Modelo predictivo sobre la deserción en educación media por municipio en Colombia

J. Juan*, S. Camilo§
Optimización, Universidad del Norte, Barranquilla, Colombia
*djulioj@uninorte.edu.co
§cjsinning@uninorte.edu.co

***Resumen*—La base de datos MEN ESTADÍSTICAS EN EDUCACIÓN EN PREESCOLAR, BÁSICA Y MEDIA POR MUNICIPIO fue escogida para la realización del proyecto de analítica de datos. Lo anterior, por la importancia que amerita ser puesta en la educación de nuestro país y, así con ello, analizar estos datos de manera gráfica y desarrollar un modelo predictivo para estimar la deserción en educación media por municipio en el país, y responder hipótesis útiles para la determinación de posibles estrategias a tomar con el objetivo de mejorar el acceso y la calidad de la educación impartida en el territorio colombiano.**

***Index Terms*—Educación media, Aprobación, Deserción, Departamento.**

***Abstract*—The MEN STATISTICS ON PRESCHOOL, BASIC AND HIGH SCHOOL EDUCATION BY MUNICIPALITY database was chosen to carry out the data analytics project. The foregoing, due to the importance that deserves to be placed on education in our country and, thus, analyze these data graphically and develop a predictive model to estimate dropout in secondary education by municipality in the country, and answer useful hypotheses. for the determination of possible strategies to take in order to improve access and quality of education provided in the Colombian territory.**

***Index Terms*—Medium eduaction, Approval, Desertion, Department.**

I. Introducción

El presente informe se hizo con el objetivo de dejar evidencia sobre el procedimiento llevado a cabo en el momento de desarrollar el modelo predictivo. La base de datos seleccionada es MEN ESTADÍSTICAS EN EDUCACIÓN EN PREESCOLAR, BÁSICA Y MEDIA POR MUNICIPIO ubicada en el sitio web gubernamental https://www.datos.gov.co/Educaci-n/MEN_ESTADISTICAS_EN_EDUCACION_EN_PREESCOLAR-B-SICA/nudc-7mev. La mencionada base de datos nos presenta información estadística correspondiente a los niveles educativos de preescolar, básica y media por municipios, con información oficial recolectada entre los años 2011, 2019 y datos preliminares del año 2020[1].

El modelo predictivo desarrollado fue limitado a analizar y predecir datos referentes a la educación media, Siendo más concreto, se usa la aprobación y el departamento en cuestión para predecir una deserción baja o alta en cierto municipio.

A continuación se ahondará en los objetivos y motivos por el cual se llevó a cabo este modelo, además se plantean las hipótesis y se da una descripción del modelo en cuestión, para finalmente realizar un análisis de este y extraer conclusiones.

II. Objetivos

II-A. Objetivo general

Desarrollar un modelo predictivo con el cual se pueda pronosticar la deserción en la educación media de cierto municipio.

II-B. Objetivos específicos

- Seleccionar una base de datos
- Analizar datos provenientes de la base de datos
- Desarrollar el modelo predictivo

III. Justificación

El presente informe es necesario para dejar evidencia en lo que respecta al proyecto de analítica de datos, donde se analizó la información, se diseñó e implementó un modelo predictivo, esto con el objetivo de aprobar o rechazar hipótesis.

Este modelo, por su naturaleza y por la base de datos usada, tiene el poder de ser útil en la construcción de políticas y administración de recursos destinados a la educación del país.

IV. Hipótesis

Se plantea entonces con estos datos la siguiente pregunta: ¿Existe alguna relación entre el porcentaje de aprobación de un municipio, su departamento y la deserción escolar en educación media?

Y por consiguiente se propone la hipótesis: Las zonas del país con mayor índice de ruralidad son aquellas en los cuales se presenta una mayor deserción en la educación media.

