Modelación de la Deserción Estudiantil en Colombia: Un Análisis Basado en Indicadores Educativos

Ricardo William Sazalar
Santiago Vera Quiceno

Universidad Nacional Sede Medellin
Medellin

Resumen

La deserción escolar en Colombia varía significativamente entre departamentos y representa un desafío clave para el sistema educativo. Este estudio analiza sus determinantes a partir de datos del Ministerio de Educación Nacional (2011-2023), considerando factores como el presupuesto educativo, la tasa de aprobación y reprobación, y la clasificación geográfica (Andino vs. No Andino).

Para capturar la heterogeneidad territorial, se emplea un modelo Beta Jerárquico Bayesiano, permitiendo estimaciones más precisas. Los resultados revelan una mayor probabilidad de deserción en los departamentos no andinos y destacan factores críticos para orientar estrategias de intervención.

Este trabajo contribuye al diseño de políticas públicas enfocadas en la reducción de la deserción y la mejora de la continuidad educativa.

Palabras clave: deserción escolar, modelo Beta Jerárquico, educación en Colombia, análisis bayesiano.

Introducción

La deserción escolar es un desafío persistente en Colombia, con tasas que varían ampliamente entre regiones. Factores como la infraestructura educativa, la calidad docente y los recursos financieros influyen en este fenómeno, afectando la equidad y el acceso a la educación.

Este estudio examina los determinantes de la deserción a partir de datos del Ministerio de Educación Nacional (2011-2023), con especial atención al presupuesto educativo, las tasas de aprobación y reprobación, y la clasificación geográfica de los departamentos (Andino vs. No Andino).

Para abordar la variabilidad entre territorios, se emplea un modelo Beta Jerárquico Bayesiano, que permite generar estimaciones más robustas y precisas. Los hallazgos ofrecen información clave para la formulación de políticas educativas que promuevan la permanencia escolar y reduzcan las brechas territoriales en el acceso a la educación.

Objetivo

El objetivo de este estudio es modelar la deserción escolar en Colombia utilizando un enfoque bayesiano jerárquico, con el fin de determinar:

El impacto del presupuesto educativo en la deserción escolar.

Si la clasificación geográfica de los departamentos (Andino vs No Andino) influye en la deserción.

El efecto de las tasas de aprobación y reprobación en la permanencia escolar.

Este estudio contribuye al análisis de políticas públicas en educación, al proporcionar una estimación más robusta de los factores que afectan la deserción escolar en el país.

Metodología

Datos

Los datos utilizados en este estudio fueron obtenidos del Ministerio de Educación Nacional de Colombia, abarcando el período 2011-2023. Se recopilaron indicadores clave relacionados con la educación, como:

Tasa de deserción escolar (variable de interés). Tasa de aprobación y reprobación (indicadores de desempeño académico). Presupuesto educativo ajustado (factor económico clave).

Cálculo del Presupuesto Educativo Ajustado

El presupuesto educativo para cada departamento y año no estaba directamente disponible, por lo que fue calculado en varias etapas:

Asignación inicial: Se tomó el presupuesto total asignado por la Nación a educación en cada año.

Distribución por departamento: Se ponderó en función de la cantidad de estudiantes matriculados en cada departamento durante ese año.

Ajuste por inflación: Finalmente, el presupuesto fue ajustado al año 2023 utilizando el Índice de Precios al Consumidor (IPC), con el fin de hacer comparaciones consistentes a lo largo del tiempo.

Este ajuste permite analizar la evolución real del presupuesto en términos comparables, considerando el efecto de la inflación y las variaciones en la cantidad de estudiantes por departamento.

Modelo Estadístico

Para modelar la deserción escolar, se utilizó un modelo bayesiano jerárquico con distribución Beta, ya que la variable de interés es una proporción en el intervalo (0,1).

Los parámetros fueron estimados mediante el método de Monte Carlo vía Cadenas de Markov (MCMC), implementado en Stan.

El modelo jerárquico se construyó de la siguiente manera:

$$\mu_i = \mu_\alpha + \gamma_{\text{region}[i]} + \beta_1 \cdot \text{presupuesto}_i + \beta_2 \cdot \text{aprobación}_i^2 + \beta_3 \cdot \text{reprobación}_i \tag{1}$$

$$\text{deserción}_i \sim \text{Beta}(\text{inv\_logit}(\mu_i) \cdot \phi, (1 - \text{inv\_logit}(\mu_i)) \cdot \phi) \tag{2}$$

Donde:
- $\alpha_{\text{region}[i]}$ representa el **efecto base de la deserción** según la clasificación del departamento en *Andino* o *No Andino*.
- β_1, β_2, β_3 son los coeficientes asociados a las variables explicativas *presupuesto, reprobación y aprobación*, respectivamente.
- ϕ es un parámetro de dispersión que ajusta la varianza de la distribución Beta.

El modelo incorpora un **efecto jerárquico**, lo que permite capturar la variabilidad entre las regiones mediante un intercepto específico para cada grupo. Esto proporciona estimaciones más robustas y evita el sobreajuste.

Priorización de Parámetros

En un enfoque bayesiano, los parámetros del modelo requieren distribuciones previas (priors) que reflejen el conocimiento previo o asunciones razonables sobre los valores esperados antes de observar los datos. El uso de priors bien definidas ayuda a estabilizar las estimaciones y a reducir el sobreajuste del modelo.

Las priors utilizadas en este estudio fueron seleccionadas con base en principios de regularización y flexibilidad, permitiendo que el modelo capture patrones sin imponer restricciones excesivas.

Table 1: Distribuciones previas (*priors*) utilizadas en el modelo.

Parámetro	**Distribución**	**Descripción**
μ_α	$\mathcal{N}(0.1, 0.1)$	Media base de la deserción. Se asume que está cerca del 10%.
σ_α	Gamma$(2, 1)$	Controla la variabilidad entre regiones (Andino vs No Andino).
σ_{id}	Gamma$(2, 1)$	Modela la desviación estándar de los efectos aleatorios γ.
γ	$\mathcal{N}(0, \sigma_{id})$	Modela efectos aleatorios adicionales en el modelo como el año.
β	$\mathcal{N}(0, 0.5)$	Coeficientes para las covariables presupuesto, reprobación y aprobación.
ϕ	Gamma$(2, 0.5)$	Controla la dispersión de la distribución Beta.

