Determinación de factores causales de deserción escolar en el Municipio de Zipaquirá mediante un Modelo de Regresión Logit.

Proyecto de grado presentado como requisito para la obtención del título de profesional en estadística de la Universidad Santo Tomás.

Estudiante: Matthew Enrique Rentería Guzmán

Tutores: Lida Fonseca, Oscar Beltrán

Universidad Santo Tomás

Bogotá

2023

Resumen

La deserción escolar es un problema que incide en el desarrollo de una nación, pues repercute en el futuro de niños, adolescentes y jóvenes adultos, ya que aumenta la dificultad en la obtención de empleos estables y/o bien remunerados. El estudio de las características de la deserción escolar ha sido objeto de investigación permanente realizado por el Ministerio de Educación Nacional (MEN) cuyos avances y retrocesos a nivel nacional han sido divulgados a través de informes técnicos tales como sus notas técnicas.

La Secretaría de Educación de Zipaquirá no es ajena al tema de la deserción, por ello solía buscar medir dicha situación a través de la plataforma SIMPADE que funcionó hasta el 2019. Desde el 2020 a la fecha no cuenta con un sistema adecuado para medir la deserción, situación lo cual ha repercutido en la matrícula anual y en otros indicadores.

El presente estudio busca identificar cuáles son los factores que más influyen en la deserción escolar del municipio de Zipaquirá tanto en colegios privados como oficiales, a través de la aplicación de un modelo Logit para determinar estos factores.

Abstract

School dropout is a problem that affects the development of a nation, since it affects the future of children, adolescents and young adults, since it increases the difficulty in obtaining stable and/or well-paid jobs. The study of the characteristics of school dropout has been the subject of permanent research carried out by the Ministry of National Education (MEN) whose advances and setbacks at the national level have been disclosed through technical reports such as its technical notes.

The Zipaquirá Secretary of Education is no stranger to the issue of desertion, which is why it used to seek to measure this situation through the SIMPADE platform that worked until 2019. From 2020 to date, it does not have an adequate system to measure desertion, a situation which has had an impact on annual enrollment and other indicators.

The present study seeks to identify the factors that most influence school dropout in the municipality of Zipaquirá, both in private and official schools, through the application of a Logit model to determine these factors.

1 Introducción

En Colombia de acuerdo con MEN la deserción escolar “es un fenómeno que implica el abandono del proceso educativo de niños, niñas, adolescentes y jóvenes” (MEN. 2022), y es una problemática de suma importancia para el desarrollo de un país esto debido a que según Cervini. (2005) este factor es uno de los principales al momento de mantener el esquema de pobreza en la región.

Sobre el concepto de la deserción existen múltiples definiciones o formas de cálculo, para ser específicos la deserción puede calcularse a partir de lo que ocurre dentro de un año mediante la Tasa de deserción intraanual, entre años consecutivos a través de la Tasa de deserción interanual (MEN. 2020). De acuerdo con el MEN en Colombia se mide la deserción de forma intraanual para los niveles de educación preescolar, básica y media la cual ofrece información valiosa para conocer la proporción de estudiantes que no terminan un año lectivo, lo cual permite conocer la evolución a corto plazo de este fenómeno, pero con la limitante de que no ofrece información acerca de la evolución de esta situación a mediano y largo plazo.

Para el registro de la deserción escolar durante la década pasada el ministerio usó las plataformas SIMAT (Sistema Integrado de Matrícula) y SIMPADE (Sistema de Información para el Monitoreo, Prevención y Análisis de la Deserción Escolar), las cuales se enfocaban en el registro y control de los estudiantes que ingresaban a institutos de educación preescolar básica y media, en colegios tanto públicos como privados.

De acuerdo con investigaciones propias realizadas en la Secretaria de Educación de Zipaquirá el municipio desde el año 2009 (-MEN, Portal de modernizacion de Secretarias de Educación, s.f.) posee la certificación en educación por lo cual es autónomo en la administración de su sistema educativo en los niveles educativos de preescolar, básica, media y educación para el trabajo y desarrollo humano que está sujeta a seguir los lineamientos establecidos por MEN. Esto ha permitido mayor libertad del manejo de la información en la base de datos de la plataforma del Sistema Integrado de Matricula- SIMAT.

El estudio del presente trabajo se centra en la identificación de algunos factores que resultan influyentes en el momento en que un estudiante entre los 5 y los 18 años decide abandonar el estudio. Para ello se empleará la base de datos de SIMAT con los registros entre 2016 y 2022 para, una vez identificados esos factores, construir un modelo probabilístico que permita identificar la influencia de estas variables sobre la mismo.

2 Antecedentes

En las investigaciones que se han realizados respecto a la deserción escolar a nivel nacional se destacan las realizadas por el MEN a través de su nota técnica para el año 2022, en la cual se investigó el comportamiento de la tasa de deserción intraanual (o sea la que ocurre dentro de un año), y el método utilizado fue manejando la base de la plataforma del Sistema Integrado de Matricula- SIMAT con la información de los años 2015 a 2018 y para el análisis de los factores asociados se utilizó un modelo de regresión lineal en donde construyeron un conjunto de variables para obtener nuevos factores relevantes a tener en cuenta. Estas variables que fueron construidas teniendo en cuenta el cambio de colegio, si estudió el año anterior y la disponibilidad del siguiente grado. Para la variable respuesta usaron una desagregación de la variable

estado en donde tomaron la variable “retirado” como la condición para desertor y como variables independientes usaron las tres (3) anteriormente mencionadas y un conjunto de variables que usan factores individuales de la persona tales como lo son el sexo, la edad o discapacidades y un conjunto de factores de las instituciones educativas tales como lo son si es urbana o rural, la jornada o la metodología. Como tal los resultados de esta investigación mostraron que los factores que más influyen en la deserción son el género, la edad, y la repitencia.

Por su parte Zapata, D. (2021), en la cual investigó métodos para la detección de estudiantes en riesgo de deserción, para ello uso una base obtenida de la plataforma SIMAT y la IEBS en la etapa de preprocesamiento de los datos a través de métodos de minería de datos en donde para la selección de variables relevantes uso un método de análisis discriminante conocido como mCMC, luego procedió con el balanceo de la base de estudio y por último realizó la construcción de varios modelos de clasificación como lo son el SVM, Random Forrest, SVM-RBF, ANN-MLP y el Gradient Boosting hasta llegar al definitivo. El resultado muestra que entre las características identificadas están el género, la edad, el número de hermanos y la distancia a la institución educativa.