V. Metodología

Primeramente, haciendo uso de la página gubernamental Datos Abiertos, se seleccionó la base de datos más adecuada para el proyecto, ya con la base de datos seleccionada se realizó un análisis gráfico de los datos y las relaciones existentes entre las variables que este posee, teniendo toda está información entonces se procedió a plantear una hipótesis para ser comprobada haciendo uso del modelo. Finalmente, se desarrolló el modelo predictivo y se usó este para la comprobación de la hipótesis.

VI. Base de datos

VI-A. Características

La base de datos MEN ESTADÍSTICAS EN EDUCACIÓN EN PREESCOLAR, BÁSICA Y MEDIA POR MUNICIPIO, es propiedad del Ministerio de Educación Nacional. La base de datos fue actualizada por última vez el 17 de diciembre del 2021, se encuentra en español, tiene una cobertura de tipo nacional, la actualización es de frecuencia anual y fue emitida el 31 de diciembre del 2015.

Adicionalmente, la base de datos cuenta con 41 columnas, es decir, cada fila (tupla) cuenta con 41 atributos considerados indicadores educativos. También, cuenta con 11,200 filas, las cuales superan en número los 10,000 registros mínimos requeridos para realizar el proyecto de analítica de datos.

Aunado a todo lo anterior, es notable que la base de datos posee cierta popularidad, ya que cuenta con 79,300 visitas y 14,500 descargas.

VI-B. Descripción

Los atributos de la tabla de datos comprenden diversos tópicos, pero este proyecto solo usará:

1. DEPARTAMENTO: Contiene el nombre del departamento.
2. APROBACIÓN_MEDIA: Tasa de aprobación de estudiantes del sector oficial en media. Identifica el porcentaje de alumnos en este nivel educativo que aprueba de acuerdo con los planes y programas de estudio vigentes [1].
3. DESERCIÓN_MEDIA: Tasa de deserción intra-anual del sector oficial en media. Identifica la proporción de alumnos matriculados que, por factores culturales, coyunturales o de prestación del servicio educativo, abandonan sus estudios durante el año lectivo. [1].

Por otro lado, lo que respecta a los registros recolectados, se tiene que cada una corresponde a un municipio de Colombia en un tiempo específico y, por lo tanto, la base de datos posee once mil dos cientos registros que representan mediciones hechas en pueblos de Colombia a lo largo de diferentes años.

Para información exacta de las definiciones de cada atributo, su utilidad y explicación de la metodología de recolección, visitar el sitio web https://www.datos.gov.co/Educaci-n/MEN_ESTADISTICAS_EN_EDUCACION_EN_PREESCOLAR-B-SICA/nudc-7mev.

VII. Análisis exploratorio

Primeramente, mediante el uso de la librería seaborn disponible para Python, se graficó la relación entre las columnas relevantes para el modelo, dando como resultado lo mostrado en la figura 1.

Esta gráfica mostró que existía una posible relación entre estos dos valores, ya con esto se orientó el análisis en un sentido en el que se relacionarían estas variables. En este proceso se descartaron otras variables como la cantidad de sedes conectadas a internet porque mostró no ser relevante para la deserción y además de esto la reprobación fue descartada dado que era una variable complementaría a aprobación.

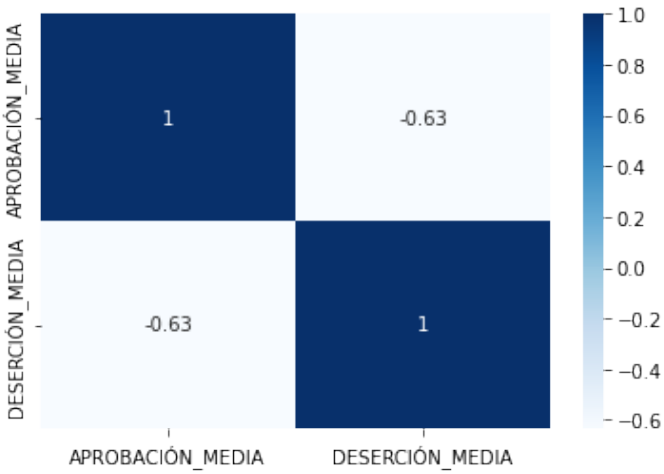

Figura 1. Gráfico de correlación entre la columna aprobación y la columna de deserción

