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:

Violi Pablo Ezequiel

Tutores:

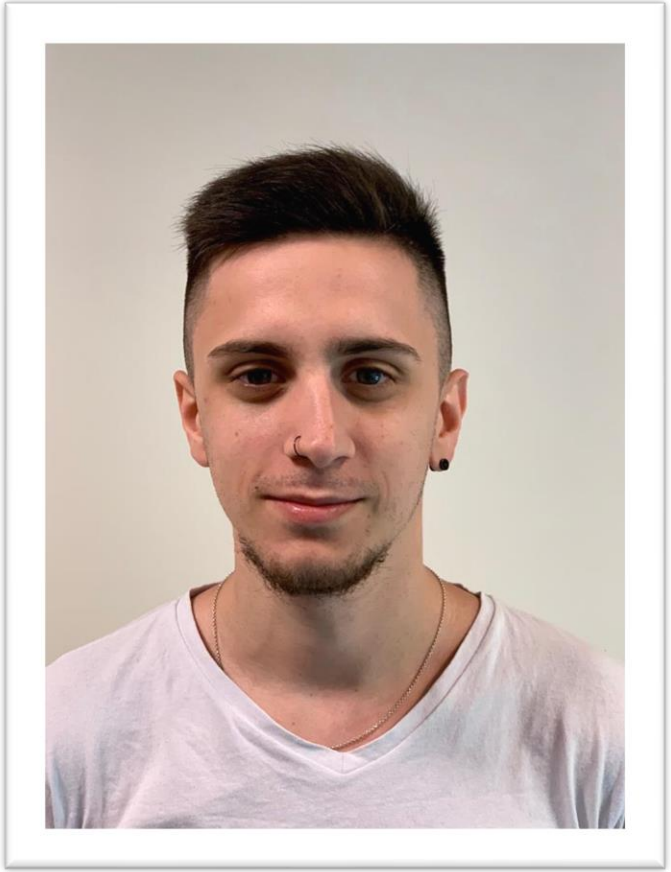
Dr. Pytel Pablo

Lic. Loidi Laura Gabriela

Departamento de Desarrollo Productivo y Tecnológico

2021

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

CRoss Industry Standard Process for Data Mining Framework

Herramientas

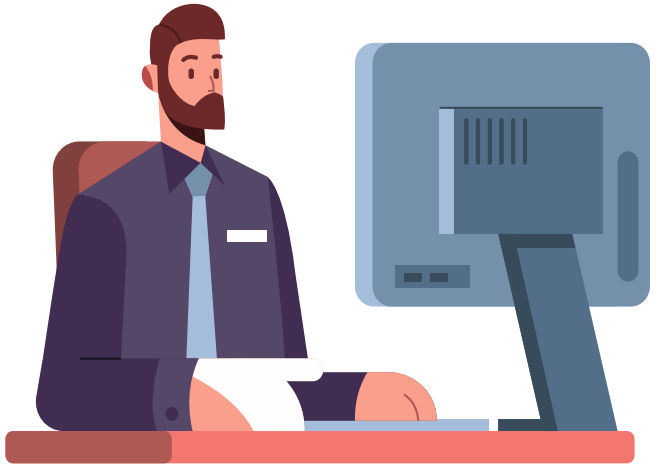
Para llevar a cabo el proyecto, se trabaja:

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

Desarrollo del Modelo Predictivo

Departamento de Desarrollo Productivo y Tecnológico

1. Comprensión del Negocio



Objetivos del negocio:

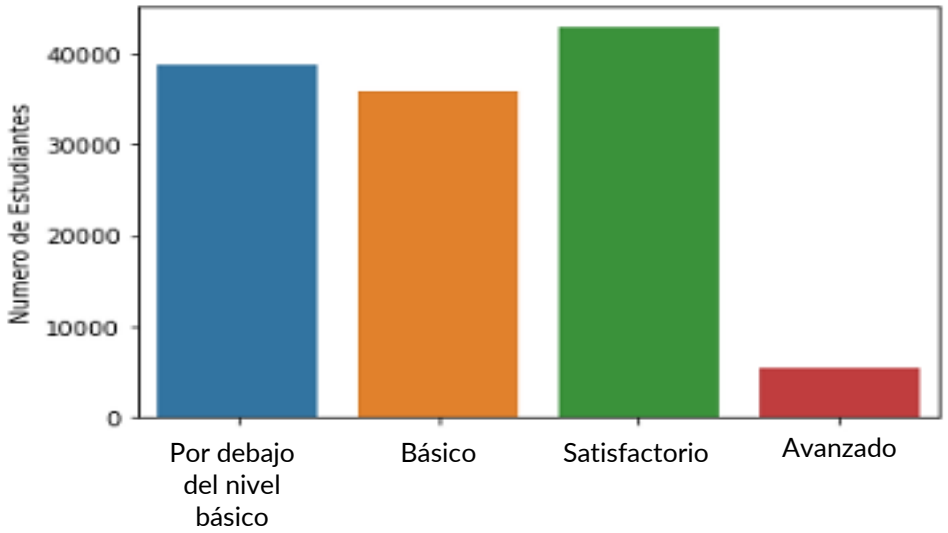
- ¿Importa el acceso a internet para la performance del estudiante?
- ¿El resultado del examen varía según los ingresos económicos del hogar?
- ¿Importa el nivel de educación de los padres para que el estudiante incorpore correctamente los conceptos de Matemática?

