Girmar Laucho
Magdalena Romano

Tarea final - Introducción a la Ciencia de Datos
FING, 2025

Descripción del set de datos seleccionado y problemas de calidad

Se eligió un set de datos que contiene información sobre la evaluación nacional de 2024 de acreditación de ciclo básico (AcreditaEBI). Estos datos incluyen información de contexto de la población que se inscribió para la prueba (proporcionada en el momento de la inscripción) y los resultados obtenidos en la prueba. Se incluye en el anexo una tabla con las primeras filas del dataset a modo de ejemplo. El objetivo del proyecto es diseñar un sistema que prediga las personas que tienen mayor probabilidad de obtener resultados insuficientes en esta prueba (en riesgo de no aprobar). Esto permitiría en los años siguientes, identificar oportunamente, al momento de inscripción, la población que presenta mayores obstáculos o que se encuentra en condiciones más desfavorables para enfrentar la prueba, de modo de poder proporcionar apoyos más focalizados.

En relación a las características del set de datos, se trata de un dataset relativamente grande, donde se presentan algo más de 11.000 registros de personas inscriptas para la prueba. Las variables registradas son: id, departamento, localidad, discapacidad, encuesta (esta variable que aparece integrada aquí contiene datos sobre: ascendencia, género, último año en que cursó algún tipo de enseñanza formal, último nivel educativo aprobado y subsistema, motivos (por los cuales dejó de estudiar), objetivo o motivación para dar la prueba, aspiraciones de seguir estudiando, nivel de ocupación y cómo se enteró de la prueba), observaciones, datos sobre los resultados de las 3 áreas que evaluaba la prueba y el resultado global, si había confirmado que iba a asistir y si concurrió efectivamente a dar la prueba.

Una ventaja del set de datos relacionada al proyecto es que las clases están bastante balanceadas donde hay aproximadamente un 60% de aprobación y 40% de no aprobación.
En relación a la calidad de los datos, en un análisis preliminar se observan están considerablemente depurados, sin embargo se identifican algunas inconsistencias y datos faltantes.

Respecto a las inconsistencias se identifica que la columna de localidad es una de las que presenta mayores errores, uno de ellos es que debe existir una correlación entre el campo departamento y localidad, un ejemplo es que el departamento de Montevideo no tiene localidades (tiene: municipios, barrios y secciones censales), debido a esto el campo de localidad debe ser homónimo al departamento, sin embargo hay registros de barrios en el campo localidad, lo que puede dificultar su correcta georreferenciación. Igualmente es importante cotejar que la localidad corresponda al departamento, de manera exploratoria se observó al menos un dato en donde la localidad no correspondía con el departamento (la localidad era pando pero el departamento seleccionado fue Montevideo). Hay algunos errores de ortografía del campo “localidad” que igualmente se deben corregir ya que están asociadas a variables geográficas que si no se corrigen, al mapear el dato se generarían registros incorrectos e inconsistencias al comparar con otras fuentes de datos que permitan

contextualizar la información como variables demográficas del Instituto Nacional de Estadística.

Por otra parte, los datos de la encuesta están en un mismo campo separados por punto y coma, se entiende que cada dato de la encuesta debe estar separado para posteriormente procesar y utilizar mejor los datos la aplicación del modelo. Al estar toda la información en una misma celda se pierde la visualización de los valores y dificulta la agrupación por variables (ascendencia, género, etc), representando un obstáculo adicional para el procesamiento de los datos.

Por último, hay inconsistencias en la representación de valores nulos, en algunos casos hay celdas vacías y en otros se representan con “\N”, esto es importante considerarlo al hacer el data profiling ya que la representación de dichos datos no es la misma.

Planteamiento de preguntas a resolver

La pregunta que nos surge a partir de los datos y que entendemos se puede abordar con algunas de las herramientas presentadas en el curso es si es posible detectar o predecir de antemano, en el momento de la inscripción, qué personas, según su perfil, pueden tener desempeños insuficientes en la prueba. A su vez resulta de interés entender cuáles de las variables mencionadas anteriormente se relacionan con los resultados de la prueba. Este problema se enmarca dentro del aprendizaje supervisado, particularmente conformaría un problema de clasificación. Aquí se busca predecir con precisión casos no vistos y también entender qué entradas o variables afectan al resultado y en qué medida.

Metodología y Proceso de Análisis de Datos

Como primer problema a resolver según lo mencionado anteriormente están los relacionados al ruido e inconsistencias de los datos que mediante la aplicación de data profiling se pueden resolver para posteriormente aplicar los métodos de predicción. Inicialmente en el campo de localidad determinar las inconsistencias más graves como la incorrecta relación con el campo departamento de esta forma se mejoraría el factor de exactitud del dato.

Posteriormente se puede resolver el problema de la estructura del campo “encuesta” y convertirlo en varias columnas con el objetivo de separar los campos de los valores. Se utilizarán algunas herramientas de la biblioteca pandas vistas en el curso.

Con el objetivo de entender cómo influyen las variables involucradas se puede realizar un análisis exploratorio. Aquí por ejemplo se pueden explorar y visualizar tasas de aprobación por grupos y aspectos vinculados a las correlaciones de los resultados según distintas variables como departamento o localidad, nivel educativo alcanzado o año de desvinculación del sistema educativo.