Retavizca, M. (2016), realizó un estudio de la deserción escolar con información del año 2015 de la encuesta de calidad de vida del DANE en donde utilizó un Modelo Logit en el cual ya se posee una variable respuesta que es si el núcleo familiar posee un niño entre 5 y 17 años que no estudie, pero sí se encuentre trabajando, y el resto son características del hogar y de sus padres. Los resultados obtenidos muestran que factores tales como la edad tanto del menor como del jefe del hogar, el género tanto del menor como del jefe del hogar o si es de sector urbano o rural influyen en la permanencia o no en los estudios.

Rodríguez, D. (2021), investigó los factores de deserción en las localidades de Bogotá durante los años 2011, 2014 y 2017 usando la información de la encuesta multipropósito de Bogotá, el modelo utilizado para esta investigación fue el Modelo Logit en donde usó la variable deserción global de la base obtenida y cuyos resultados dieron que hay disparidad educacional entre localidades, y que las de mayor nivel son Bosa, Usme, Ciudad Bolívar y San Cristóbal, en tanto que la que se encuentra en un nivel preocupante es La Candelaria.

3 Problema

De acuerdo con entrevistas realizadas a funcionarios de la Secretaria de Educación de Zipaquirá, en el año 2019 el gobierno desactivó la plataforma del Sistema de información para el Monitoreo, Prevención y Análisis de la Deserción Escolar - SIMPADE , la cual , era la encargada del registro, análisis y elaboración de informes sobre la deserción estudiantil y toda la operación fue transferida al Ministerio de Educación Nacional -MEN, el cual envía a las secretarias de educación de cada uno de los municipios informes ya resumidos de las cifras de deserción escolar, que no va más allá de los porcentajes obtenidos.

En atención a la información consignada en el SIMAT - se buscará a través de un modelo -identificar factores causales de la deserción escolar, que permitan establecer una idea de cuáles pueden ser algunos de los elementos que hacen al estudiante tomar la decisión de desertar del sistema escolar.

4 Justificación

Las investigaciones ya mencionadas se han desarrollado hasta el año 2019, teniendo como característica información limitada y desactualizada dada la situación de la pandemia, es allí donde la presente investigación a través de la información consignada en el SIMAT puede facilitar el análisis, contraste y la observación del comportamiento de la deserción escolar en el Municipio de Zipaquirá.

5 Pregunta problema

Dada la situación se busca responder a la pregunta ¿Qué factores son los más influyentes en el fenómeno de la deserción escolar, en los niveles de preescolar, básica y media en el Municipio de Zipaquirá?

6 Objetivos

6.1 Objetivo general

Identificar un conjunto de factores relevantes que influyen en la deserción escolar en el Municipio de Zipaquirá, a través de un modelo Logit.

6.2 Objetivos específicos

- Evaluar, a través de un modelo estadístico, cada una de las variables para identificar cuáles son influyentes en la deserción escolar en el municipio de Zipaquirá.
- Comparar los distintos tipos de balanceo de la base, y los resultados del modelo de acuerdo con el tipo de base balanceada.
- Realizar un análisis de las variables relevantes del modelo que permita explicar por qué estas pueden afectar la continuidad de los estudiantes dentro del sistema educativo.

7 Marco teórico

7.1 Deserción escolar

De acuerdo con MEN. (2022) se entiende como deserción escolar “el abandono del sistema escolar por parte de los estudiantes, provocado por la combinación de factores que se generan tanto al interior del sistema como en contextos de tipo social, familiar, individual y del entorno”. Su estudio a través del tiempo puede ser de 2 maneras, intraanual que consiste en el estudio del abandono del estudio durante el transcurso de un año e interanual que consiste en el abandono del estudio a través de varios años consecutivos.

7.2 SIMAT

El Sistema Integrado de Matrícula (SIMAT) es una plataforma que permite organizar y controlar el proceso de matrícula en todas sus etapas, así como tener una fuente de información confiable y disponible para la toma de decisiones. La plataforma contribuye a mejorar la gestión del proceso de matrícula de cada secretaría de educación, permitiendo realizar consolidar información, generar reportes y realizar seguimiento a todo el proceso.

7.3 Nivel de educación

De acuerdo con la Ley 115 de 1994 en su artículo 11 indica que:

7.3.1 Preescolar: Que comprenderá mínimo un grado obligatorio (transición).

7.3.2 Básica: Con una duración de nueve grados que se desarrollará en dos ciclos: La educación básica primaria de cinco grados y la educación básica secundaria de cuatro grados.

7.3.3 Media: Con una duración de dos grados, son los últimos grados del periodo lectivo escolar.

7.4 Modelo Logit

La Regresión logística, es un método de regresión de respuesta cualitativa dicotómica desarrollado por David Cox en 1938 que permite estimar la probabilidad de una variable cualitativa binaria en función de 1 o más variables cuantitativas y cualitativas estas con convertidas en variables *Dummies*. Una de las principales aplicaciones de la regresión logística es la de clasificación binaria, en el que las observaciones se clasifican en un grupo u otro dependiendo del valor que tome la variable empleada como predictor. Por ejemplo, clasificar a un individuo desconocido como hombre o mujer en función del tamaño de la mandíbula.

Es importante tener en cuenta que, aunque la regresión logística permite clasificar, se trata de un modelo de regresión que modela el logaritmo de la probabilidad de pertenecer a cada grupo. La asignación final se hace en función de las probabilidades predichas. Incluso cuando la respuesta de interés no es originalmente de tipo binario, algunos investigadores han dicotomizado la respuesta para que la probabilidad de éxito pueda ajustarse mediante regresión logística.

El modelo logístico posee la forma:

$$E(y) = \frac{e^{x^T\beta}}{1 + e^{x^T\beta}} = \frac{1}{1 + e^{-x^T\beta}}$$

Donde:

X: Es el vector de variables explicativas.

B: Es el vector de parámetros.

Debido a eso, la función de probabilidad $\pi(\mathbf{x})$ es de la forma:

$$\pi_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}} = \frac{1}{1 + e^{-x^T \beta}}$$

Donde:

P: Son los valores observados de las variables explicativas.

Con esta función de probabilidad podemos hallar la relación de probabilidades de la siguiente forma:

$$1 - \pi_i = \frac{1}{1 + e^{x^T \beta}} \rightarrow \frac{\pi_i}{1 - \pi_i} = \frac{1 + e^{x^T \beta}}{1 + e^{-x^T \beta}}$$

El resultado obtenido de aquí es conocido como transformación logit para la probabilidad π_i y de esa relación podremos sacar la razón de probabilidades (odds ratio). Sacando del odds ratio su logaritmo natural podemos obtener lo siguiente:

$$Ln\left(\frac{\pi_i}{1 - \pi_i}\right) = x^T \beta$$

Lo cual nos permite concluir que el log del odds ratio es lineal tanto en variables como en los parámetros, y que dado a que el comportamiento de los datos en valores 0 o 1 el comportamiento en la curva de tendencia es en forma de s o sigmoidal (ejemplo en la siguiente imagen).