Justificación del Modelo Jerárquico

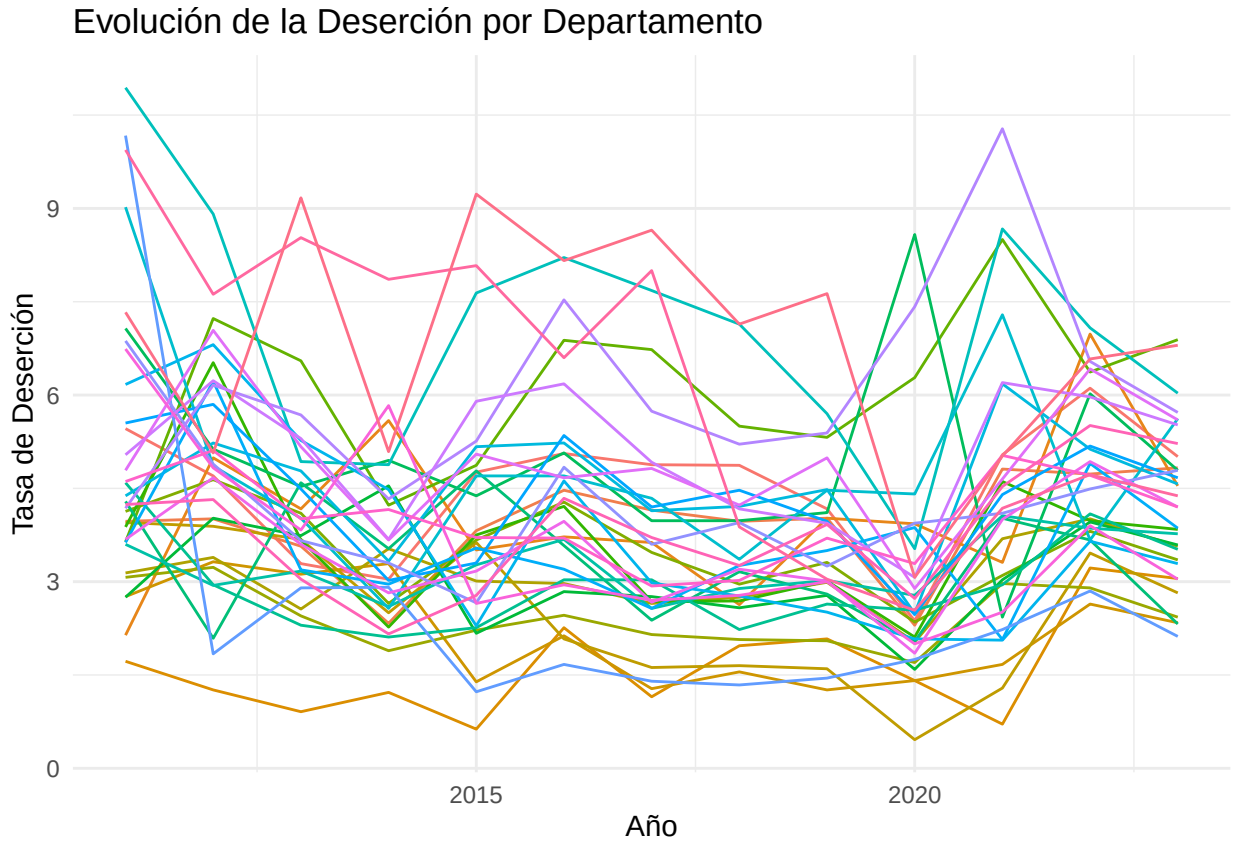
En el estudio de la deserción escolar, los datos utilizados no solo presentan variabilidad entre departamentos, sino que al agruparlos en regiones (Andino vs No Andino) y mantener registros para cada año (2011-2023), se genera una estructura de datos anidada. Esto implica que tenemos múltiples observaciones para cada región en distintos años, lo que motiva el uso de un modelo jerárquico con efectos aleatorios.

