
Reincidencia de la población carcelaria a cargo del Instituto Nacional Penitenciario en Colombia: caracterización y modelamiento estadístico

Recidivism of prison population in Colombia sampling the National Penitentiary Institute: characterization and statistical modeling

Raul Andres Diaz Fonseca^a
rauldiaz@usantotomas.edu.co

Mario Jose Pacheco Lopez^b
mariopacheco@usantotomas.edu.co

Resumen

En este documento se llevo a cabo un análisis de la población carcelaria en Colombia, con un enfoque particular en la reincidencia y la cantidad de reincidencias históricas por individuo. Utilizando datos de toda la población a noviembre de 2019 proporcionados por el Instituto Nacional Penitenciario Colombiano (INPEC), se exploraron diversas características sociales, demográficas y judiciales con el objetivo de identificar perfiles reincidentes. Debido a la distribución de las reincidencias que presenta una inflación de ceros, se compararon y aplicaron diferentes modelos estadísticos, incluyendo modelos de recuento inflados con ceros y modelos de recuento truncado en ceros, como Regresión de Poisson, Regresión Binomial Negativa, Poisson Inflado con Ceros (PIC), Binomial negativa inflada por ceros, Poisson cero truncado y Binomial negativa cero truncado. Además, se estimó un score de reincidencias que permitirá identificar perfiles propensos a la reincidencia delictiva. Este estudio arroja luz sobre la dinámica de la reincidencia carcelaria en Colombia y proporcionar información que puede contribuir a la formulación de políticas y programas de reinserción en el sistema penal del país, con una adecuada predicción y una estimación de conteo preciso de las reincidencias.

Abstract

This document presents a comprehensive analysis of the prison population in Colombia, with a particular focus on prison recidivism and the historical count of recidivism per individual. Using data provided by the National Penitentiary Institute of Colombia (INPEC), various social, demographic, and judicial characteristics will be explored to identify recurring profiles. Given the inflation of zeros in the recidivism distribution, different statistical models, including zero-inflated count models and zero-truncated count models, such as Poisson Regression, Negative Binomial Regression, Zero-Inflated Poisson (ZIP), Zero-Inflated Negative Binomial, Zero-Truncated Poisson, and Zero-Truncated Negative Binomial, will be compared and applied. Additionally, a recidivism score will be estimated to identify profiles prone to criminal recidivism. This study aims to shed light on the dynamics of prison recidivism in Colombia and provide insights that can contribute to the formulation of policies and reintegration programs within the country's penal system, aiming for the best prediction and accurate recidivism counts.

Palabras claves: Reincidencia carcelaria, Resocialización, Reincidencia delictiva, perfilamiento

^aEstudiante

^bDirector

1. Introducción

El aumento de los índices de reincidencia delictiva en Colombia, junto con los problemas de hacinamiento carcelario y la oleada mediática de delitos cometidos por personas con prisión domiciliaria, representa un desafío significativo para el sistema penitenciario del país. En el Informe de Derechos Humanos del Sistema Penitenciario en Colombia (2017-2018) realizado por la Universidad de los Andes, se destaca este fenómeno, subrayando la preocupación que genera en varios sectores de la sociedad. Además, esta problemática no es exclusiva de Colombia; es un tema que afecta a nivel regional.

Millones de personas en el mundo permanecen en estos momentos en un sitio de reclusión. Como puede ser previsible, entre mayor es la población que tiene un país, mayores probabilidades existen de que haya cientos de miles de sus habitantes encarcelados (EOM (2019)). Estados Unidos, por ejemplo, tiene 2.161.900 personas privadas de la libertad; el motivo principal de este sobredimensionamiento es la dura política judicial en el país, con penas muy elevadas y poco orientadas a la reinserción frente a una perspectiva punitiva muy asentada (EOM (2019)). En Latinoamérica, también existen altas tasas de encarcelamiento y de sobrepoblación de las cárceles, pues casi todos los países presentan alzas constantes en la cantidad de reclusos, con la excepción de Honduras y Paraguay, que se han mantenido estables en los últimos años, y Argentina, que ha logrado disminuir su tasa llegando a 144 encarcelados por cada 100.000 habitantes. En cuanto a la sobrepoblación, destacan Perú y El Salvador cuya población penal dobla la capacidad disponible, es decir, con una sobrepoblación que supera el 100 % de la capacidad instalada para alojar estos individuos (ICEPALA (2015)).