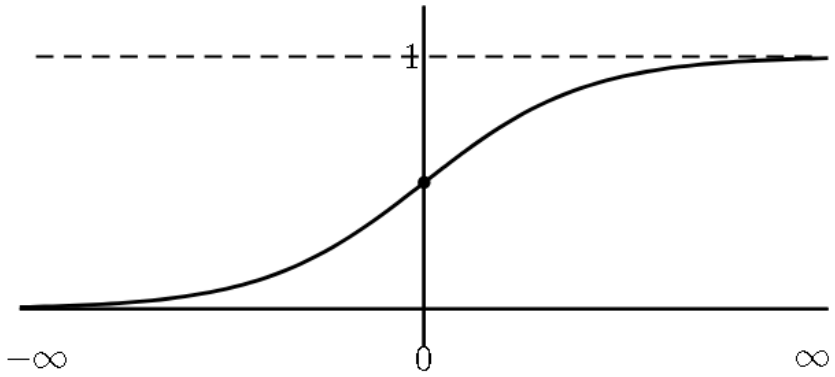

Figura 1: Tomada de: Moscote, O., & Rincón, W. (2012).

Algunas de las características para tener en cuenta es el que modelo transformado es lineal en las variables pero que las probabilidades no son lineales, de igual manera supone que el logaritmo de la razón de probabilidades esta linealmente relacionado con las variables explicativas y que los coeficientes de regresión expresan el cambio en el logaritmo de las probabilidades, cuando una de las variables explicativas cambia en una unidad, permaneciendo constantes las demás.

7.5 Algoritmo Stepwise

De acuerdo con Ferrero. (2017) el algoritmo Stepwise también conocida como regresión paso a paso consiste en la construcción paso a paso un modelo de regresión y puede operar de 3 maneras, la primera es usando y seleccionando las variables más

relevantes (también conocido como forward stepwise regression), la segunda manera es buscando y eliminando aquellas variables que resulten inútiles o redundantes (este proceso se conoce como backward stepwise regression), y el ultimo es el que utiliza los 2 métodos anteriormente mencionados (conocido simplemente como Stepwise regression).

El criterio de selección de las variables que utiliza el algoritmo puede variar, pero entre los principales son el criterio AIC, R^2, $R^2_{Ajustado}$ o el estadístico el C_p de Mallows. Para el desarrollo de este proyecto se utiliza el comando Step del programa R el cual utiliza el criterio AIC para seleccionar las mejores variables para el modelo.

8 Marco metodológico

8.1 Contextualización de la base

Para identificar cada uno de los variables empleados en las bases de datos de las plataformas SIMAT Y SIMPADE se le solicitó a la Secretaria de Educación de Zipaquirá apoyo en la aclaración de estos y producto de ello dieron a conocer que el SIMPADE fue desactivado y que la información sensible acerca de los estudiantes se limitó bastante y la perfilación así como las cifras exactas de estos quedo a cargo del Ministerio de Educación Nacional, debido a esto la información que hay disponible es muy limitada, y esta información fue extraída desde el SIMAT con información desde el año 2016 hasta el 2022 lo cual permitió una visualización más completa para el estudio del comportamiento de los estudiantes antes y después de la pandemia, esta base posee 110106 registros, fue anonimizada para protección de los estudiantes, cuyas variables y características son las siguientes:

La siguiente es la única variable numérica:

- **Edad:** Edad del estudiante durante ese año lectivo.

Y de las siguientes variables categóricas:

- **Año:** Año en el que el estudiante se encuentra al momento del registro.
- **Estado:** Variable que define el estado del estudiante a final del año, puede ser:
  - **Reprobado:** Consiste en que el estudiante reprobó el año lectivo.
  - **Graduado:** Consiste en estudiantes de grado 11 que culminaron sus estudios.
  - **Trasladado:** Consiste en estudiantes que cambiaron de establecimiento educativo.
  - **Matriculado:** Consiste en estudiantes que culminaron con éxito un año lectivo escolar.
  - **Retirado:** Consiste en la variable que indica si un estudiante se retiró de una institución educativa.

- **Motivo:** De acuerdo con lo explicado por la secretaria de Educación consiste en una variable relacionada con la variable **Estado** en la cual si el estudiante está en la categoría retirado aquí se indica el porqué, sus categorías pueden ser:
    - **Transferencia:** El estudiante fue retirado de la institución educativa debido a que se trasladará a una nueva institución educativa.
    - **Cambio de ubicación:** El estudiante fue retirado de la institución educativa debido a que se trasladará a una nueva residencia en la ciudad y decidieron irse a una institución educativa más cercana.
    - **Cambio de ciudad:** Similar a lo explicado para cambio de ubicación, pero con la diferencia de que lo que ocurre es un traslado a otra ciudad o país.
    - **Deserción:** Consiste en que el estudiante en cuestión durante lo que quedo de año decidió abandonar sus estudios, esta última calificación es asignada por la secretaria después de confirmar con las instituciones educativas.
- **Institución:** Consiste en las 10 instituciones educativas oficiales que posee el municipio.
- **Zona Sede:** Consiste en una variable dicotómica que define si la sede del colegio en la que el estudiante esta es urbana o rural.
- **Jornada:** Consiste en la jornada a la que pertenece el estudiante, posee las categorías Mañana, Tarde y Única.
- **Grado:** Consiste en el grado al que pertenece el estudiante y esta desde grado 0 (Transición) hasta el grado 11.
- **Modelo:** Consiste en el tipo de modelo educativo al cual pertenece el estudiante, puede ser:
    - **Tradicional:** Consiste en modelo educativo habitual.
    - **Caminemos en secundaria:** Consiste en un programa de apoyo educativo enfocado en estudiantes que vivan en el sector rural.
    - **Escuela nueva:** Consiste en sedes escolares nuevas cuya categoría no ha sido determinada.
- **Estrato:** El estrato al que pertenece el estudiante puede ser desde estrato 1 hasta el estrato 6.
- **Tipo de documento:** Consiste en el tipo de documento que posee el estudiante, puede ser:
    - **Tarjeta de identidad:** Consiste en el registro para estudiantes de nacionalidad colombiana que sean menores de 18 años.
    - **Cedula de ciudadanía:** Consiste en el registro para estudiantes de nacionalidad colombiana que sean de 18 años en adelante.
    - **Certificado de cabildo:** Consiste en identificación de una persona perteneciente a un cabildo indígena.
    - **Visa:** Consiste en ciudadanos de otras nacionalidades que poseen un permiso de estadía por un tiempo limitado, pueden ser o no mayores de 18 años.