Análisis Descriptivo de los Datos

Primero se realizara un analisis de las variables utilizadas en este proyecto

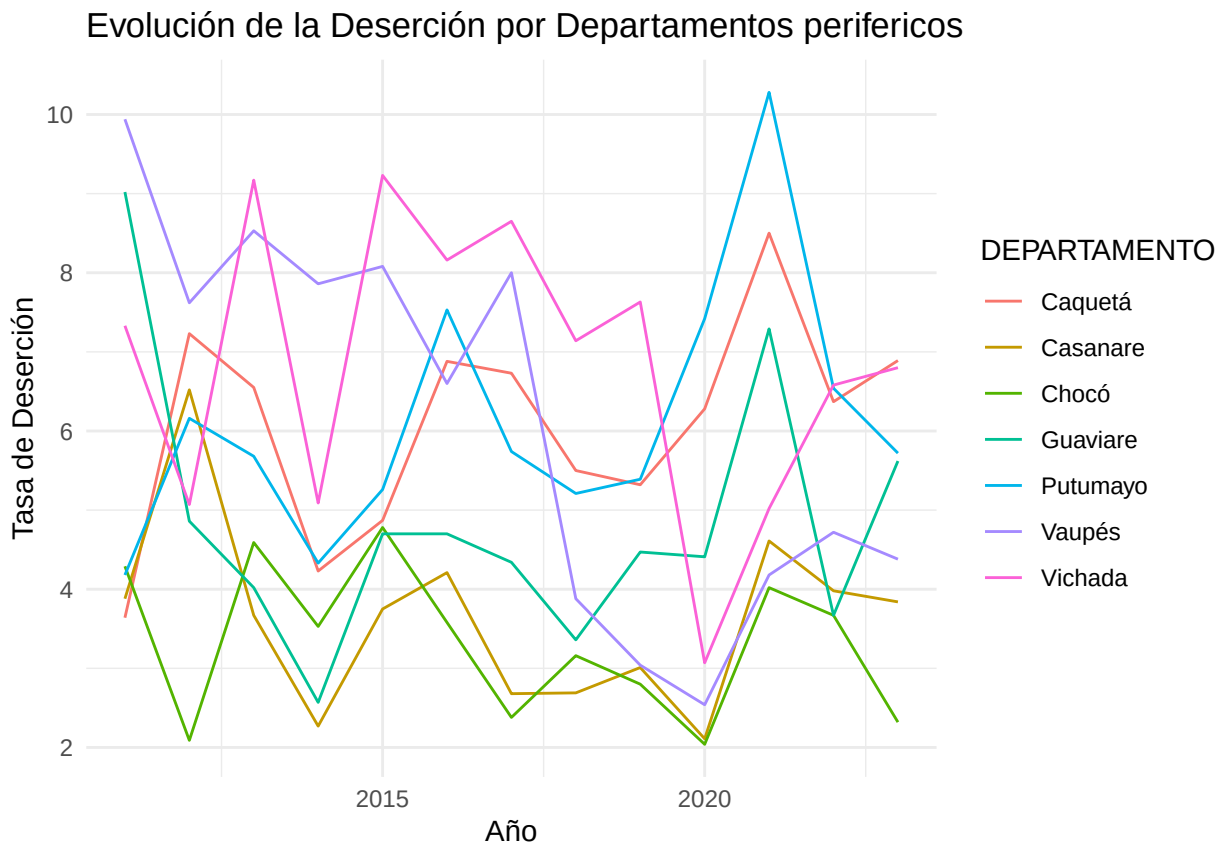
Desercion

Tenemos en el siguiente grafico la variacion de la desercion para cada departamento en el periodo de tiempo de 2011 a 2023, podemos notar que el grafico en cuestion es caotico, a un asi podemos notar que el intercepto de cada uno de los departamentos es aleatorio lo cual refuerza la idea de que un modelo jerarquico es posible para modelar el problema.

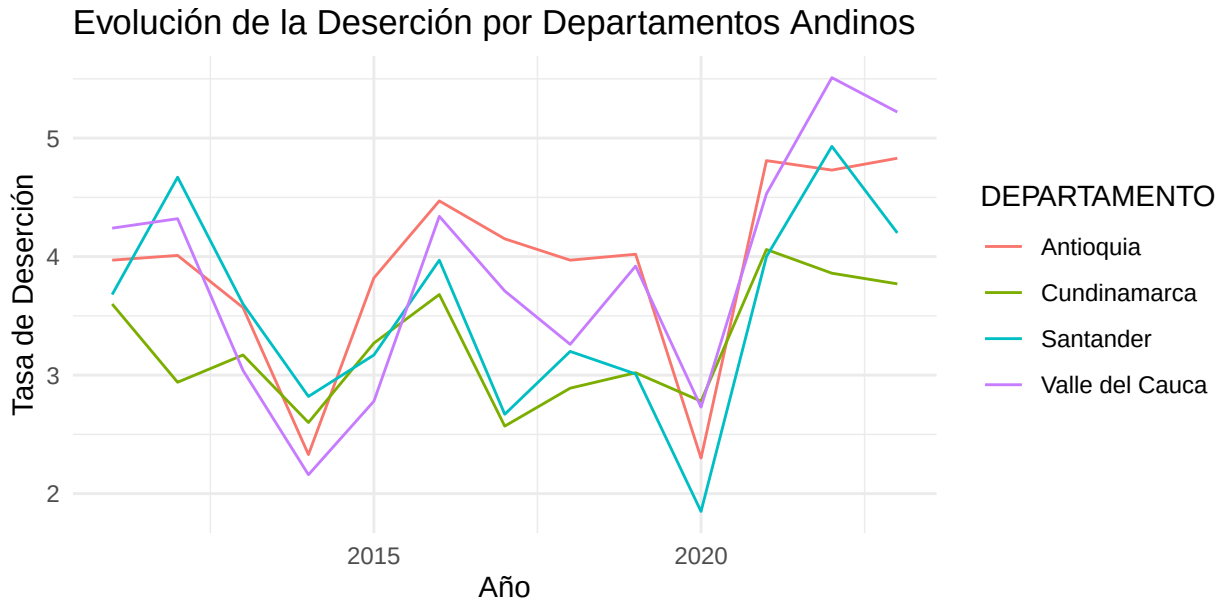


Posteriormente cuando segregamos por regiones obtenemos los siguientes graficos para algunos departamentos de interes.

Primero observamos como evoluciona la desercion en departamentos no andinos durante el tiempo, podemos notar que todos tienen periodos donde hay aumentos y donde decrece la desercion, podemos notar que los interceptos parecen ser aleaotorios y parece que se mantiene lo caoitico en la grafica pero se pueden ver algunas tendencias mas claras en algunos años.