En algunos países de Latinoamérica, un tercio de la población ya ha sido condenada previamente. Muchos estudios coinciden en que las cárceles se convierten en una escuela del crimen, un espacio que potencia la violencia, los abusos a los derechos humanos y la conformación de redes criminales (CELIV (2015)). En un contexto en el que las personas llegan cada vez más jóvenes a las cárceles, aquellos que son apresados por delitos menores y que tal vez no tenían vínculos con el crimen organizado, tienen muchas probabilidades de terminar vinculados a él durante el encierro (CELIV (2015)).

Algunas explicaciones que se desarrollan respecto a la reincidencia en el delito en Latinoamérica están basadas en el tipo de contexto social y ambiental donde se socializaron estas personas. De modo que dicho contexto funciona como un factor coadyuvante en la reincidencia delictiva (CELIV (2015)). Por ende, se plantea la relación entre la proporción de reincidencia según si hay o no antecedentes de familiares presos. De modo tal que, dentro de la población en reclusión, se aumentan las probabilidades de reincidencia delictiva dentro del grupo de quienes han tenido antecedentes de familiares que han estado presos.

En Colombia, de acuerdo con información de la Fiscalía General de la Nación, entre enero de 2017 y octubre de 2018, se recibieron aproximadamente 29.000 notificaciones criminales de hurtos cometidos por personas con antecedentes penales. El 31 % de los capturados tenían entre dos y cuatro detenciones previas; otros 592 habían sido capturados y puestos en libertad entre cinco y seis veces en los últimos meses, mientras que 240 tenían de siete a diez anotaciones” (FIP (2019)). Por ejemplo, una persona puede entrar al sistema carcelario luego de haber sido condenado por hurto y, al mismo tiempo, tener antecedentes penales por un caso de inasistencia alimentaria. En este contexto, surge la pregunta de investigación central de este trabajo: ¿cuáles son los factores determinantes que influyen en la reincidencia carcelaria de la población bajo la supervisión del Instituto Nacional Penitenciario en Colombia (INPEC)?

En el presente trabajo se diseñó un modelo estadístico que permite estimar la probabilidad de reincidencia entre los reclusos del INPEC y la cantidad posible de reincidencias por recluso. Para ello, se emplearon modelos estadísticos que se ajusten a la naturaleza de los datos y la complejidad de la reincidencia carcelaria. Entre los enfoques metodológicos, se incluirán modelos de regresión con respuesta dicotómica (Probit y Logit) y modelos de conteo. El objetivo es proporcionar herramientas que contribuyan a la identificación de posibles casos de reincidencia, lo que a su vez podría respaldar futuras políticas de resocialización.

Para abordar esta problemática, se utilizarán diversos enfoques, como la Regresión de Poisson, la Regresión Binomial Negativa, el Poisson Inflado con Ceros (PIC), la Binomial Negativa Inflada por Ceros, el

Poisson Cero Truncado y la Binomial Negativa Cero Truncada. La base de datos proporcionada por el INPEC abarca la población carcelaria hasta noviembre de 2019, y se contrastarán los modelos con una población correspondiente a noviembre de 2020.

Esta investigación arroja luz sobre la reincidencia carcelaria en Colombia y ofrecer una base sólida para la toma de decisiones informadas en el ámbito penitenciario. Los modelos estadísticos aplicados se basarán en la realidad de la población carcelaria colombiana y se utilizarán para identificar factores cruciales que pueden contribuir a la comprensión de la reincidencia delictiva en el contexto del INPEC.

2. Antecedentes

Riaño (2016) explica la importancia de analizar los factores de riesgo que pueden jugar un papel importante en la reducción de la reincidencia y del retorno a la ilegalidad en un grupo de población desmovilizada reincidente de las Autodefensas Unidas de Colombia, AUC, recluido en el Complejo Penitenciario y Carcelario Metropolitano de Bogotá (COMEB). En ese aspecto, manifiesta que metodológicamente la relevancia del trabajo deriva de conocer, desde entrevistas a profundidad, los factores de riesgo relacionados con la reincidencia delictiva de este grupo de desmovilizados de las AUC, que ha limitado el éxito en la reintegración positiva a la sociedad.

La reintegración de esta población requiere de atención especial en el diseño de programas de resocialización desde el enfoque de la capacidad de agencia y ciudadanía, lo cual constituye un reto estructural para el Estado Colombiano de cara a la construcción de un escenario de postconflicto y de paz (Riaño (2016)). Por esta razón, el análisis de los factores de riesgo y de control de la reincidencia penitenciaria, sugiere que los programas de resocialización deben fomentar la participación de agentes de cambio comprometidos con proyectos de vida y el acoplamiento a la promoción de una ciudadanía responsable que contribuya al mejoramiento del entorno, la sociedad y el desarrollo del país (Riaño (2016)).