Objetivo de la minería de datos:

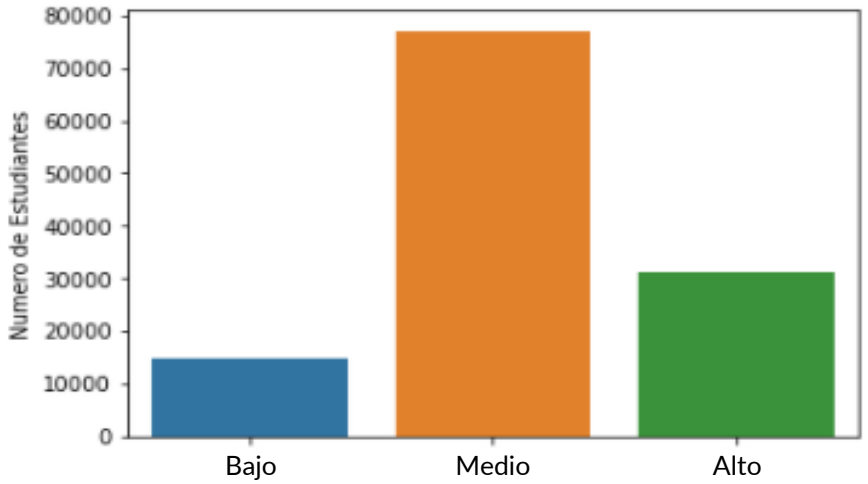
Obtener un modelo que tenga un Accuracy y un área bajo la curva ROC performantes.

2. Comprensión de los Datos

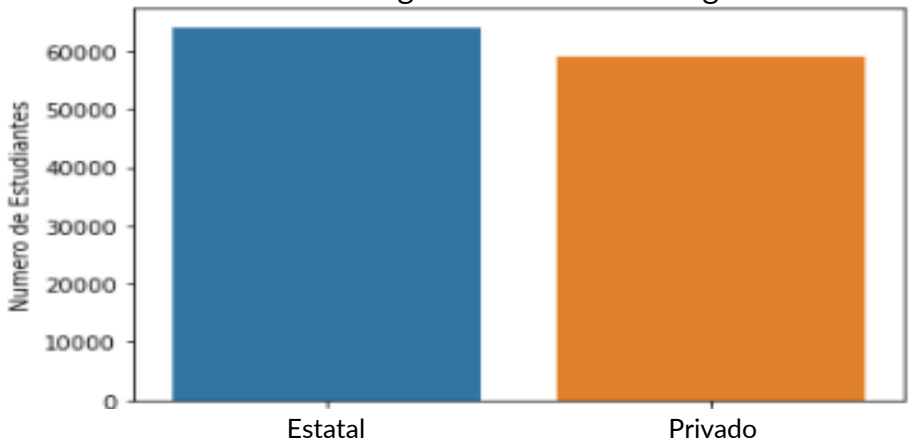
Distribución de desempeño de estudiantes en Matemática



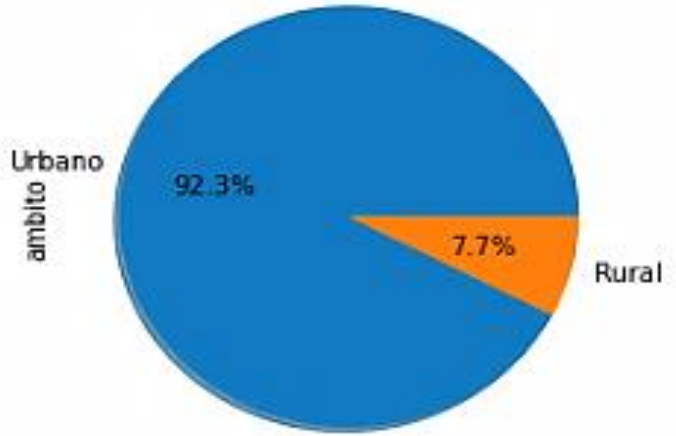
Distribución según índice Socioeconómico



