Proyecto final

Juárez Osorio Sandra Leticia

Resumen

I. Base de Datos

La base de datos que se analizará en este proyecto es la Encuesta Nacional sobre Acceso y Permanencia en la Educación (ENAPE) 2021 realizada por el Instituto Nacional de Estadística y Geografía, INEGI. Dicha encuesta se realizó en la población de 0 a 29 años en el Sistema Educativo Nacional (SEN), partiendo de la identificación de la población que estuvo inscrita en el anterior ciclo escolar (2020-2021), así como aquella inscrita en el ciclo escolar actual (2021-2022). Asimismo, buscaron identificar a la población que no estuvo inscrita en el anterior ni en el actual ciclo escolar, así como aquéllas que nunca han asistido a la escuela, las razones por las que no lo hicieron, entre otra información relevante sobre el tema educativo en el país.

Dicha encuesta fue realizada en 22719 hogares de las 32 entidades del país. Las preguntas realizadas toman en consideración los siguientes aspectos generales:

- Edad y sexo
- Entidad
- Nivel educativo y grado en el que se encuentra la persona (desde maternal hasta doctorado).
- Formas de evaluación de la escuela.
- Utilización de la tecnología con fines didácticos por parte de la escuela.
- Apoyo recibido en casa para la realización de tareas.
- Problemas ocasionados por la escuela en la persona.
- Desempeño escolar (materias reprobadas, con extraordinario o recursadas).
- Abandono escolar y las razones que lo ocasionaron.
- Trabajo (solo aplica para mayores de 17 años)
- Acceso en la vivienda a tecnología.
- Opinión del encuestado acerca del papel de la educación.
- Modelo de asistencia seguido por la institución educativa después de la pandemia de COVID-19.

II. Descripción del problema

En este trabajo se analizará la manera en que el apoyo obtenido en casa, el acceso a la tecnología, la necesidad de trabajar y las condiciones escolares influyen en el desempeño escolar de la persona. Para ello, se crearán dos clases:

- Mal desempeño escolar $\longrightarrow$ 0
- Buen desempeño escolar $\longrightarrow$ 1

Estas dos categorías se tomarán de las respuestas a las siguientes preguntas:

Pregunta	Respuesta	Categoría	Respuesta	Categoría
¿ Realizó algún examen extraordinario?	Si	0	No	1
¿Tuvo que recursar alguna materia?	Si	0	No	1
¿Concluyó el ciclo escolar?	No	0	Si	1
¿Reprobó, aprobó solo algunas materias, aprobó?	Reprobó / aprobó solo algunas	0	Aprobó	1

Para el análisis de este problema se eliminarán las muestras con nivel especialidad, maestría y doctorado. Debido a la poca cantidad de individuos que contestaron por separado haber realizado un examen extraordinario o haber recursado una materia o no concluir el año escolar, sería difícil tomar cada una de estas como una categoría y por ello se tomó el caso de clasificación binaria.

Sería interesante realizar un análisis de desempeño escolar a nivel primaria, pero el actual sistema educativo no permite reprobar alumnos de dicho nivel. Por ello, en la etapa de preprocesamiento será necesario remover los renglones correspondientes a personas de primaria. La encuesta abarca al sector de la población en especialidad, maestría y doctorado pero dichas personas tienen en general condiciones socio-económicas distintas, por lo que tampoco se tendrán en cuenta en este estudio.

III. Metodología

A. Preprocesamiento

Se removieron las columnas correspondientes a los menores de edad que estuvieron o fueron inscritos en educación maternal y guarderías (columnas `PA3.3MODMAT, P32, PB3.5MODMAT`). Además se eliminaron los renglones correspondientes a las personas que acuden a kínder, primaria, especialidad y maestría (en la columna`PA3.3NIVEL`).