Caicedo (2014) explica cuáles son los factores que ayudan a explicar la duración del periodo entre la salida de un centro de reclusión y el posterior reingreso a uno de estos. Esto es, los determinantes de la duración hasta la reincidencia carcelaria. Estos factores son explicados a partir del análisis econométrico de variables que son teóricamente importantes, en consideración del modelo de acumulación diferenciada de capital humano lícito e ilícito de Mocan et al. (2004). Así pues, los resultados técnicos se interpretan a la luz de la teoría de la disuasión y de la carrera criminal. El objetivo de dicho trabajo consiste en permitir ver cómo las características individuales y de la pena asociadas a ambos tipos de capitales, influyen en la decisión de reincidir, y en este sentido, a la duración hasta el reingreso a un centro de reclusión. El modelo considerado por Caicedo (2014) es un modelo Weibull y las variables que logran explicar la duración hasta el reingreso a un centro de reclusión son: edad, edad al cuadrado, sexo, días en la sentencia condenatoria, días efectivamente cumplidos, número de capturas previas, delitos contra la vida e integridad personal, delitos contra personas y bienes protegidos por el Derecho Internacional Humanitario, delitos contra la libertad, integridad y formación sexuales, delitos contra el patrimonio económico, delitos contra la seguridad pública, si el individuo es divorciado y si reportó que tenía una unión marital de hecho.

En conclusión, Caicedo (2014) realiza un estudio econométrico basado en una muestra de internos en una cárcel de Colombia. El autor al respecto argumenta que, los resultados reflejan el estudio sobre el sector de la justicia penal en el país, debe incorporar los efectos del conflicto interno. Por otra parte, encuentra que los delitos asociados al tráfico de armas son estado-absorbentes, es decir, que una vez que los individuos entran en este camino, les costará salir de él, de lo contrario, los sujetos tendrán un alto potencial de reincidencia carcelaria.

Por su parte Bales, et al (2015) estudian la reincidencia y analizan ciertos resultados (nuevo encarcelamiento, detención y reconvicción) para delincuentes liberados de prisión. Sin embargo, los autores sugieren que, dentro de la población de individuos capaces de reincidir, sino se excluyen las poblaciones de corrección de la cárcel y la comunidad se omite gran parte del análisis. Además, los autores discuten

las herramientas de evaluación de riesgos como medio para predecir la reincidencia entre las poblaciones supervisadas y encarceladas.

Brandon K. Applegate Ray Surette Bernard J. McCarthy (1999), evalúan la influencia de una nueva generación carcelaria sobre la reincidencia. En esta investigación se explica que las cárceles de nueva generación tienen comportamientos distintos a los tradicionales en temas de reincidencia delictiva, las cuales representan una reforma fundamental de la estructura y del estilo de gestión en las cárceles tradicionales. A partir de métodos cuantitativos, Los resultados mostraron que el modelo de nueva generación no aumenta la reincidencia carcelaria y puede reducir la delincuencia posterior a la liberación de los presos.

FIP (2018), desarrolla en su investigación: el qué hacer con la reincidencia delincencial, por tanto, se cuestiona cómo ha evolucionado el tema de la reincidencia carcelaria en Colombia, para ello se recurrió a una metodología cuantitativa en la cual se tomaron datos estadísticos y la aplicación de encuestas para realizar la recolección de los datos. En ese sentido, se encontró que en febrero de 2018, la Fiscalía presentó el top de reincidentes, una lista de 825 personas que acumulan entre 20 y 24 entradas a centros de reclusión y tienen pendientes nuevas órdenes de captura para cumplir sentencias impuestas por jueces (FIP, 2018).

De acuerdo con información de la Fiscalía General de la Nación, entre enero de 2016 y octubre de 2017, esta institución recibió casi 29.000 noticias criminales de hurtos cometidos por personas con antecedentes penales. El 31 % de los capturados tenían entre dos y cuatro detenciones previas; otros 592 habían sido capturados y puestos en libertad entre cinco y seis veces en los últimos meses, “mientras que 240 tenían de siete a diez anotaciones. Según lo registrado por Fiscalía, los seis delitos de mayor reincidencia son, en su orden, el hurto (28 %); la fabricación y porte de armas (16 %); delitos relacionados con drogas (13,5 %); la violencia intrafamiliar (7 %); las lesiones personales (7 %), y los homicidios (6 %)” (FIP (2018)).