Distribución según el sector del Colegio



Distribución según Ámbito Escolar



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 trabaja sobre atributos atípicos y faltantes
- 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

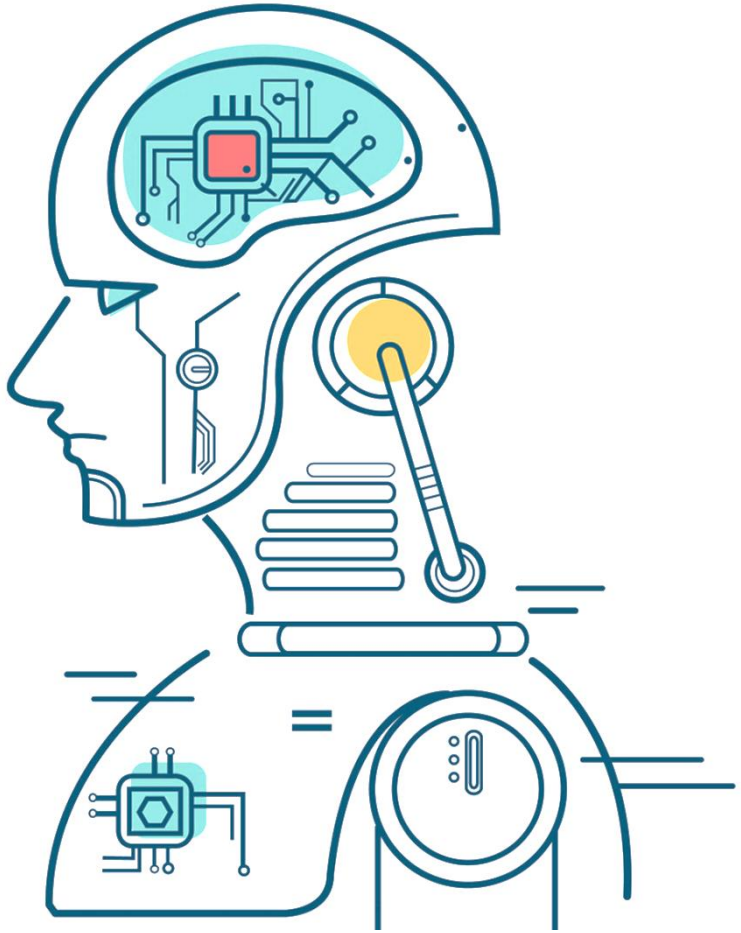
- Se separan los datos en dos datasets, uno para el entrenamiento, y otro para validación:

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

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

4. Modelado

Algoritmos utilizados durante esta etapa:



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

✓ Resultado Aceptable

✗ Resultado No Aceptable

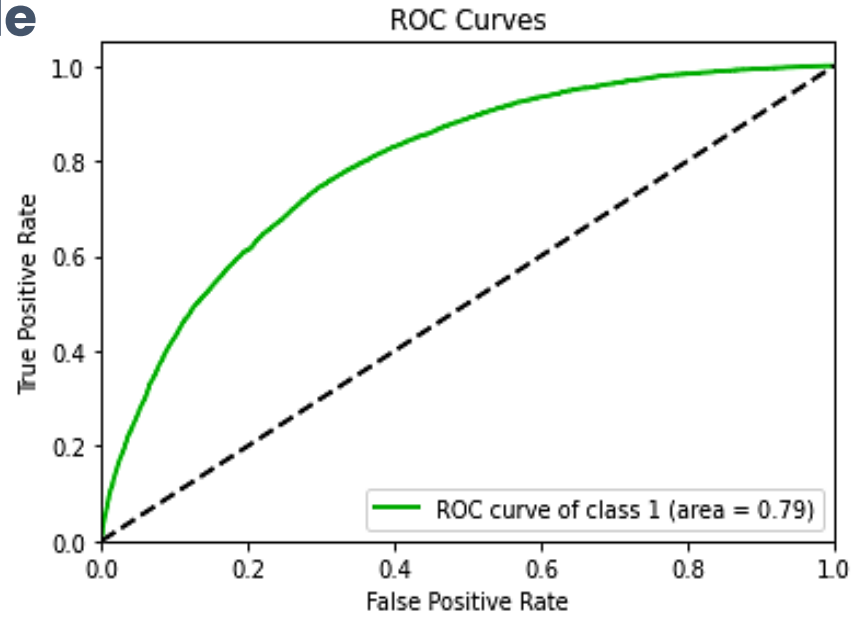
5. Evaluación

El mejor resultado fue obtenido utilizando el algoritmo LightGBM

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 Performante
- ✓ Área bajo la curva ROC Performante