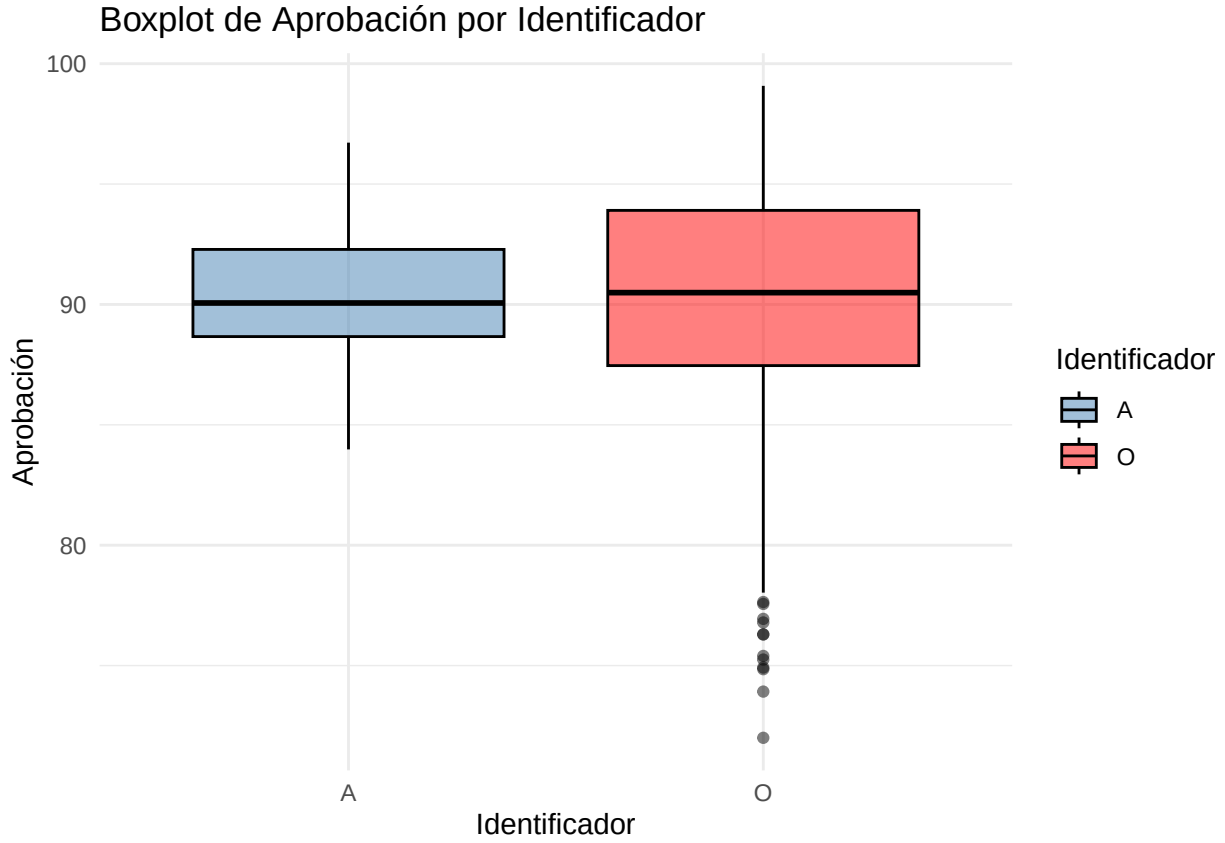


Ahora realizamos el mismo analisis pero para departamentos andinos lo primero es notar que estos tienen un comportamiento mas comun, en terminos generales la desercion baja en el año 2020 probablemente por pandemia(COVID-19), tambien una caida en el año 2014 probablemente por el cambio de gobierno y en terminos generales se observa un comportamiento similar entre departamentos andinos.

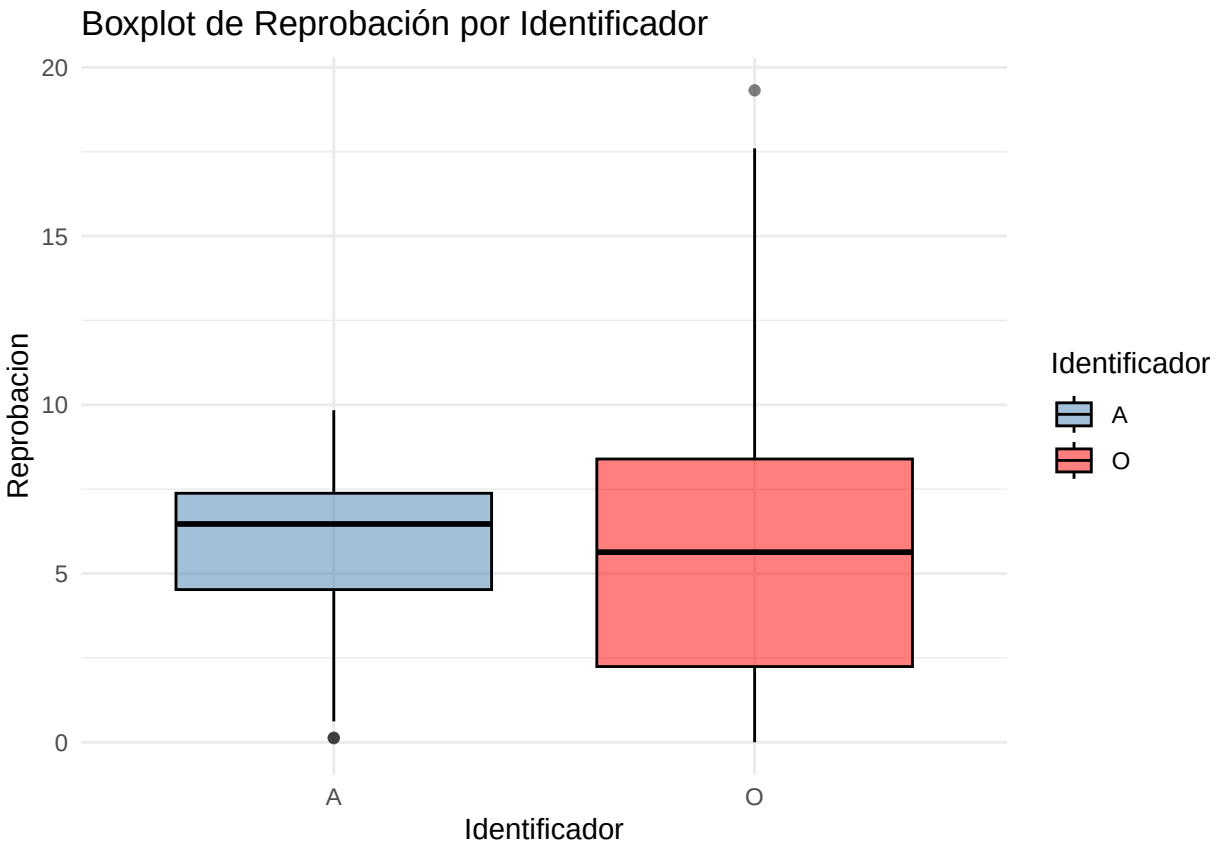


Aprobacion y Reprobacion

El boxplot para la vairiable *aprobacion* muestra la distribución de la variable según el Identificador (A: Andino, O: Periferia). Se observa que ambas categorías presentan medianas similares, indicando que los valores centrales de aprobación no varían significativamente entre los dos grupos. Sin embargo, la categoría "O" (Periferia) exhibe una mayor dispersión en los datos, con una presencia notable de valores atípicos por debajo del límite inferior, lo que sugiere una mayor variabilidad en los niveles de aprobación en esta región. En contraste, la categoría "A" (Andino) muestra una distribución más compacta, con un menor rango intercuartílico y menos valores extremos, lo que indica que la aprobación en esta zona es más homogénea. En general, aunque ambas categorías tienen niveles de aprobación relativamente altos, la periferia presenta más variabilidad y casos de aprobación baja que podrían requerir un análisis más detallado.



El siguiente boxplot muestra la distribución de la variable Reprobación según el Identificador (A: Andino, O: Periferia). Se observa que la mediana de la reprobación es ligeramente mayor en la categoría "O" (Periferia) en comparación con la categoría "A" (Andino), lo que indica que la reprobación tiende a ser más alta en las regiones periféricas. Además, la dispersión de los datos es mayor en la categoría "O", con un rango intercuartílico más amplio y valores atípicos más elevados, lo que sugiere una mayor variabilidad en los niveles de reprobación en esta región. En contraste, la categoría "A" presenta una distribución más compacta y con menos valores extremos, lo que indica una reprobación más homogénea. Este análisis sugiere que las regiones periféricas podrían enfrentar mayores desafíos en términos de desempeño, con casos en los que la reprobación es considerablemente alta en comparación con la región andina.