Se destaca que en los últimos seis años la población reincidente ha aumentado en un 47 % (tomando como base los 10.956 reincidentes registrados en 2010). Según las estimaciones realizadas por el economista Santiago Tobón para “el Ministerio de Justicia y del Derecho, a partir de información del INPEC, el 7,9 % de los internos condenados vuelve a ingresar al año siguiente de su salida. Esta cifra aumenta a 13 % para los dos años; 16,6 % para tres; 19,2 % para cuatro años, y 21 % para cinco” (FIP (2018)). Los internos condenados por hurto o delitos relacionados con porte de armas de fuego presentan tasas de reingreso y recondena mayor, en comparación con quienes son condenados por otros delitos. Además, los presos que tienen condenas previas y que” han recibido alguna sanción durante el cumplimiento de la pena, tienen un riesgo de reincidencia entre 4 y 12 veces mayor que quienes no tienen condenas previas o sanciones” (FIP (2018)).

Para el análisis de la reincidencia la aplicación de la estadística es fundamental para el entendimiento de las diferentes variaciones, comportamientos masivos y atípicos de una población a través del tiempo. Maltz (2001) realiza un análisis estadístico, donde aplica una comparación de modelos paramétricos y no paramétricos encontrando una diferencia significativa entre ellos, en el mismo análisis resalta la importancia de discriminar la población reincidente y no reincidente para poder entrenar lo modelos estadísticos.

Hilbe (2014) En los diferentes estudios donde se busca estimar un conteo es necesario definir cuáles son las distribuciones que representan mejor la variable objetivo. Para esto es necesario probar varias distribuciones que se ajusten al comportamiento de la variable a estudiar controlando la sobre-dispersión que se puede modelar mediante las distribuciones Binomial Negativa, Poisson y Gaussiana inversa de Poisson. Martines (2009) La distribución Binomial Negativa es una distribución discreta, cuenta el número de fracasos antes del i -ésimo éxito en “ $(N + i)$ ” experimentos independientes Bernoulli, Distribución Poisson: es una variable discreta, unimodal y se destaca por la propiedad de igualdad de media y varianza.

Según Aarón (2009) en su análisis estadístico para datos de conteo, el exceso de cero debe ser tratado mediante modelos adecuados que combinan conteos como lo son los modelos de Cero Inflados y modelos Hurdle; el modelo Cero Inflado da mayor peso a la probabilidad de que la variable objetivo sea igual a cero, el primer componente genera solo conteos de cero, pero el segundo genera el rango completo de

conteos, incluyendo los ceros. Hurdle es un modelo de dos componentes que combina un proceso binario para los valores que están por debajo del valor cero y un proceso que genera solo los conteos mayores que cero mediante un modelo Cero Truncado (Mullahy, 1986)

3. Justificación

Para Foucault (1976), la reincidencia penitenciaria como fenómeno social adquiere importancia para el estado y la sociedad colombiana por su gran impacto en el alto índice de hacinamiento, la baja efectividad en las estrategias de resocialización de los privados de la libertad y los pobres resultados en la garantía de reinserción moral y material en el mundo estricto de la economía (citado por Riaño, 2016).

La importancia de este trabajo radica en la necesidad de estudiar el fenómeno del crecimiento de la población carcelaria, para ello, se busca identificar un perfil de reincidencia carcelario para dar herramientas a la promoción y prevención de estas actitudes, identificadas por el perfil de la sociedad carcelaria. Igualmente, esta investigación contribuye en dar un aporte social a la resocialización de estos individuos, dando herramientas y recomendaciones que lleven a los privados de la libertad y así puedan tener mayores opciones de una mejor calidad de vida. Adicionalmente, esta investigación busca aportar a la academia ya que en la actualidad la literatura sobre la reincidencia carcelaria en Colombia es escasa, ya sea por la confidencialidad de los datos o falta de interés, con lo cual se espera que el aporte de esta investigación permita el desarrollo de futuras investigaciones sobre este tema.

4. Marco teórico

Algunas definiciones necesarias para la comprensión del problema a tratar son presentadas a continuación:

Reincidencia delictiva: Desde la concepción más simple, la reincidencia se refiere a la repetición de un acto delictivo por parte del delincuente. Este concepto puede inferirse desde la simple observación y la reunión de algunos conocimientos básicos, pues sin hacer uso de herramientas idiomáticas es posible aclarar dicho significado (Ossa, 2012).