Luego, se ilustró mediante el uso de gráfico de violines la deserción en correspondiente a cada departamento, como se muestra en la figura 2

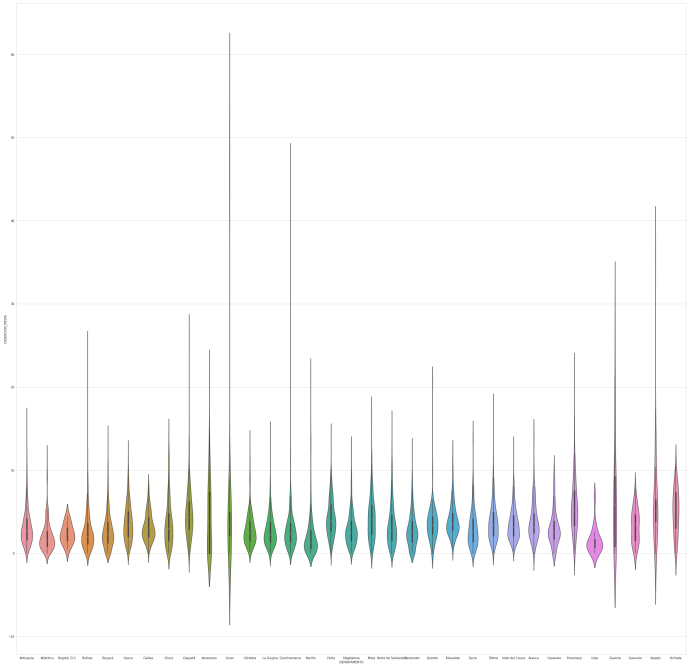

Figura 2. Gráfica de Deserción vs. Departamento

A partir de este gráfico de violines es posible notar los departamentos en los cuales hay una mayor deserción y teniendo esto se pueden plantear ciertas hipótesis del porqué de este comportamiento.

VIII. Modelo

VIII-A. Procedimiento

Para el desarrollo del modelo se usó la librería pandas para cargar la base de datos y poder modificarla y usarla en Python.

Después de cargar la base de datos mediante pandas está se modificó y se dejaron solo las columnas DEPARTAMENTO, DESERCIÓN_MEDIA Y APROBACIÓN_MEDIA.

Usando la función get_dummies ofrecida por pandas, se transformó la variable categórica departamento en varias columnas que si pueden ser usadas para entrenar el modelo.

A la base de datos se le agrega una columna que corresponde a una variable categórica que tendrá 1 si la deserción es lo que consideramos alta (deserción $\geqslant 4\,\%$) o baja, y se remueve la columna original de DESERCIÓN_MEDIA.

Se definieron las variables de entrada y salida y haciendo uso de la librería sklearn se dividió la base de datos en datos de entrenamiento y datos de validación. Luego se usó la función de LogisticRegression y se entrenó y validó con los datos anteriormente definidos.

VIII-B. Forma

El modelo hecho puede ser representado como se observa en la figura 3. En este se representan las entradas y la salida que representa 1 si se considera deserción alta o 0 si está es baja.

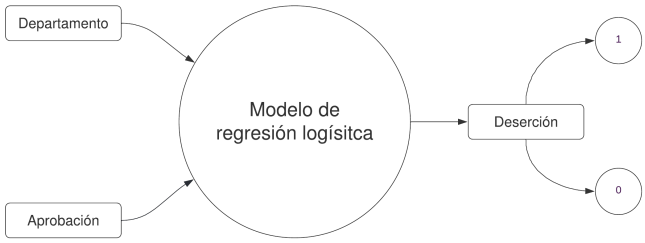

Figura 3. Modelo de regresión logística

VIII-C. Resultados

El modelo se entrenó y válido mil veces, obteniendo que el puntaje promedio que obtiene este es de un 75 % y la varianza de un 4 %, como se observa en la figura 4.