En esta línea, podría ser útil hacer un gráfico de barras del campo departamento y resultado global de la prueba, esto con el fin de observar en cuáles departamentos hubo mayor tasa de no aprobación. Igualmente al corregir los datos inconsistentes detectados anteriormente en el campo localidad se podría hacer un gráfico a un nivel geográfico más específico (el objetivo

es abordar los datos desde lo macro (departamento), a los datos más específicos como localidad o barrios).

Otra visualización preliminar sería al momento de separar los valores del campo encuesta hacer un gráfico tomando como referencia el último año en el que los aplicantes cursaron algún tipo de enseñanza formal y los campos de resultados de las tres áreas que contempla la prueba (lectura, escritura y resolución de problemas), esto con el fin de observar si existe una relación entre el último año de estudio formal y los resultados en la prueba. Igualmente se pueden establecer rangos (≥20 años, 20-10,10-5 y <5) para poder observar cómo ha sido el resultado de la prueba en los aplicantes según los rangos de tiempo transcurridos desde su último año de educación formal.

En relación al problema de clasificar y predecir resultados con el objetivo de identificar situaciones de vulnerabilidad, es indispensable trabajar en el preprocesamiento de los datos. Por ejemplo, transformar atributos categóricos y también para ser más eficientes en los tiempos de aprendizaje y eliminar algo de ruido, se podrían eliminar atributos que a diferencia de los analizados anteriormente no parecen ser demasiado significativos, como por ejemplo "cómo se enteró de la prueba" o "ascendencia". Ya que antes de realizar los entrenamientos es importante distinguir las características más discriminatorias para evitar el sobreajuste (que puede ocurrir especialmente en los árboles de decisión).

Para abordar este problema de clasificación consideramos se podrían aplicar dos técnicas o modelos. Por un lado, el método de *Regresión logística*, que como base estima la probabilidad de pertenencia a cada clase usando una función logística y luego la adapta para dar una respuesta binaria (aprueba o no aprueba). Por otro lado podría usarse la técnica de *Árboles de decisión* o en particular los árboles de clasificación, que también pueden ser apropiados para los problemas de clasificación.

Previamente al modelado, se realizará la división del conjunto de datos en el conjunto de entrenamiento y el conjunto de test (que se usará luego para evaluar el modelo).
También se ajustarían en la fase de entrenamiento los hiperparámetros (por ejemplo la profundidad máxima del árbol de decisión), mediante la técnica de validación cruzada, para no tener que separar datos específicos para la validación.
Para la evaluación del modelo se usarán las métricas de *accuracy*, *precision* y *recall*. En este caso se considera importante priorizar mejores valores de recall, es decir minimizar falsos negativos, que serían personas "en riesgo o vulnerabilidad" que el sistema no detectó. También se obtendrá la matriz de confusión para visualizar y evaluar de forma completa el desempeño del modelo.

Consideraciones éticas y estrategias de mitigación de riesgos

Respecto a la privacidad de los datos, cabe mencionar que los datos están anonimizados, sin embargo, el dataset tiene información que puede considerarse sensible tales como etnia, género y discapacidad que si se combinan con información geográfica muy específica sería posible identificar a personas. Para esto es importante plantearse si es necesario aplicar técnicas que minimicen este riesgo. Por ejemplo, podría ser conveniente utilizar una escala

geográfica más amplia contrarrestando lo planteado anteriormente sobre afinar los datos de localización.

Igualmente, el algoritmo puede tender a predecir “no aprueba” con mayor frecuencia, por algún sesgo en los datos. Por ejemplo, se puede incurrir en una subrepresentación de grupos, especialmente en personas con discapacidad, ascendencia afro y género (representación no binaria) que se pueden generalizar y hacer conclusiones no representativas a la realidad.

Es importante también asegurar que las predicciones de aquellos perfiles que correspondan a personas en riesgo de no aprobar se utilicen únicamente para el objetivo propuesto, diseñar intervenciones, apoyos, tutorías, recursos adicionales, etc. previas a la prueba, no con fines estigmatizantes ni de preselección de las personas habilitadas para dar la prueba.

Anexo:

Fig 1: Primeras 5 filas del set de datos

id	departamento	localidad	discapacidad	encuesta	observaciones	lectura	escritura	res_problemas	resultado	confirma	Concurre
447	montevideo	montevideo		Asc:Afro o Negra;Gen:Masculino;Ult:2008;Grado:1º secu	\N	2	2	2	1	1	1
468	montevideo	montevideo		Asc:Afro o Negra;Gen:Femenino;Ult:1998;Grado:1º UTU	\N	2	2	2	1	1	1
593	canelones	pando		Asc:Afro o Negra;Gen:Femenino;Ult:2005;Grado:1º secu	\N	2	2	2	1	1	1
655	montevideo	montevideo		Asc:Blanca;Gen:Femenino;Ult:1998;Grado:1º secundaria	\N	\N	\N	\N	\N	1	\N
753	montevideo	montevideo		Asc:Blanca;Gen:Masculino;Ult:2003;Grado:2º secundaria	\N	\N	\N	\N	\N	\N	\N