La reincidencia delictiva es un fenómeno de vieja data en la historia de la criminalidad. El término proviene del latín *reincidere*, que significa volver a hacer. Desde la antigüedad, la reincidencia se ha considerado como un agravante de la pena en procesados que poseen estos antecedentes; sin embargo, en muchas legislaciones no se considera como tal. El aumento o disminución de la criminalidad ha dependido de las políticas en materia de seguridad y criminológicas en los países, aunado esto al tipo, calidad y cantidad de la publicidad que ejercen los medios de comunicación. La exacerbación de políticas de “mano dura” de las administraciones gubernamentales y de los medios de comunicación tienen efectos nocivos en el desarrollo de la criminalidad (López, 2013).

La aplicación de la reincidencia en Colombia dedica especial atención a la reincidencia penitenciaria y hace referencia a aquellos individuos que, habiendo sido condenados, han vuelto a ser privados de la libertad o se les ha impuesto una pena en establecimientos penitenciarios (Ossa, 2012). En ese sentido, la reincidencia penitenciaria es del 24 % para nuestro país, lo cual no es un dato alentador si se tiene en cuenta que quedan al margen los delitos que no llegan a ser judicializados, que no se reportan, que no se visualizan; por ello, no hay datos específicos para definir un indicador preciso al respecto.

En Colombia, según la Corte Constitucional en su sentencia C-077 de 2006: La reincidencia es una especie de las circunstancias modificativas agravantes de responsabilidad, prevista en algunos ordenamientos penales y, más ampliamente, en algunos ordenamientos sancionatorios, en virtud de la cual se agrava la sanción impuesta al infractor cuando ha sido sancionado anteriormente por la comisión de otras infracciones.

En nuestra legislación actual, y para concretar de qué manera opera el instituto de la reincidencia en

nuestro país, el artículo 55 del Código Penal consagra las Circunstancias de menor punibilidad, y en el numeral primero se encuentra taxativamente la carencia de antecedentes penales, circunstancia que se tiene en cuenta para que el juez escoja el cuarto dentro del cual deberá moverse para imponer la correspondiente pena según lo manda el artículo 61 del mismo estatuto, al consagrar los fundamentos para la Individualización de la pena, quiere decir esto que si el acusado posee antecedentes penales. El juez va a estar impedido a la hora de realizar la ejecución penal para ubicarse en el cuarto mínimo de movilidad, vislumbrándose concretamente dicho instituto como una causal de agravación de la pena, aunque no esté tipificado de dicha forma como en el caso de las agravantes específicas del tipo penal, así muchos sectores en el campo del derecho sigan insistiendo que al reincidente la pena no se le agrava porque no es una causal de agravación (Ossa, 2012)

En suma, la sentencia C-181 de 2016 define la reincidencia delictiva como una recaída en el delito, por parte de quien ya había sido condenado penalmente con anterioridad, lo cual genera una reacción social y jurídica en términos punitivos, pues agrava la pena del nuevo delito (Martínez, Peña y Peña (2017)). Afirma además la Corte en esta misma sentencia, que la reincidencia es un elemento accidental y accesorio a la pena y “al delito en sí mismo considerado, ya que no condiciona la existencia de los elementos dogmáticos que son: la conducta, la tipicidad, la antijuricidad y la culpabilidad” (Ossa (2012)).

Resocialización: Para Arias (2017) La Corte Constitucional colombiana concibe la resocialización como un derecho al que deben tener acceso todos los individuos privados de la libertad en establecimientos carcelarios en Colombia: El derecho a participar en los programas de educación y trabajo representa una actividad que tiene como finalidad la resocialización y el refuerzo de la concepción del mismo como valor fundante de la sociedad. El objetivo principal de la participación del recluso en programas de educación y trabajo es preparar al interno para su vida en libertad; por lo tanto, las actividades laborales y de educación se tornan de carácter obligatorio para aquellos reclusos que tengan la calidad de condenados, teniendo en cuenta su finalidad de resocialización (Jiménez (2017))

De conformidad con lo señalado por el organismo que administra el sistema penitenciario y carcelario colombiano, la resocialización es una “técnica de tratamiento clínico que pretende cambiar la conducta del interno. Volver a socializarse, lo que significa aprender las expectativas sociales e interiorizar normas de conducta. Resocializarse es volver a valer como ser social conforme quiere la sociedad.

Aspectos de la resocialización dentro del tratamiento penitenciario colombiano.