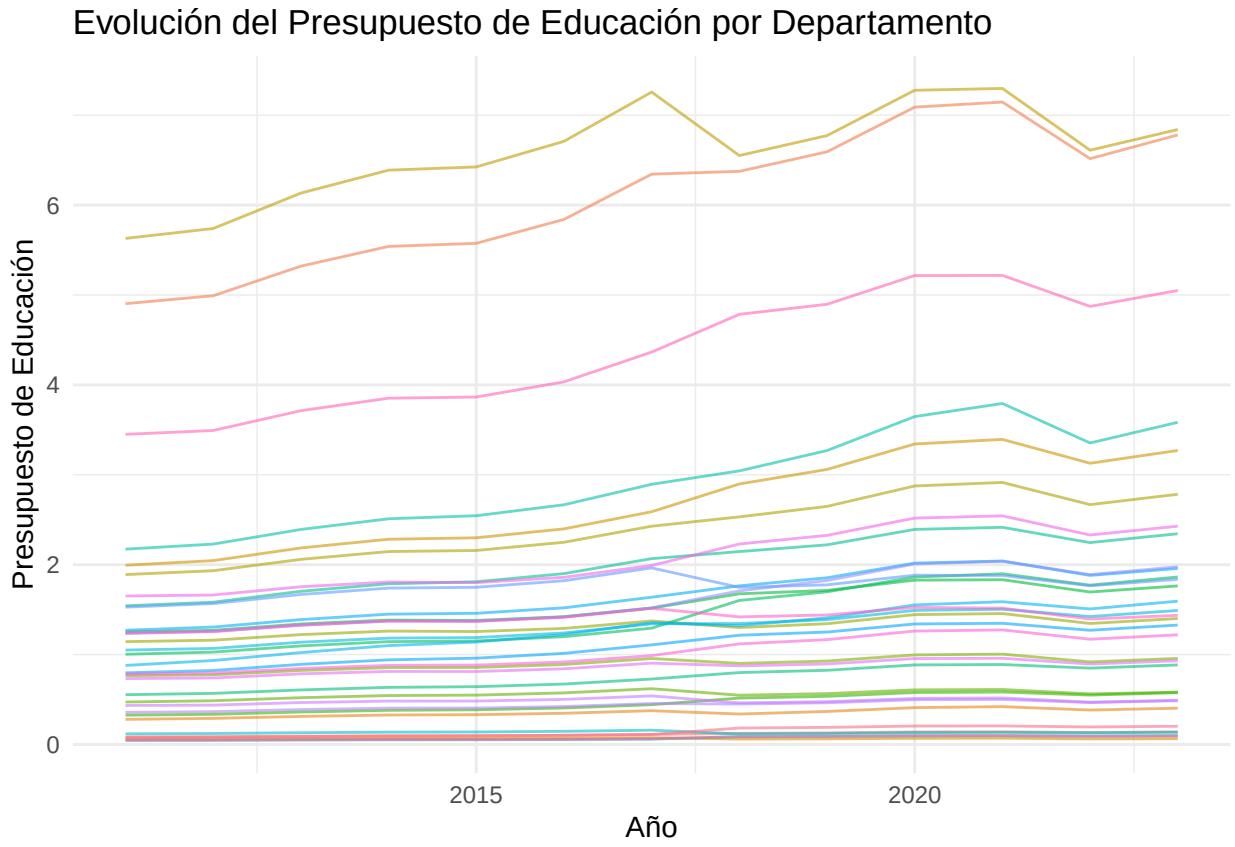


En conclucion en la categoría A (Andino), tanto la aprobación como la reprobación presentan distribuciones más compactas, con menor dispersión y menos valores atípicos. Esto sugiere que el desempeño en términos de aprobación y reprobación es más homogéneo en esta región, con menos casos extremos.

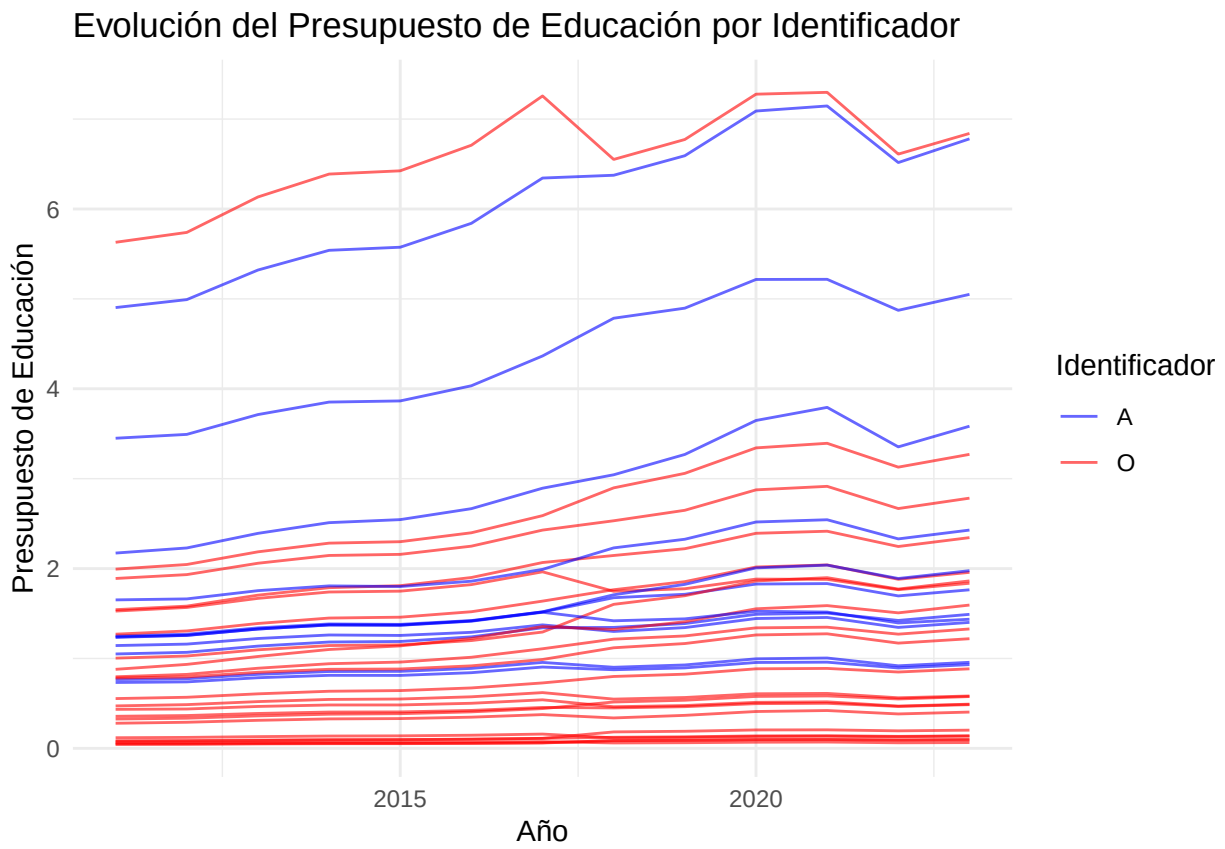
En la categoría O (Periferia), se observa una mayor dispersión en ambas métricas. Hay más valores atípicos en la aprobación (en la parte baja) y en la reprobación (en la parte alta), lo que sugiere que en esta región existen instituciones o grupos con un desempeño significativamente inferior a la media.