  - **Cédula de extranjería:** Consiste en ciudadanos de otras nacionalidades que han conseguido la residencia permanente en Colombia y son mayores de 18 años.
- **Género:** Variable dicotómica que define si el estudiante es hombre o mujer.
- **Discapacidad:** Consiste en si el estudiante posee algún tipo de discapacidad física o mental y de qué tipo.
- **Nacionalidad:** Consiste en la nacionalidad del estudiante.

8.2 Delimitación de la investigación

Aunque se intentó añadir la característica de la distancia del hogar del estudiante hasta su institución educativa, no fue posible trabajarla debido a que más de la mitad de los registros no poseen información acerca del barrio en donde viven. Otras dificultades que se presentaron fue la limitada información que posee la base de datos del SIMAT, el acceso a la información familiar del estudiante, el ingreso familiar, acceso a internet, estado civil de los padres y si cuenta con hermanos.

8.3 Procesamiento de la base

Dado el contexto de la base, lo objetivos del trabajo, así como las limitaciones de este, se decidió trabajar con la información tanto de los colegios públicos como privados en los niveles de preescolar (Grado 0 o Transición), básica (1 a 9 grado) y media (10 y 11 grado) entre las edades de 5 a 19 años respectivamente, dado que después de esas edades los estudiantes son transferidos del programa de educación tradicional a programas de aprendizaje para adulto y en extra-edad.

De acuerdo con la metodología utilizada por MEN. (2022) y por Zapata, D. (2021), en ésta ellos explican que para la base del SIMAT se construye la variable **deserción** a través del uso de la variable **estado** y se dicotomiza de manera tal que 1 es si el estudiante está retirado y 0 para todo lo demás.

Por otro lado, en la investigación del Ministerio de Educación Nacional - MEN (MEN (2022)) se utiliza la variable de **estado** para definir la situación académica del estudiante y la cantidad de cursos pendientes.

Además, se debió construir una nueva variable denominada **Nivel educativo** a partir de la variable Grado_Cod, conformada por las categorías preescolar, básica y media de acuerdo con lo explicado por MEN. (2022).

Teniendo en cuenta los parámetros empleados por MEN. (2022) se adopta esta misma para construir una variable dicotómica que permita identificar si el estudiante reprobó en un año y se guarda en una variable aparte teniendo de esta forma una idea del comportamiento del estudiante durante el año lectivo.

Luego, de una manera similar a la realizada por Zapata, D. (2021), para el pre procesamiento de los datos se construye la variable de **años pendientes** a partir de la diferencia entre el curso actual y el grado 11 y otra variable para definir si el estudiante

está en extra-edad respecto a su grado, para esto se construirá una variable dicotómica en donde 1 que está en la edad que corresponde y 0 para el caso contrario.

8.4 Balanceo de la base, análisis descriptivo y selección de variables del modelo

Una vez construidas las variables adicionales necesarias para el conjunto de datos se procede con el análisis descriptivo de la base comenzando por la observación de la proporción de estudiantes desertores, allí se observa que hay un claro desbalance de nuestra variable respuesta (figura 2), por lo que, según lo explicado por Zapata, D. (2021) se recomienda el balanceo de la base.

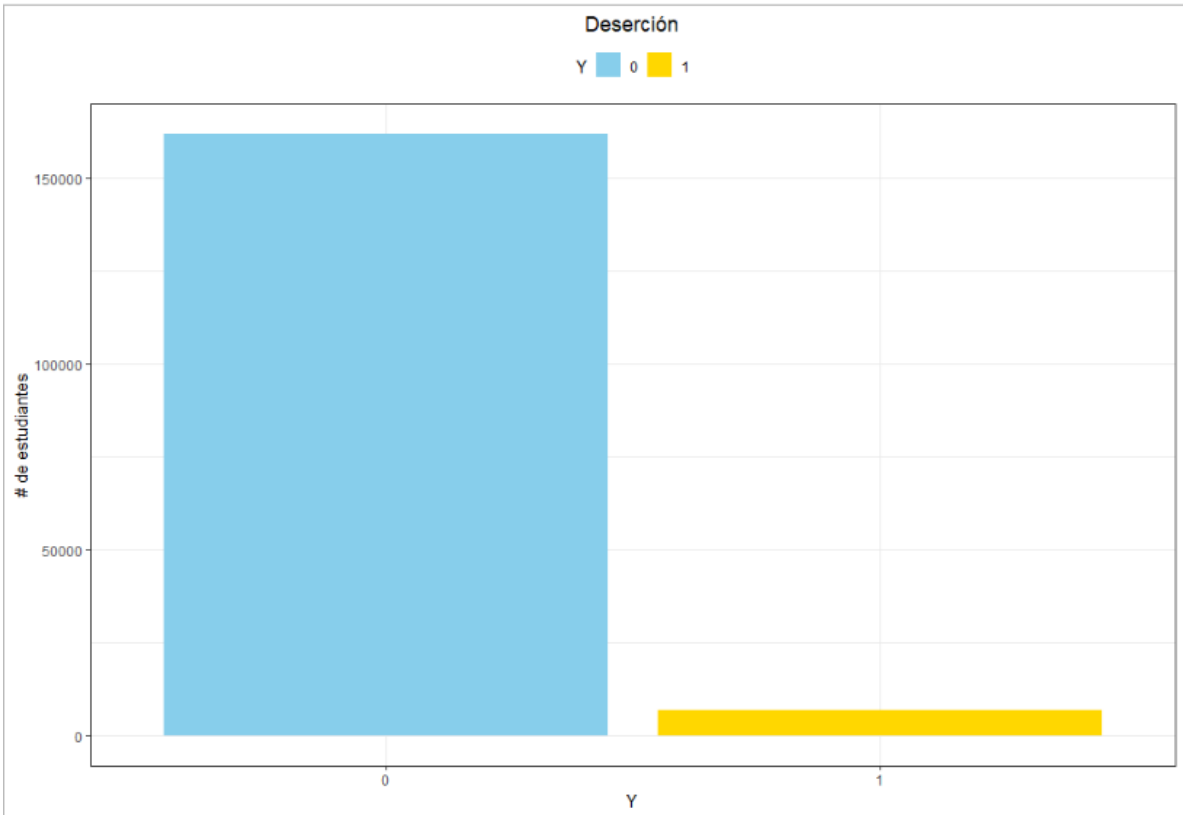

Figura 2: Fuente: Elaboración propia

Luego, se realizó una observación del comportamiento de sus 2 variables cuantitativas EDAD y Años_pendientes a través de un histograma (figura 3) la cual muestra la similitud en los comportamientos de edad y años pendientes.