1. Es un tratamiento que se brinda al condenado tras la sentencia de responsabilidad penal; 2. Este tratamiento le permitirá retornar a la sociedad con el conocimiento del actuar errado y la concientización por el respeto a las normas establecidas por la sociedad; 3. El tratamiento genera un bloqueo frente a la comisión de nuevos delitos en el futuro. De manera operativa, este proceso se obtiene a través del trabajo, el estudio, la disciplina, la instrucción, la cultura, arte, la recreación y las relaciones de familia (Jiménez (2017)).

Asimismo, para preparar al individuo hacia el tránsito a la vida en libertad, se le deben ofrecer opciones de contacto con la sociedad extramuros, por lo que adquieren importancia los diferentes permisos y beneficios penitenciarios, que le permiten salir de la prisión con anterioridad al cumplimiento de la pena (Jiménez (2017)). Se desprende de lo anterior que, dentro de los programas de resocialización, que, a su vez, les permiten redimir pena por trabajo, estudio y enseñanza. Pero incluso señala la legislación penitenciaria, que se puede redimir pena por actividades literarias, deportivas, artísticas y en comités de internos, las cuales se asimilan al estudio.

Los modelos estadísticos de regresión Poisson suelen ser aplicados para estimar eventos de conteo, específicamente, en contar y estimar datos en forma discreta de enteros no negativos, determinando la cantidad de veces que puede ocurrir un evento durante una frecuencia establecida (Issac y Rivera, 2021), de la misma forma el modelo de regresión binomial negativa busca un resultado similar, con la diferencia que este permite que la varianza condicional de la variable estimada sea mayor que la media por lo que sería más preciso para distribuciones sobre dispersas, por otro lado la regresión binomial nos permite una mejor estimación para varianzas menores que la media, es decir, distribuciones con baja dispersión (Delgado, 2015). En muchas ocasiones la cantidad de ceros en la población se presenta con alta frecuencia por

lo que se denomina “exceso de ceros” o que están “inflados con ceros”, por lo que se debe tener especial cuidado, ya que cualquier estimación puede ser inexacta debido a sus parámetros. Por lo que es necesario analizar estas distribuciones mediante el modelo de regresión de Poisson con exceso de ceros (ZIP por sus siglas en inglés) propuesta Lambert (1992) y Binomial Negativa inflada por ceros que buscan resolver este inconveniente publicada por Greene (1994).

En muchas investigaciones y análisis de datos, se enfrentaba al desafío de tratar con el “exceso de ceros” en la población de interés. Este fenómeno se refería a la presencia de un gran número de observaciones con valores iguales a cero, lo que podía distorsionar significativamente cualquier estimación estadística. En otras palabras, los datos estaban inflados con ceros, lo que planteaba interrogantes sobre la precisión de cualquier análisis subsiguiente. Estos casos especiales requerían un enfoque estadístico más refinado para garantizar que las estimaciones fueran lo más precisas posible.

En respuesta a este desafío, se desarrollaron diversos modelos estadísticos que buscaban abordar el problema del exceso de ceros. Entre estos modelos, dos enfoques ampliamente utilizados eran el “Modelo de Regresión de Poisson con Exceso de Ceros” (ZIP, por sus siglas en inglés) propuesto por Lambert (1992) y el “Modelo de Regresión Binomial Negativa Inflada por Ceros” introducido por Greene (1994). Estos modelos habían demostrado ser herramientas valiosas para analizar datos con exceso de ceros, permitiendo una estimación más precisa de los parámetros de interés.

En el contexto del estudio, se exploró la aplicación de estos modelos y otros enfoques probabilísticos y de conteo en el contexto del estudio. El objetivo era comprender cómo estos modelos podían contribuir a resolver los desafíos planteados por el exceso de ceros y cómo se podían aplicar de manera efectiva a los datos. Cada modelo tenía sus propias características y suposiciones, y se analizó cómo se comportaban en el entorno de investigación específico.

Se presentó cada uno de estos modelos en detalle, discutiendo sus fundamentos teóricos, su aplicabilidad y sus ventajas y desventajas. A través de este análisis, se esperaba encontrar la mejor manera de abordar el problema del exceso de ceros y obtener estimaciones más precisas en el estudio.

4.1. Asociación de variables

Con el fin de analizar qué tan asociadas están las variables entre sí y con la variable de estudio; la reincidencia de la población penitenciaria, se calcula una medida de correlación. Se utiliza para comprender las interacciones entre variables y predecir el comportamiento de una variable basándose en la otra. Sin embargo, es importante destacar que la correlación no implica causalidad, es decir, el hecho de que dos variables estén correlacionadas no significa necesariamente que una causa la otra.