Presupuesto

Ahora analizaremos el presuesto de la nacion asignado a cada uno de los departamentos, este se puede observar en el siguiente grafico, podemos observar que el presupuesto tiene una tendencia creciente, pero parece que su pico mas alto fue en el año 2017, posteriormente este a decrecido y se a mantenido fluctuando desde entonces, es importante recordar que este presupuesto esta ajustado al IPC del años 2023.



Posteriormente realizaremos el analisis por regiones observando como es el comportamiento entre departamentos Andinos y no Andinos.



El gráfico muestra la evolución del presupuesto de educación por departamento a lo largo del tiempo, diferenciando entre las regiones Andinas y Periféricas. Se observa una tendencia general de aumento en la inversión educativa en la mayoría de los departamentos, aunque con fluctuaciones en ciertos periodos. Bogotá, D.C. destaca como el departamento con el presupuesto más alto, lo cual es consistente con su alta densidad poblacional y su rol como centro administrativo y económico del país. Asimismo, se evidencia que los departamentos Andinos tienden a tener una distribución presupuestaria más homogénea, mientras que en la Periferia hay una mayor variabilidad, con algunos departamentos recibiendo presupuestos considerablemente altos y otros con asignaciones más reducidas. Esta disparidad sugiere que la distribución de los recursos educativos podría no estar únicamente determinada por la ubicación geográfica, sino también por factores como la cantidad de estudiantes, la infraestructura educativa y las políticas gubernamentales de asignación de recursos.

Conclucion analisis exploratorio

Las gráficas muestran diferencias claras en la evolución de la deserción entre departamentos Andinos y Periféricos.

Estas diferencias justifican el uso del modelo Beta Jerárquico, donde α_i permitira capturar variaciones estructurales.

La alta variabilidad observada refuerza la importancia de modelar la deserción con una distribución Beta, que es flexible para datos de proporciones.

Validacion del modelo

Analisis de convergencia y estimacion de parametros

La siguiente tabla muestra el resumen posterior de los parámetros estimados, con estadísticas como la media , desviación estándar , cuartiles (5%, 50% y 95%), el número efectivo de muestras (n_eff) y el diagnóstico de convergencia (Rhat).

	Estadísticas							
Parameter	Mean	SEMean	SD	P5	P50	P95	Neff	Rhat
μ_α	0.09	0.00	0.10	-0.07	0.09	0.25	3922	1
σ_α	2.00	0.01	1.37	0.37	1.71	4.65	9750	1
β_1	-0.12	0.00	0.03	-0.16	-0.12	-0.08	8243	1
β_2	0.04	0.00	0.02	0.02	0.04	0.06	7044	1
β_3	0.08	0.00	0.02	0.04	0.08	0.12	6968	1
ϕ	114.44	0.09	7.89	101.66	114.28	127.65	7986	1
σ_{id}	3.10	0.01	1.09	1.74	2.89	5.18	6559	1
γ_1	-3.24	0.00	0.11	-3.41	-3.24	-3.06	4015	1
γ_2	-3.30	0.00	0.10	-3.47	-3.30	-3.13	4117	1

El modelo presentado corresponde a una regresión Beta jerárquica, lo que implica que la variable de interés sigue una distribución Beta y que se incorporan niveles jerárquicos para capturar la variabilidad entre grupos. En este contexto, los coeficientes estimados nos permiten entender cómo influyen el presupuesto, la aprobación y la reprobación en la respuesta, así como las diferencias sistemáticas entre regiones Andinas y No Andinas.

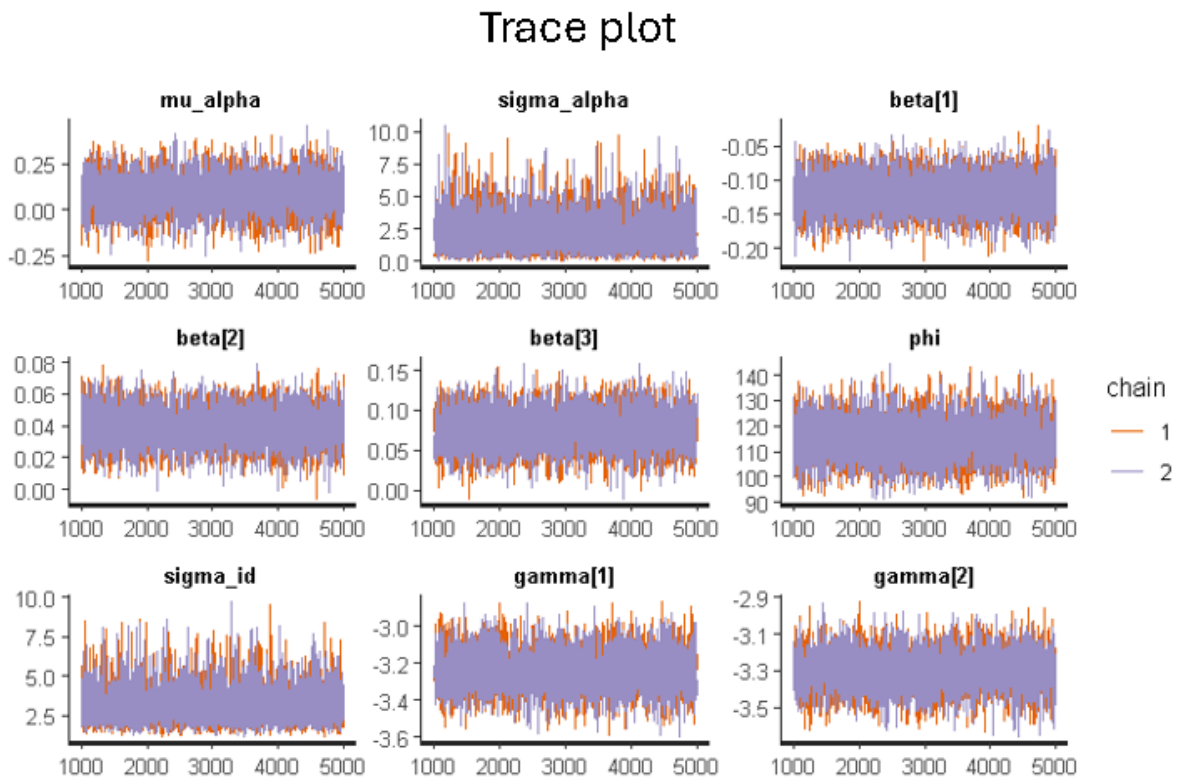
El parámetro β_1 ($\mu = -0.12$) representa el efecto del Presupuesto, indicando que un aumento en el presupuesto tiene un efecto negativo en la respuesta del modelo. Aunque el efecto es pequeño, su dirección sugiere que, una vez controladas otras variables, un mayor presupuesto no siempre se traduce en una mejora inmediata en la variable objetivo, este impacto puede tardar dado que el presupuesto se puede ver reflejado en infrastructura, capacitaciones y reformas en instituciones, cosas que pueden tomar tiempo en ver sus efectos.