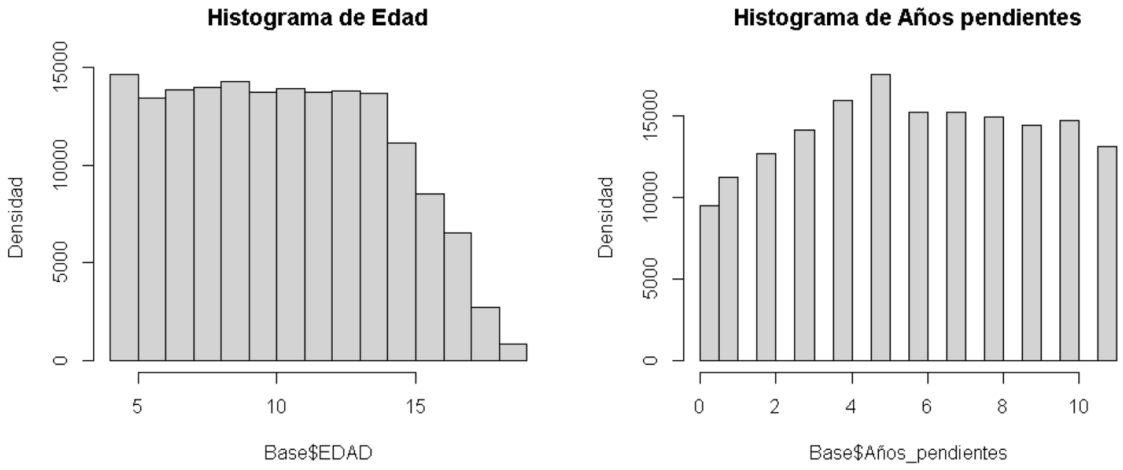

Figura 3: Fuente: Elaboración propia

Posteriormente se construyó una tabla que mostrara las cifras exactas del comportamiento de deserción por año, sector, sede y género.

				2016	2017	2018	2019	2020	2021	2022
0	NO OFICIAL	RURAL	FEMENINO	758	827	853	878	882	868	939
			MASCULINO	882	892	909	960	966	944	964
		URBANA	FEMENINO	3632	3582	3647	3615	3457	3161	3403
			MASCULINO	3350	3356	3393	3396	3314	2988	3231
	OFICIAL	RURAL	FEMENINO	1190	1194	1207	1244	1194	1168	1187
			MASCULINO	1265	1260	1289	1294	1179	1124	1219
		URBANA	FEMENINO	6304	6261	6087	6095	5950	5854	6163
			MASCULINO	6180	6096	6002	6050	5774	5813	6168
1	NO OFICIAL	RURAL	FEMENINO	0	0	2	8	14	40	35
			MASCULINO	0	1	2	13	29	42	31
		URBANA	FEMENINO	6	4	2	61	210	227	140
			MASCULINO	7	7	12	39	198	229	191
	OFICIAL	RURAL	FEMENINO	7	29	28	29	118	127	84
			MASCULINO	11	38	45	55	162	172	98
		URBANA	FEMENINO	130	90	156	160	411	547	465
			MASCULINO	184	98	236	188	520	635	627

Tabla 1: Fuente: Elaboración propia

Luego de realizada el análisis exploratorio de la base se procedió con el balanceo de la base, el cual se realizará a través de la librería ROSE del software R, y para la comparativa de la investigación se realizará una comparativa a través del método "Both" y "Under" y con un tamaño de muestra de 14.000 para de esta manera seleccionar la base más adecuada para la construcción de los modelos a comparar para selección.

Para el método "Under" se llegó a observar una proporción equitativa de 50% estudiantes desertores y 50% estudiantes que culminaron el año lectivo, mientras que

para el método “Both” la proporción quedo 49.8% estudiantes desertores y 50.2% estudiantes que culminaron el año lectivo. Posteriormente a las 2 bases se procede a construirles sus modelos correspondientes con todas las variables posibles que no resulten redundantes o innecesarias, las cuales son fueron:

- ZONA_SEDE
- JORNADA
- MODELO
- ESTRATO
- GENERO
- SECTOR
- EDAD
- REP_AÑO
- DISCAPACIDAD
- PAIS_ORIGEN
- Nivel_Educativo
- Años_pendientes

Dadas las características de la base y de acuerdo con Zapata, D. (2021), MEN. (2022), Manzano. (2011), Rodríguez, D. (2021) y Retavizca, M. (2016), el modelo más adecuado sería el modelo Logit debido a que la principal característica de nuestra variable respuesta es que es una variable dicotómica, además de la ventaja que algunas de las investigaciones antes mencionadas usan precisamente bases otorgadas por la plataforma SIMAT lo cual, permite tener una idea más clara del cómo operar con esta información.

De acuerdo con ello para la construcción del modelo, primero se divide la base en una de entrenamiento que empleará el 80% de la base balanceada y otra de prueba que usará el 20% restante.