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

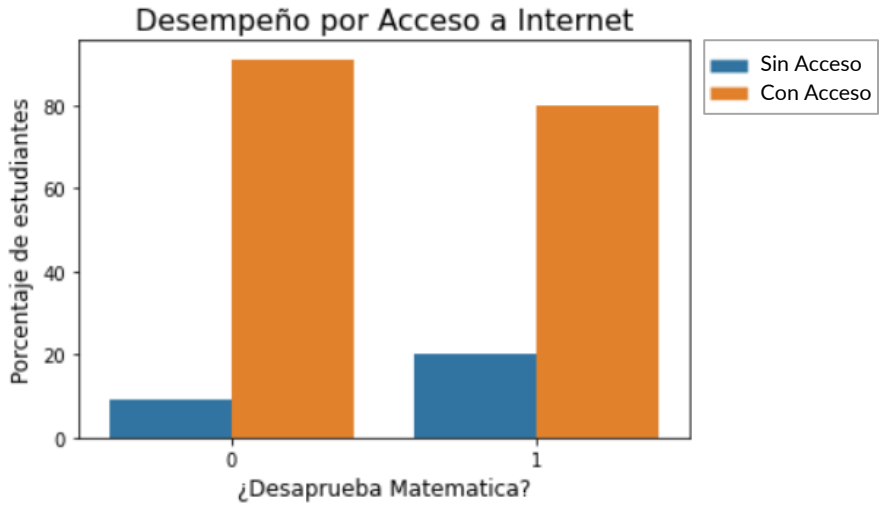
Metricas de Performance – Modelo LightGBM

5. Evaluación

Objetivo del negocio:

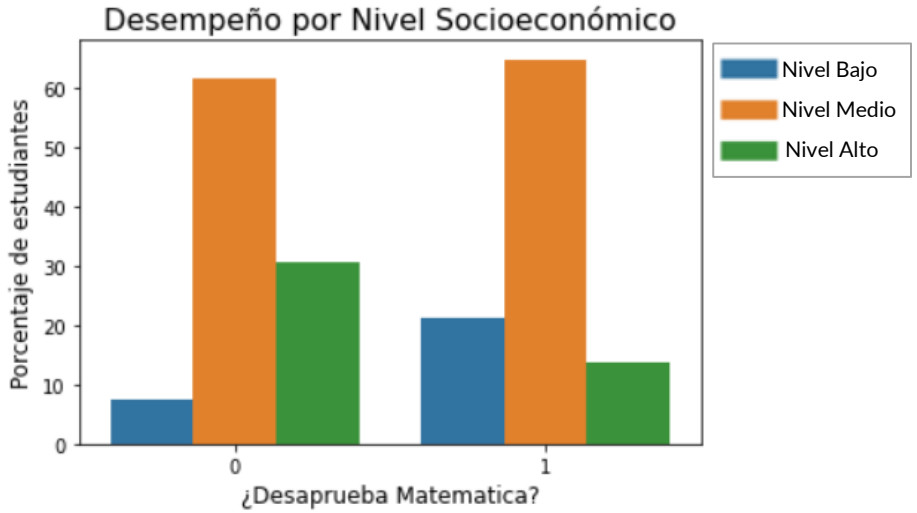
¿Importa el acceso a internet para la performance del estudiante?

El atributo tiene una correlación marcada con la performance del estudiante. Sin embargo, el modelo predictivo no la consideró entre las mas importantes



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

Efectivamente es una de las variables con mayor correlación con la performance del estudiante, y a su vez una variable con importancia alta para el modelo predictivo.