VIII-D. Análisis

En el desarrollo del modelo se logró comprobar la relación entre las variables 'DEPARTAMENTO' y 'APROBACIÓN_MEDIA' para la estimación de la variable 'DESERCIÓN', y como el uso de regresión logística fue ideal para este problema dato que este corresponde a uno de clasificación. Aunado a esto se obtuvieron puntajes consistentes con poca varianza entre los diferentes conjuntos de datos generados.

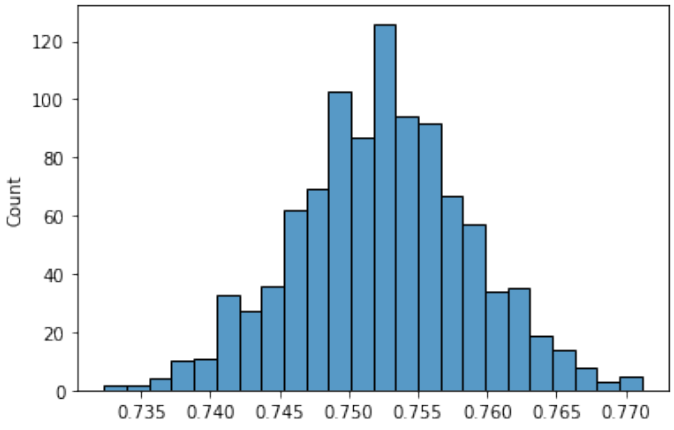

Figura 4. Gráfico de puntajes del modelo

IX. Conclusión

El modelo fue desarrollado y se llegó a un resultado que se considera exitoso, ahora con este modelo entrenado sé estable la veracidad de la hipótesis hecha.

Se toma como base el estudio hecho por las naciones unidas en el que se clasifican las diferentes provincias que tiene el país según su ruralidad. En la figura 5 se puede observar gráficamente esta clasificación.

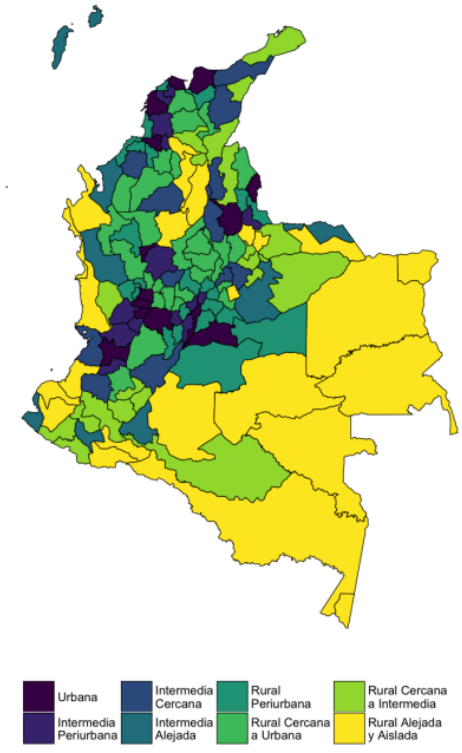

Figura 5. Clasificación de provincias por ruralidad [2]

Se tomarán los departamentos que el informe considera

como rural, alejada y aisladas y se usaran municipios pertenecientes a estas para comprobar la hipótesis usando el modelo.

Se usarán los departamentos: Vichada, Guainía, Vaupés y Putumayo. Los datos indican que en un 69 % el modelo predijo que diversos municipios pertenecientes a estos departamentos tendrán deserción alta en la educación media.

REFERENCIAS

1 de Educación Nacional, M., "Men_estadisticas_en_educacion_en_preescolar, bÁsica y media_por_municipio," Dec 2015. [Online]. Available: https://www.datos.gov.co/Educaci-n/MEN_ESTADISTICAS_EN_EDUCACION_EN_PREESCOLAR-B-SICA/nudc-7mev

2 para América Latina y el Caribe Oficina de la CEPAL en Bogotá, C. E., Rámirez, J. C., and de Aguas, J. M. Naciones Unidad, 2017.