8.4.1 Caso “Under”

Lo primero que se realizó fue la construcción de un modelo que tuviera todas las variables posibles, los resultados fueron los siguientes:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	2.103e+00	3.956e+03	0.001	0.99958	
ZONA_SEDEURBANA	7.036e-02	5.896e-02	1.193	0.23273	
JORNADATARDE	2.833e-01	2.759e-01	1.027	0.30448	
JORNADAÚNICA	3.133e-01	5.933e-02	5.280	1.29e-07	***
MODELOEDUCACIÓN TRADICIONAL	-7.937e-01	3.713e-01	-2.137	0.03256	*
MODELOESCUELA NUEVA	-1.462e+00	5.194e-01	-2.814	0.00489	**
ESTRATOESTRATO 1	2.028e-01	1.682e-01	1.206	0.22780	
ESTRATOESTRATO 2	8.763e-02	1.664e-01	0.527	0.59842	
ESTRATOESTRATO 3	6.997e-03	1.731e-01	0.040	0.96775	
ESTRATOESTRATO 4	3.292e-01	2.254e-01	1.460	0.14417	
ESTRATOESTRATO 5	8.326e-01	5.705e-01	1.459	0.14445	
ESTRATOESTRATO 6	1.695e+00	1.288e+00	1.316	0.18817	
GENEROMASCULINO	5.315e-02	4.539e-02	1.171	0.24162	
SECTOROFICIAL	1.054e-01	5.615e-02	1.877	0.06058	.
EDAD	9.581e-01	2.480e-02	38.635	< 2e-16	***
REP_AÑO1	-1.879e+01	1.469e+02	-0.128	0.89825	
DISCAPACIDADDISCAPACIDAD FÍSICA	-1.216e+00	1.262e+00	-0.963	0.33539	
DISCAPACIDADDISCAPACIDAD INTELECTUAL	1.782e-01	1.029e+00	0.173	0.86247	
DISCAPACIDADDISCAPACIDAD MÚLTIPLE	6.514e-01	1.096e+00	0.594	0.55221	
DISCAPACIDADDISCAPACIDAD PSICOSOCIAL (MENTAL)	1.290e+00	1.084e+00	1.191	0.23367	
DISCAPACIDADDISCAPACIDAD VISUAL BAJA VISIÓN IRREVERSIBLE	1.833e+01	3.956e+03	0.005	0.99630	
DISCAPACIDADDISCAPACIDAD VISUAL CEGUERA	1.885e+00	1.604e+00	1.175	0.23999	
DISCAPACIDADNO APLICA	1.103e+00	1.003e+00	1.100	0.27135	
DISCAPACIDADOTRA DISCAPACIDAD	-5.928e-01	1.467e+00	-0.404	0.68613	
DISCAPACIDADSISTEMICA	1.080e+00	1.904e+00	0.567	0.57063	
DISCAPACIDADTRANSTORNO PERMANENTE DE VOZ Y HABLA	8.677e-01	1.663e+00	0.522	0.60181	
DISCAPACIDADTRASTORNO DEL ESPECTRO AUTISTA	1.944e+00	1.554e+00	1.250	0.21115	
PAIS_ORIGENCOLOMBIA	-1.922e+01	3.956e+03	-0.005	0.99612	
PAIS_ORIGENECUADOR	-7.547e-01	4.845e+03	0.000	0.99988	
PAIS_ORIGENNO ESPECIFICADO	-1.890e+01	3.956e+03	-0.005	0.99619	
PAIS_ORIGENVENEZUELA	-1.748e+01	3.956e+03	-0.004	0.99647	
Nivel_EducativoMedia	-2.259e-01	9.325e-02	-2.423	0.01541	*
Nivel_EducativoPreescolar	3.788e-01	8.957e-02	4.229	2.35e-05	***
Años_pendientes	1.031e+00	2.921e-02	35.284	< 2e-16	***

Figura 4: Fuente: Elaboración propia

Dados los resultados de este modelo inicial se determinó que de acuerdo al p-value del modelo, así como a través del algoritmo Stepwise las variables Estrato, Discapacidad, Rep_Año, Genero y Zona son irrelevantes para determinar si un estudiante decide abandonar sus estudios o no por lo que se procedió a construir el modelo con las variables más relevantes y los resultados fueron los siguientes:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-13.54233	0.51827	-26.130	< 2e-16	***
JORNADATARDE	0.53415	0.25830	2.068	0.038646	*
JORNADAÚNICA	0.33176	0.05454	6.083	1.18e-09	***
SECTOROFICIAL	0.17771	0.04837	3.674	0.000239	***
EDAD	0.80275	0.02150	37.329	< 2e-16	***
Años_pendientes	0.89420	0.02585	34.589	< 2e-16	***
Nivel_EducativoMedia	-0.01077	0.08745	-0.123	0.902018	
Nivel_EducativoPreescolar	0.42203	0.08552	4.935	8.03e-07	***
MODELOEDUCACIÓN TRADICIONAL	-0.76526	0.34028	-2.249	0.024517	*
MODELOESCUELA NUEVA	-1.58388	0.47806	-3.313	0.000923	***

Figura 5: Fuente: Elaboración propia

Al tener la mayoría de variables con un p-value menor a 0.05 se procede a observar los resultados de la evaluación del modelo, los cuales son los siguientes:

```
               Accuracy : 0.4988
                 95% CI : (0.4895, 0.5081)
    No Information Rate : 0.5046
    P-Value [Acc > NIR] : 0.8886

                  Kappa : -0.0034

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.5580
            Specificity : 0.4386
         Pos Pred Value : 0.5030
         Neg Pred Value : 0.4935
              Precision : 0.5030
                 Recall : 0.5580
                     F1 : 0.5291
             Prevalence : 0.5046
         Detection Rate : 0.2815
   Detection Prevalence : 0.5596
      Balanced Accuracy : 0.4983

       'Positive' Class : 0
```

Figura 6: Fuente: Elaboración propia

En el resultado inicial se observa un nivel regular en la clasificación y asignación de un posible perfil de deserción estudiantil, esto puede llegar a justificarse hasta cierto punto dado que se está operando con la base de entrenamiento. Aun así, el hecho de que el resultado sea de un nivel así de regular deja mucho a desear.