El coeficiente de correlación más comúnmente utilizado es el coeficiente de correlación de Pearson, que se basa en la covarianza entre las variables y sus desviaciones estándar. Sin embargo, existen otros coeficientes de correlación, como el coeficiente de correlación de Spearman, que se basa en el rango de las observaciones, pero para realizar el presente estudio se utiliza la correlación V de Cramer dado que se estudian variables cuantitativas y cualitativas, las cuales se convierten en factores y se procede a realizar una asociación.

La probabilidad de reincidencia es una medida de asociación utilizada para evaluar la relación entre dos variables categóricas. Es una extensión del coeficiente de contingencia de Pearson y se utiliza específicamente cuando las variables categóricas tienen más de dos categorías, varía entre 0 y 1, donde 0 indica ausencia de asociación y 1 indica una asociación completa. Cuanto más cerca esté el valor de 1, mayor será la asociación entre las variables. Es importante tener en cuenta que la correlación V de Cramer solo evalúa la asociación entre variables categóricas y no puede utilizarse para variables numéricas. Esta medida es ampliamente utilizada en el campo de la estadística y la investigación social para determinar la fuerza de la relación entre variables categóricas en análisis de datos, como encuestas, estudios demográficos y análisis de datos cualitativos.

La fórmula correspondiente es:

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(r-1, c-1)}} \quad (1)$$

Donde:

- V representa la correlación de V de Cramer.
- χ^2 es el valor de la prueba de chi-cuadrado asociado a la tabla de contingencia.
- n es el tamaño total de la muestra o el número total de observaciones.
- r es el número total de categorías de la primera variable.
- c es el número total de categorías de la segunda variable.

4.2. Synthetic Minority Over-sampling Technique (SMOTE): Balanceo de Muestra

El desbalance de clases es una problemática común en problemas de clasificación, donde una de las clases (generalmente la minoritaria) tiene una representación significativamente menor que la otra. En el contexto de la reincidencia carcelaria en Colombia, esta asimetría se manifiesta en una cantidad sustancial de individuos no reincidentes, en comparación con un número considerablemente inferior de reincidentes. En términos específicos, la base de datos revela que una mayoría de individuos están clasificados como no reincidentes, mientras que solo una minoría de la población ha sido catalogada como reincidente. Es necesario entender la magnitud de esta disparidad para abordar de manera efectiva los desafíos asociados con el desbalance de clases.

Para abordar este desequilibrio, se emplea la técnica SMOTE desarrollada por Chawla et al. (2002). SMOTE se basa en la idea de generar instancias sintéticas de la clase minoritaria mediante una interpolación ponderada de ejemplos existentes. La metodología de SMOTE sigue los siguientes pasos:

Selección de Vecinos: Para cada ejemplo de la clase minoritaria, se seleccionan k vecinos más cercanos dentro de la misma clase.

Generación de Instancias Sintéticas: Se generan nuevas instancias sintéticas mediante una combinación lineal de un ejemplo de la clase minoritaria y uno de sus vecinos. Esto se logra mediante la fórmula:

$$Nuevo_Ejemplo = Ejemplo_Min_Clase + (Ejemplo_Min_Clase - Vecino)\lambda$$

Donde:

- *Nuevo_Ejemplo*: Representa la instancia sintética generada, que se añadirá al conjunto de datos original para equilibrar la representación de la clase minoritaria.
- *Ejemplo_Min_Clase*: Corresponde a un ejemplo de la clase minoritaria seleccionado para la generación de instancias sintéticas.
- *Vecino*: Indica uno de los vecinos más cercanos al *Ejemplo_Min_Clase* dentro de la misma clase minoritaria. La elección de vecinos es fundamental para preservar las características distintivas de la clase minoritaria.
- λ : Es un valor aleatorio en el rango $[0, 1]$ que determina la proporción de la combinación lineal entre el *Ejemplo_Min_Clase* y su *Vecino*. Este valor introduce variabilidad en la generación de instancias, contribuyendo a la diversidad de las nuevas muestras.

Al comprender cada componente de la ecuación, se destaca la importancia de *Nuevo_Ejemplo* como una instancia sintética que conserva las características de la clase minoritaria, facilitando así un equilibrio en la distribución de clases en el conjunto de datos.

Incorporación de Instancias Sintéticas: Incorporación de Instancias Sintéticas: Las instancias sintéticas se añaden al conjunto de datos original para equilibrar la representación de ambas clases, la no reincidente y la reincidente.