Además, se excluyeron las columnas correspondientes a `FILTRO C`, las cuales corresponden a preguntas hechas a las personas que contestaron no haber estado inscritos ni en el ciclo anterior ni en el siguiente. Si bien resultaría interesante analizar las causas de que

una persona no se encuentre asistiendo a la escuela, las preguntas de esta encuesta se enfocaron mayoritariamente a analizar el desempeño escolar en el ciclo anterior. Únicamente se contarían con 8 features para analizar dicha situación.

Los datos de la encuesta están organizados en dos tablas, una correspondiente a la vivienda y otra correspondiente a las preguntas relacionadas a la educación. Ambas tablas se conectan a través de la columna `FOLIO`, la cual es una variable única para la tabla de vivienda y puede estar repetida en la tabla de educación. Esto debido a que en una misma vivienda pueden habitar más de una persona en etapa escolar. Entonces, es necesario añadir a cada renglón de la tabla de educación sus correspondientes valores de vivienda.

Para tratar con los valores faltantes, se aplicará una estrategia que depende de cada categoría. En algunos casos, es posible rellenar la información faltante con la ayuda de la información en otra columna. Por ejemplo en las columnas `PB3.13 (1-7)` en donde se pregunta quién ayudó a la persona en sus tareas, para las personas que respondieron *Nadie* es posible añadir un 0 en las partes donde dice *Mamá*, *papá*, etc. En otras preguntas, el encuestador no siguió preguntando cierto bloque de preguntas dependiendo de ciertas respuestas y dichos espacios se dejaron en blanco por lo que fue necesario rellenarlos con ceros. Finalmente, se descartarán las muestras que aún después de dichas consideracioens aún tengan 10 o más valores faltantes.

B. Data Augmentation

La clase correspondiente a los alumnos con bajo desempeño escolar se encontraba desbalanceada con respecto a los alumnos con buen desempeño. Después de la limpieza de datos se encontraron 3365 muestras de bajo desempeño y 7882 muestras de alumnos con buen desempeño. Por ello, se decidió aplicar la técnica SMOTE (Bowyer y cols., 2011) en dicha clase, con lo que se obtuvieron finalmente 6728 muestras de dicha clase. Las muestras fueron divididas en training, validation y testing en una proporción de 4/6,1/6,1/6, respectivamente.

C. Feature Egineering

Se realizó la técnica de backwards selection en los 96 features. El número de features elegidos será guardado como un hiperparámetro en *ML flow* para realizar una comparativa de cómo se relaciona esta cantidad con las métricas empleadas.

D. Clasificadores

Modelos Lineales.
Se empleó un modelo de clasificación lineal utilizando Least Squared Error como función de coste. Además, se utilizó el modelo de regularización de Ridge monitoreando como hiperparámetro a la constante λ que será tomada de un grid search. Las métricas consideradas

son precisión, recall y accuracy para training, validation y testing.

Mixture of Gaussians.
Utilizando los datos de entrenamiento se obtuvo el ajuste a dos funciones gaussianas. Utilizando los valores de μ y σ obtenidos para cada una, se calculó la probabilidad de que cada muestra del set de datos de validation perteneciera a la clase. En el cálculo de la función gaussiana fue necesario aplicar una regularización sobre la matriz de covarianza para evitar obtener matrices singulares.

Logistic Regression.
Fue necesario emplear la versión regularizada de este método para evitar tener matrices singulares en el algoritmo de Newton-Raphson. Se monitoreará dicho parámetro de regularización para relacionarlo con las métricas.

Multilayer Perceptron.
Se introdujeron los features seleccionados un MLP. Como función de activación se utilizó $\tanh x$. Los hiperparámetros a monitorear serán la arquitectura de la red, el learning rate y como métricas se considerarán los pasos en los que el método convergió así como la precisión, recall y accuracy de training y validation y posteriormente de testing.

Referencias

Bowyer, K. W., Chawla, N. V., Hall, L. O., y Kegelmeyer, W. P. (2011). SMOTE: synthetic minority over-sampling technique. *CoRR*, *abs/1106.1813*. Descargado de `http://arxiv.org/abs/1106.1813`