----------	-------

Metricas de Performance - Modelo LightGBM



5. Evaluación

Objetivo del negocio:

¿Importa el acceso a internet para la performance del estudiante?

Si bien el atributo tiene una correlación marcada con el modelo, no es de los atributos más críticos al momento de determinar el éxito del estudiante en el examen

¿El resultado del examen varía según los ingresos económicos del hogar?

Efectivamente, es una de las variables con mayor importancia con respecto a la nota del examen, por lo que es determinante al momento de evaluar al estudiante.

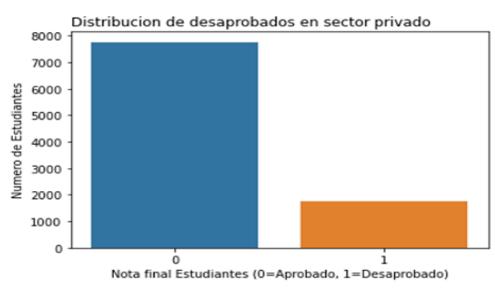
¿Importa el nivel de educación de los padres para que el estudiante incorpore correctamente los conceptos de Matemática?

Los resultados muestran que, si bien el nivel de los padres tiene una incidencia alta sobre el desempeño del estudiante, el nivel de la madre es más importante que el nivel del padre al momento de determinar el éxito del estudiante en el examen.

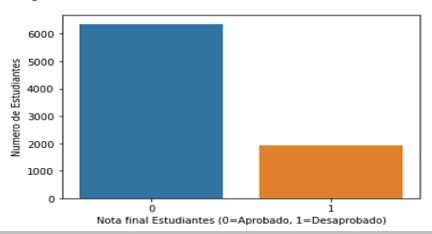


5. Evaluación

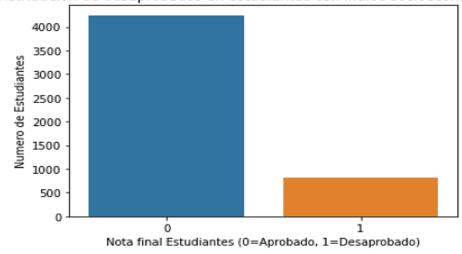
Variables con mayor importancia:



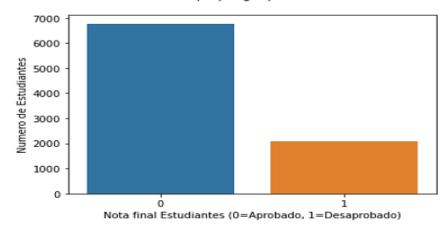
Distribucion de desaprobados en alumnos que respondieron Positivamente a la pregunta: ¿Vas a Seguir estudiando en educación universitaria cuando termines el secundario?



Distribucion de desaprobados en estudiantes con indice socioeconomico Alto



Distribucion de desaprobados en alumnos que respondieron Positivamente en la afirmacion: Si me lo propongo, puedo ser bueno en Matemática







6. Implementación

Utilizando los datos de validación se divide a la población en 10 partes iguales (o deciles) de igual cantidad de estudiantes:

Ranking Riesgo	Estudiantes Totales	Estudiantes Desaprobados	Tasa de Desaprobación	Captura acumulada de desaprobados
1	3.076	2.230	72%	23%
2	3.075	1.840	60%	42%
3	3.076	1.483	48%	58%
4	3.075	1.201	39%	70%
5	3.075	972	32%	80%
6	3.076	721	23%	88%
7	3.075	551	18%	93%
8	3.076	332	11%	97%
9	3.075	214	7%	99%
10	3.076	89	3%	100%
Total	30.755	9.633	31%	-

Se propone aplicar acciones de corrección temprana especiales sobre los 12.302 estudiantes con mayor riesgo de desaprobación (resaltados con rojo). De esta manera, se estaría dando apoyo anticipado al 70% de los estudiantes que desaprueban el examen



Conclusiones

Sobre los Objetivos del Modelo Predictivo:

- o Se ha conseguido identificar información innovadora en el campo de la Educación Nacional.
- o Se han conseguido los objetivos obteniendo niveles de predicción superiores a los propuestos.

Sobre el Trabajo Realizado:

- o Las técnicas de Minerías de Datos otorgan un foco de mayor detalle al análisis, y pueden mejorar, improvisar, y adelantarse a problemas sobre cualquier contexto.
- o La metodología CRISP-DM ha facilitado la organización del proyecto, así como establecer un marco práctico y metodológico apto para desarrollar y optimizar los modelos de machine learning en un entorno controlable y escalable.



Líneas Futuras

El trabajo fue realizado con las encuestas Aprender realizadas durante el 2019. Nuestra expectativa es poder probar el modelo con las encuestas Aprender 2021 y evaluar si el nivel de predicción se mantiene estable.

Por otro lado, se pueden probar algoritmos de machine learning mas complejos, como las Redes Neuronales Aritificiales, para evaluar si los niveles de predicción aumentan.



¡Muchas Gracias!

¿Preguntas?