Por otro lado, β_2 ($\mu = 0.04$) captura el efecto de la Aprobación al cuadrado. Su signo positivo implica una relación no lineal donde el impacto de la aprobación se incrementa en valores más altos, sugiriendo que los sistemas con mejores niveles de aprobación pueden experimentar mejoras acumulativas en el desempeño educativo. β_3 ($\mu = 0.08$), que representa la Reprobación, tiene un efecto positivo, indicando que mayores niveles de reprobación están asociados con un aumento en la variable dependiente, lo que puede reflejar dinámicas en las que la reprobación afecta la estabilidad del sistema educativo, aumentando la desercion.

En el nivel jerárquico, los parámetros γ_1 ($\mu = -3.24$) y γ_2 ($\mu = -3.30$) representan las medias de los efectos aleatorios para los grupos Andino y No Andino, respectivamente. Aunque la diferencia entre estos valores parece pequeña, en un modelo Beta jerárquico, incluso diferencias sutiles pueden ser significativas en la distribución de la variable respuesta. El hecho de que No Andino tenga una media ligeramente menor sugiere que, en términos generales, su desempeño es más bajo en comparación con el grupo Andino.

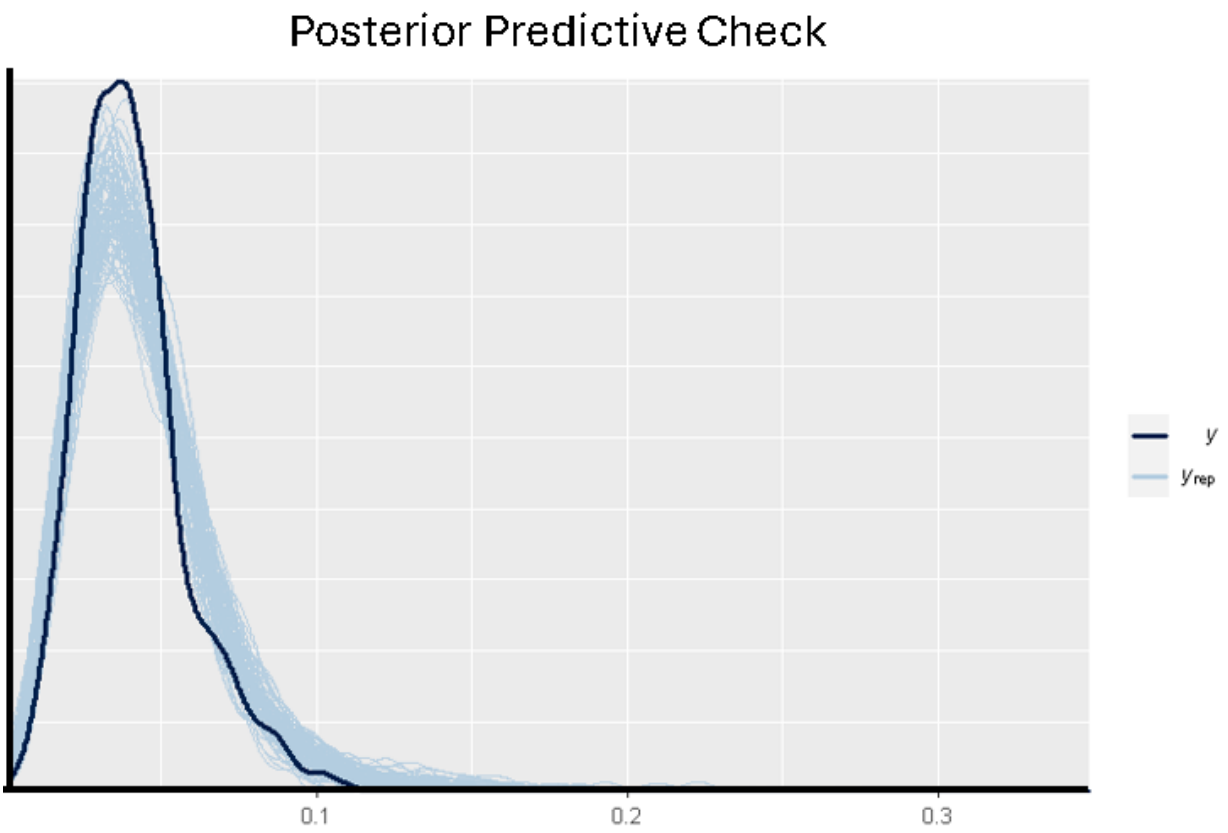
Finalmente, la buena convergencia del modelo se confirma con valores de Rhat cercanos a 1 en todos los parámetros, lo que indica que las cadenas han alcanzado una distribución estacionaria y que las inferencias son confiables. Sin embargo, la dispersión en algunos parámetros, como σ_α (2.00, sd = 1.37) y σ_{id} (3.10, sd = 1.09), sugiere que existe variabilidad entre unidades individuales y que podría haber heterogeneidad en la asignación y uso del presupuesto. En resumen, el modelo sugiere que la asignación de recursos y el desempeño educativo están influenciados por factores no solo individuales sino también estructurales, como la región geográfica y la dinámica de aprobación y reprobación.