8.4.2 Caso “Both”

Para el método “Both” al igual a como se realizó en “Under” lo primero que se realizó fue un modelo inicial que usara todas las variables y cuyos resultados se muestran a continuación:

```
Coefficients:
                                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                                       -14.53663 4268.36433  -0.003   0.9973
ZONA_SEDEURBANA                                     0.10810    0.06078   1.778   0.0753 .
JORNADATARDE                                        0.42276    0.25338   1.668   0.0952 .
JORNADAÚNICA                                        0.28988    0.05905   4.909 9.14e-07 ***
MODELOEDUCACIÓN TRADICIONAL                        -0.20094    0.34118  -0.589   0.5559
MODELOESCUELA NUEVA                                -1.00910    0.52141  -1.935   0.0530 .
ESTRATOESTRATO 1                                    0.22500    0.16731   1.345   0.1787
ESTRATOESTRATO 2                                   -0.01013    0.16532  -0.061   0.9511
ESTRATOESTRATO 3                                   -0.10189    0.17238  -0.591   0.5545
ESTRATOESTRATO 4                                    0.10216    0.21813   0.468   0.6395
ESTRATOESTRATO 5                                   -0.21213    0.49205  -0.431   0.6664
ESTRATOESTRATO 6                                    0.21471    1.43051   0.150   0.8807
GENEROMASCULINO                                     0.07697    0.04582   1.680   0.0930 .
SECTOROFICIAL                                       0.00375    0.05687   0.066   0.9474
EDAD                                                0.99311    0.02552  38.909  < 2e-16 ***
REP_AÑO1                                          -19.05702  147.36195  -0.129   0.8971
DISCAPACIDADDISCAPACIDAD FÍSICA                    14.49840 1602.37172   0.009   0.9928
DISCAPACIDADDISCAPACIDAD INTELECTUAL               15.78239 1602.37150   0.010   0.9921
DISCAPACIDADDISCAPACIDAD MÚLTIPLE                  15.42015 1602.37156   0.010   0.9923
DISCAPACIDADDISCAPACIDAD PSICOSOCIAL (MENTAL)      16.52443 1602.37154   0.010   0.9918
DISCAPACIDADDISCAPACIDAD VISUAL BAJA VISIÓN IRREVERSIBLE 32.83162 4268.36704   0.008   0.9939
DISCAPACIDADDISCAPACIDAD VISUAL CEGUERA            18.32450 1602.37186   0.011   0.9909
DISCAPACIDADNO APLICA                              16.65191 1602.37149   0.010   0.9917
DISCAPACIDADOTRA DISCAPACIDAD                      -2.88945 2288.41494  -0.001   0.9990
DISCAPACIDADSÍNDROME DE DOWN                        1.59517 2714.37312   0.001   0.9995
DISCAPACIDADSISTEMICA                              14.39192 1602.37333   0.009   0.9928
DISCAPACIDADTRANSTORNO PERMANENTE DE VOZ Y HABLA   16.33501 1602.37203   0.010   0.9919
DISCAPACIDADTRASTORNO DEL ESPECTRO AUTISTA         15.92454 1602.37169   0.010   0.9921
PAIS_ORIGENCOLOMBIA                               -19.23006 3956.17736  -0.005   0.9961
PAIS_ORIGENECUADOR                                 -0.52830 4270.29504   0.000   0.9999
PAIS_ORIGENNO ESPECIFICADO                        -18.95842 3956.17737  -0.005   0.9962
PAIS_ORIGENVENEZUELA                              -17.25958 3956.17736  -0.004   0.9965
Nivel_EducativoMedia                               -0.15107    0.09305  -1.624   0.1044
Nivel_EducativoPreescolar                           0.42637    0.08854   4.816 1.47e-06 ***
Años_pendientes                                     1.06777    0.02996  35.639  < 2e-16 ***
---
```

Figura 7: Fuente: Elaboración propia

Dado los resultados del modelo inicial se llega a observar una notable cantidad de variables innecesarias, por lo que se procede a usar el algoritmo Stepwise, así como la selección de ciertas variables de relevancia obligatoria a través del p-value, el modelo final queda con las siguientes variables:

```
Call:
glm(formula = Y ~ JORNADA + SECTOR + EDAD + Años_pendientes +
    Nivel_Educativo + MODELO, family = "binomial", data = train_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.9239  -0.9413  -0.4751   0.9907   3.6109

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                -14.58179    0.50718 -28.750  < 2e-16 ***
JORNADATARDE                 0.54521    0.23821   2.289  0.02209 *
JORNADAÚNICA                 0.26320    0.05381   4.891 1.00e-06 ***
SECTOROFICIAL                0.15435    0.04888   3.158  0.00159 **
EDAD                         0.83386    0.02201  37.889  < 2e-16 ***
Años_pendientes              0.93002    0.02636  35.283  < 2e-16 ***
Nivel_EducativoMedia         0.05579    0.08670   0.643  0.51995
Nivel_EducativoPreescolar    0.46411    0.08446   5.495 3.91e-08 ***
MODELOEDUCACIÓN TRADICIONAL -0.24292    0.30872  -0.787  0.43137
MODELOESCUELA NUEVA         -1.31671    0.48446  -2.718  0.00657 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 15526  on 11199  degrees of freedom
Residual deviance: 13031  on 11190  degrees of freedom
AIC: 13051
```

Figura 8: Fuente: Elaboración propia

Las variables con las que más conviene quedarse son las mismas que con el método under por lo que se procede con el estudio de los resultados del modelo con el uso de este método de balanceo.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1072  505
         1  355  868

               Accuracy : 0.6929
                 95% CI : (0.6754, 0.7099)
    No Information Rate : 0.5096
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3842

 Mcnemar's Test P-Value : 3.757e-07

            Sensitivity : 0.7512
            Specificity : 0.6322
         Pos Pred Value : 0.6798
         Neg Pred Value : 0.7097
              Precision : 0.6798
                 Recall : 0.7512
                     F1 : 0.7137
             Prevalence : 0.5096
         Detection Rate : 0.3829
   Detection Prevalence : 0.5632
      Balanced Accuracy : 0.6917