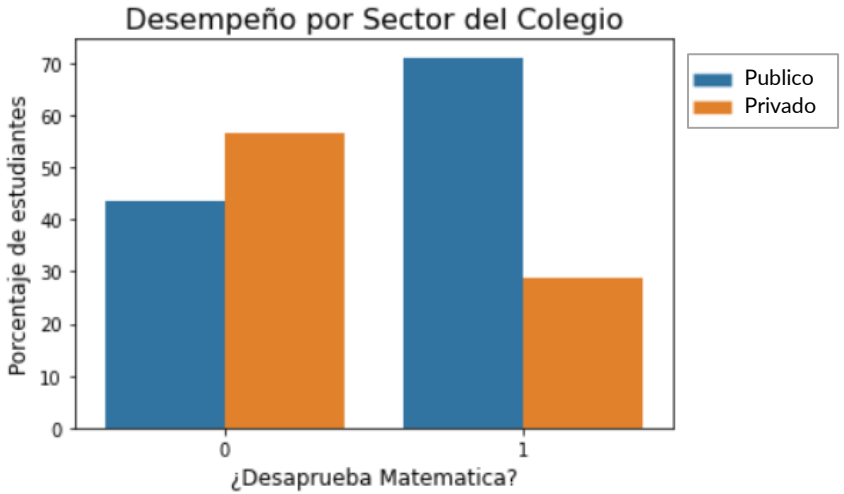
¿Importa el nivel de educación de los padres para que el estudiante incorpore correctamente los conceptos de Matemática?

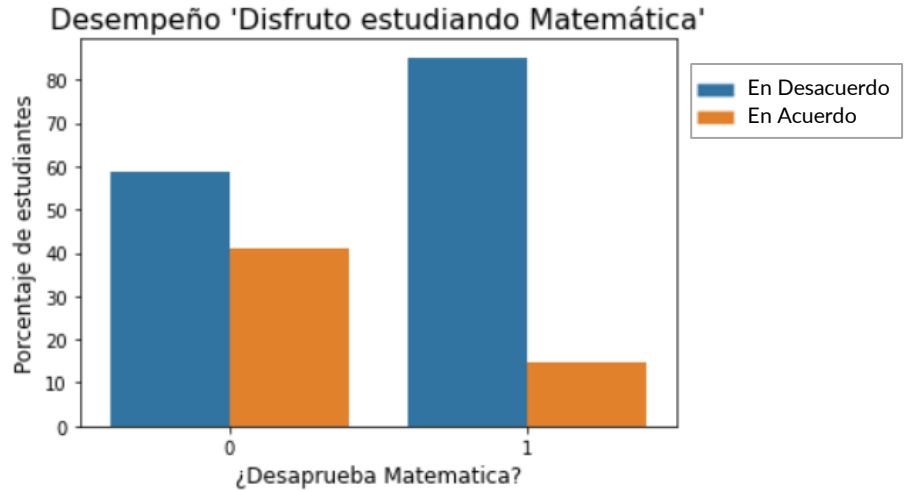
Los resultados muestran que, el nivel de educación de los padres tiene una gran influencia en la nota del estudiante. Sin embargo, el modelo predictivo considera el estudio de la madre con un mayor peso que el estudio del padre.

5. Evaluación

Variables con mayor importancia:

Ranking	Variable	Descripción
1	sector_2	¿Es sector privado o público?
2	ap40_01_1.0	¿En qué medida estás de acuerdo con la siguiente afirmación? Disfruto estudiando Matemática - Nada de Acuerdo
3	ap40_04_4.0	¿En qué medida estás de acuerdo con la siguiente afirmación? Si me lo propongo, puedo ser bueno en Matemática
4	isocioa_3.0	¿Índice socioeconómico Alto del estudiante?
5	edadA_junio2019_17.0	¿Edad de 17 años?
6	ap40_02_1.0	¿En qué medida estás de acuerdo con la siguiente afirmación? Me interesan las clases de Matemática en mi escuela
7	ap24_1.0	¿Asististe a nivel inicial? - Sí, antes de los cuatro años
8	ap39_04_3.0	¿cómo te resulta Resolver problemas y ejercicios? - Fácil
9	ap39_04_4.0	¿cómo te resulta Resolver problemas y ejercicios? - Muy Fácil
10	ap40_01_4.0	¿En qué medida estás de acuerdo con la siguiente afirmación? Disfruto estudiando Matemática - De Acuerdo





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:

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

Sobre el Trabajo Realizado:

- 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.
- 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.

También se podría realizar un A/B Testing con los estudiantes con propensión a la desaprobación del examen, para entender si accionando de manera temprana sobre los desaprobados cambia el comportamiento con respecto a los no accionados.

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

¡Muchas Gracias!

¿Preguntas?