Evaluación de Convergencia y Ajuste del Modelo



El Trace Plot muestra las trayectorias de las cadenas de Markov para diferentes parámetros del modelo, representando cada color una cadena diferente.

Se observa que las cadenas parecen mezclarse bien, es decir, oscilan alrededor de un valor estable sin tendencias claras o grandes saltos. Esto indica que no existe evidencia de no convergencia dentro de los parámetros del modelo.

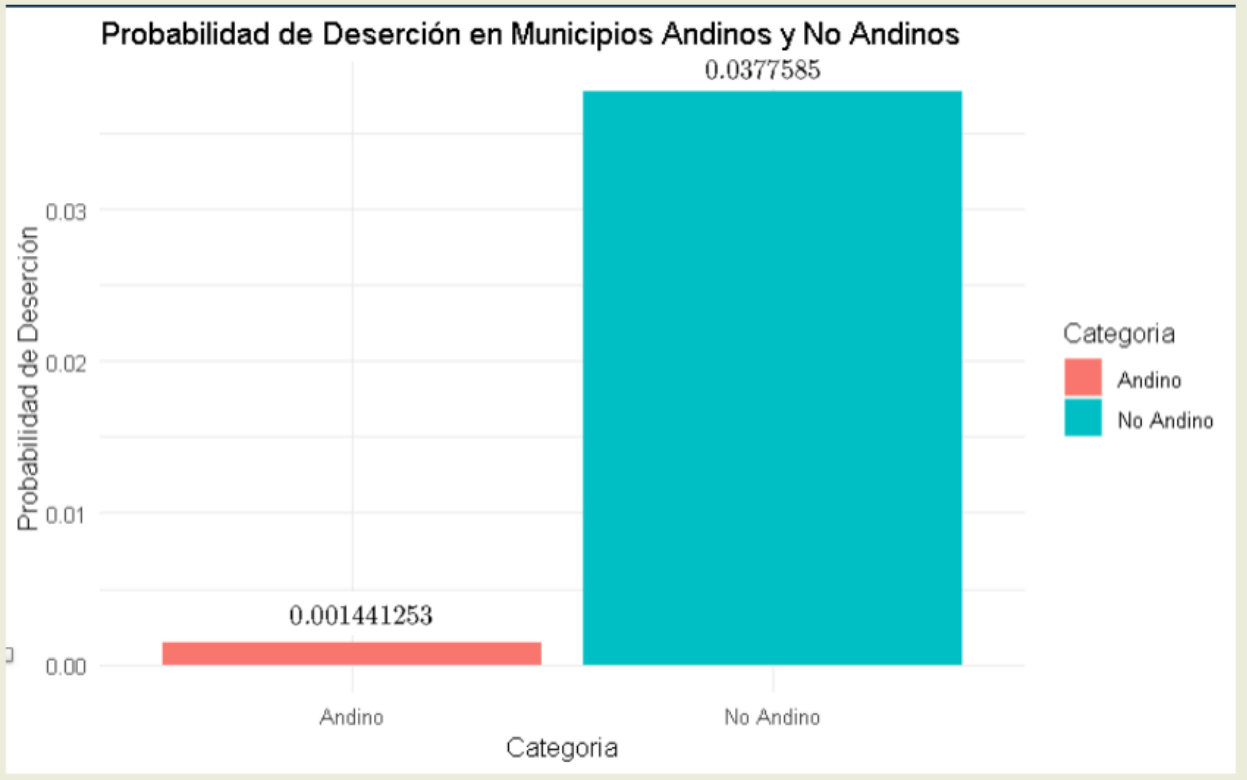


El Posterior Predictive Check permite, a partir del modelo ajustado, simular datos usando las muestras de la distribución posterior de los parámetros para comparar estas simulaciones con los datos observados. Si el modelo captura bien la estructura de los datos, las características como la forma y la variabilidad de los datos simulados serán similares a las de los datos reales.

Se observa que el modelo capta bien la distribución de los datos observados, aunque puede presentar ligeras desviaciones en las colas de la distribución.

Probabilidad de desercion

En el presente apartado abordaremos la pregunta propuesta al inicio, referente a la probabilidad de deserción discriminada por su categoría departamental. Se tratará de responder esta pregunta de interés mediante el modelo propuesto y desarrollado a lo largo del artículo.



Notamos que la probabilidad de deserción en un departamento no andino es significativamente mayor $\approx 4\%$ en comparación con la de un departamento andino $\approx 0.1\%$, lo que representa un resultado preocupante.

Además, se añade un intervalo de credibilidad del 80% para confirmar que, efectivamente, existe evidencia significativa que indica que la probabilidad de deserción en un departamento andino es diferente a la de un departamento que no lo es.

$$\begin{cases} \text{Percentil } 10\% : 0.0019 \\ \text{Percentil } 80\% : 0.1285 \end{cases}$$

Al no contener el 0, el intervalo de credibilidad del 80% nos permite concluir que existe una diferencia significativa entre la probabilidad de deserción educativa en un departamento andino y en uno que no lo es. Además, se confirma que la probabilidad de deserción en los departamentos no andinos es considerablemente más alta.

Concluciónes

- El uso de un modelo beta jerárquico para modelar la deserción estudiantil en Colombia, basado en si su departamento pertenece o no a la región andina, funcionó de manera aceptable.
- Se identificaron algunos factores de riesgo asociados a la deserción estudiantil en Colombia, los cuales permiten explicar este fenómeno de manera eficiente.
- Se observa que la relación entre factores y coeficientes es no lineal.
- Fue posible calcular la probabilidad de deserción en función de si un departamento es andino o no. Se evidencia una amplia diferencia entre ambos, lo que resulta preocupante. Esto indica que, aunque es importante seguir reduciendo la deserción en los departamentos andinos, es crucial centrar esfuerzos en los departamentos no andinos.
- Este estudio sirve como una base funcional para futuras investigaciones sobre la deserción estudiantil en Colombia.
- Para una mejor implementacion y desarrollo del modelo se podria pensar en trabajar con expertos en el campo los cuales ayuden a poponer nuevas variables de ineteres, de igual manera el contar con datos previos al 2011 y seguir con la recoleccion de estos dará como resultado una gran mejora en la calidad del modelo desarrollado
- La modelación de la deserción estudiantil permitirá enfocar esfuerzos y recursos en áreas críticas, mejorando la calidad y continuidad educativa en Colombia.