       'Positive' Class : 0
```

Figura 9: Fuente: Elaboración propia

Los resultados muestran un rendimiento superior al presentado bajo la metodología "Under", con una mejor precisión, así como clasificación del perfil del estudiante.

9 Selección del modelo y resultados

Para la comparación del modelo se construyó una tabla que compare los resultados del método "Under" y "Both" usando tanto la base de prueba, como la de entrenamiento.

	Train data				
Medida de balanceo	**Accuracy**	**F1**	**Recall**	**Precision**	**ROC performance**
Both	0.6924	0.7062	0.7326	0.6816	0.7656217
Under	0.4988	0.5291	0.5580	0.5030	0.7634699

Tabla 3: Fuente: Elaboración propia

Los resultados muestran que, en el caso de los datos de entrenamiento el uso del método de balanceo "Both" se posee una mayor certeza de clasificación de estudiantes desertores, así como de su certeza de detectar futuros casos.

	Test data				
Medida de balanceo	**Accuracy**	**F1**	**Recall**	**Precision**	**ROC performance**
Both	0.7104	0.7077	0.7116	0.7039	0.7754174
Under	0.7046	0.6943	0.6987	0.6899	0.7725792

Tabla 3: Fuente: Elaboración propia

Para el uso de la base de pruebas, aunque para el método "under" muestra una clara mejoría en sus resultados, sigue dando resultados inferiores a los presentados por el método "Both".

Por lo que en definitiva el mejor método para la construcción y selección de una población adecuada para el modelo, es el método "Both"

Dado lo anterior se pudo determinar lo siguiente:

Las variables país de origen, discapacidad y repitencia no son relevantes para determinar la deserción, sin embargo, para mejorar la selección de variables se decidió utilizar el algoritmo Stepwise el cual se basa en la valoración de su AIC para determinar las variables más importantes del modelo.

De acuerdo con los resultados mostrados por el modelo podemos observar que las variables que más influyen en el momento de deserción escolar son:

- Con relación al colegio: si este es oficial o privado y el modelo educativo adoptado por este.
- Con relación al estudiante: edad, si es repitente o no, su estrato socioeconómico, su nivel educativo y la jornada a la que pertenece.

Lo anterior, indica que hay una mayor tendencia de la deserción escolar en jóvenes que se encuentran en las etapas iniciales de su educación y cuya situación económica es vulnerable. Otro factor que influye significativamente es el tipo de jornada y el tipo de modelo educativo, lo cual explicaría la correlación que posee estas 2 variables.

Desde la perspectiva del autor y de lo dicho por Moreno, D. (2013) y MEN (2022) al presentarse la deserción escolar tanto en colegios oficiales como privados se puede concluir que estos no están ofreciendo las condiciones necesarias para retener a los estudiantes quienes por razones económicas, violencia intrafamiliar, bullying, distancia del colegio a la casa, desmotivación con el proceso de aprendizaje ya sea por temáticas o trato por parte de los docentes y directivos , entre otros factores; hacen que prefieran no terminar el ciclo escolar perdiendo así la posibilidad de mejorar sus condiciones sociales, culturales y económicas debido a que su baja escolaridad no le permitirá acceder a trabajos con una remuneración alta.

De acuerdo con lo mostrado en los resultados del modelo Logit, los modelos pedagógicos tradicionales y el de Escuela Nueva pueden llegar a influir negativamente en la toma de decisión sobre continuar o desertar de los estudios lo cual se puede atribuir según lo explicado por Sanabria, E. (2014) y por Moreno, D. (2013) a que este tipo de metodologías no proporcionan los elementos suficientes que permitan retener a los estudiantes en las instituciones.

En cuanto a la edad la deserción escolar es frecuente entre los niños de 5 a los 15 años, los cuales en encuetran entre transición a grado 10°. Estos niños están clasificados en su gran mayoría en los estratos socioeconómicos del 1 al 4, que corresponden a un nivel socioeconómico medio-bajo y bajo.

Los estudiantes repitentes son quienes presentan las tasas de deserción más altas, así como los pertenecientes a la jornada única y tarde. Lo anterior, puede estar asociado según lo dicho por MEN (2022), Sanabria, E. (2014) y por Moreno, D. (2013) a que no se está garantizando aspectos básicos como una adecuada alimentación, transporte escolar, contenidos pedagógicos acorde a las necesidades sociales y labores de los estudiantes.

10 Conclusiones

Teniendo en cuenta los resultados anteriores se puede concluir:

- Los factores que más influyen en la deserción escolar tanto en colegios oficiales como en privados en el Municipio de Zipaquirá son: edad, modelo, nivel educativo, jornada y clase de institución (Oficial o privada).
- Los mayores índices de deserción escolar se encuentran en los niveles de preescolar y media.
- La deserción escolar es más alta en los colegios oficiales que en los privados, debido a que es posible que no estén ofreciendo las condiciones necesarias para retener a los estudiantes.

- Los modelos pedagógicos con mayor posibilidad de deserción escolar son el Tradicional y el de Escuela Nueva.
- Los estudiantes pertenecientes a la jornada única y a la de la tarde son quienes presentan las tasas de deserción más altas.
- El mejor método para la optimización de la base y la selección de varibles es el método “Both”.
- ¿Qué políticas públicas educativas debería adoptar el municipio de Zipaquirá para disminuir el riesgo de deserción de los estudiantes de 5-19 años además de los programas de Alimentación y transporte escolar que ofrecen?

11 Recomendaciones

- Los establecimientos educativos tanto oficiales como privados deben diligenciar la totalidad de los campos solicitados por el SIMAT.
- Con base en los resultados de la presente investigación, formular estrategias para el apoyo a las poblaciones vulnerables a nivel socioeconómico con el fin de generar retención escolar tanto en colegios oficiales como en privados.

12 Referencias

- Ministerio de Educación Nacional. (2022). Deserción escolar en Colombia: análisis, determinantes y política de acogida, bienestar y permanencia: nota técnica.
- Manzano, D. (2011). Interpelación entre la deserción escolar y las condiciones socioeconómicas de las familias: el caso de la ciudad de Cúcuta (Colombia).
- Drysdale, R. (1972), Factores determinantes de la deserción escolar en Colombia. Estudio de un caso de escolaridad rural primaria.
- Montoya, N. (2019), Identificación de las Posibles Causas de Deserción Escolar en los Jóvenes y Niños del Colegio Departamental General Santander Sede San Benito de Sibaté, Universidad Cooperativa De Colombia.
- Ventura, J. (2021), Factores asociados a las causas de la deserción estudiantil en instituciones de educación superior de El Salvador.
- Ocampo, L. (2019), Modelo de sobrevida aplicado a deserción estudiantil de las facultades de Estadística y Diseño gráfico de la Universidad Santo Tomás.
- Suárez, Lilian et al., (2021), Técnicas estadísticas y logro de aprendizaje: revisión bibliográfica.
- Rodríguez, D. (2021), Análisis de la deserción escolar por localidades en Bogotá.
- Palacios, I. (2021), Factores determinantes en la deserción escolar de los estudiantes de la Institución Educativa INELAG del municipio de El Retorno, Guaviare.
- Hurtado, M. (2020), FACTORES ASOCIADOS A LA DESERCIÓN ESCOLAR EN BÁSICA PRIMARIA DE LA INSTITUCION EDUCATIVA SAN ISIDRO I Y SAN ISIDRO II.

- Zapata, D. (2021), Método para la Detección de Estudiantes en Riesgo de Deserción, Basado en un Diseño de Métricas y una Técnica de Minería de Datos.
- Jurado, R. (2012), CONDICIONES SOCIOCULTURALES DE LA FAMILIA SANTANDEREANA COMO ESCENARIO INFLUYENTE EN LA DESERCIÓN O PERMANENCIA ESCOLAR EN PREADOLESCENTES DE LA INSTITUCIÓN EDUCATIVA CAMPO HERMOSO DE BUCARAMANGA.
- Retavizca, M. (2016), DETERMINANTES DE LA DESERCIÓN ESCOLAR ASOCIADO AL TRABAJO INFANTIL EN EL RANGO DE EDAD ENTRE LOS 5 Y LOS 17 AÑOS EN COLOMBIA PARA EL AÑO 2015.
- Ferrero, R. (2017), SELECCIÓN PASO A PASO E IMPORTANCIA DE LOS PREDICTORES.
- Moreno, D. (2013). "La Deserción Escolar: Un problema de Carácter Social". Revista In Vestigium Ire. Vol. 6, pp. 115-124.
- Sanabria, E. (2014), LA DESERCION ESCOLAR EN EL CONTEXTO RURAL COLOMBIANO. CASO GUATEQUE – BOYACÁ.