SMOTE es una técnica ampliamente reconocida en la comunidad de aprendizaje automático para tratar desequilibrios en conjuntos de datos. Al aplicar SMOTE en el contexto de la reincidencia carcelaria utilizando la librería DMwR de RStudio, se logró mejorar la capacidad predictiva de los modelos, permitiendo una evaluación más precisa de los factores relacionados con la reincidencia, a pesar de la distribución desigual de clases en la base de datos.

4.3. Modelos Probabilísticos.

Con el fin de estimar la probabilidad de reincidencia de cada individuo se utilizarán modelos predictivos donde la variable objetivo es de carácter binario. Como son los modelos Logit, Probit y Elastic Net. Una vez calculada la probabilidad de cada modelo se procederá a calcular un Score de Reincidencia, un Score para evaluar o cuantificar el desempeño o la posición relativa de un individuo dentro de la población carcelaria. Este Score se calculará de la siguiente forma:

$$Score_i = 1000(1 - P_i) \quad (2)$$

Donde: P_i es la probabilidad estimada

Para este primer ejercicio de regresión, trabajado sobre la base de datos del INPEC de noviembre de 2019, se entrenarán los modelos utilizando el 70 % de la base de datos seleccionado de manera aleatoria sin reemplazo. Se utilizará una semilla para garantizar la replicabilidad de los resultados. El restante 30 % de la base de datos se reservará para probar la eficiencia del modelo.

4.3.1. Modelo Logit

El Modelo Logit es una técnica ampliamente utilizada en estadística y econometría para estimar la probabilidad de un evento binario, como la reincidencia en el contexto de esta investigación. Este modelo se basa en la transformación logarítmica de la razón de las probabilidades.

Denotamos L_i como el logaritmo natural de la razón entre las probabilidades P_i (probabilidad de que el evento ocurra) y las probabilidades complementarias $(1 - P_i)$ (probabilidad de que el evento no ocurra). La fórmula correspondiente es la siguiente:

$$L_i = \ln \left(\frac{P_i}{1 - P_i} \right) \quad (3)$$

Donde:

- L_i es el valor logit para el individuo i .
- P_i es la probabilidad de que el evento de interés ocurra para el individuo i .

El Modelo Logit se ajusta utilizando métodos de máxima verosimilitud, que buscan encontrar los valores de los parámetros del modelo que maximizan la probabilidad de observar los datos observados bajo el modelo. Un aspecto importante del Modelo Logit es que transforma el espacio de las probabilidades (que va de 0 a 1) a un espacio logit (que abarca todos los números reales). Esto facilita la interpretación de

cómo las variables predictoras afectan la probabilidad del evento. Un aumento en el logit de P_i implica un cambio proporcional en la probabilidad P_i . Esto permite analizar el impacto de las variables predictoras sobre la probabilidad de reincidencia de manera más intuitiva.

El Modelo Logit es una herramienta poderosa para modelar la probabilidad de eventos binarios y es ampliamente utilizado en campos como la criminología y la economía. Proporciona información valiosa sobre cómo las variables explicativas influyen en la probabilidad de reincidencia y se utiliza comúnmente en estudios relacionados con la justicia penal.

Para este proceso se calcularon varios modelos a través de un análisis de varianza (Anova), con el fin de evaluar el efecto de cada variable predictora sobre la variable dependiente. Este análisis proporcionó información sobre la significancia de cada variable en el modelo general. Posteriormente, se llevó a cabo un proceso de selección de variables utilizando la técnica de paso a paso, también conocida como *stepwise*.

El método *stepwise* es una estrategia sistemática para seleccionar variables, que puede incluir tanto la adición como la eliminación de variables del modelo. Inicia con un modelo nulo y evalúa sistemáticamente la adición o eliminación de variables en función de criterios específicos, como la significancia estadística o medidas de ajuste del modelo.

Los resultados de la selección de variables revelaron las variables que contribuyen de manera significativa a la capacidad predictiva del modelo. Cada paso del proceso mostró qué variables se incluyeron o excluyeron y proporcionó una comprensión más profunda de cómo estas variables afectan la variable dependiente.

En este contexto, la selección de variables no solo se centró en identificar las variables más importantes, sino también en entender cómo la inclusión o exclusión de cada variable afectó la capacidad del modelo para explicar la variabilidad en la variable dependiente. Este proceso permitió refinar y optimizar el modelo, mejorando su capacidad para predecir y comprender los factores que influyen en la reincidencia carcelaria en la población estudiada.

$$L_i = \ln \left(\frac{P_i}{1 - P_i} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k \quad (4)$$

Donde:

- L_i es el valor logit para el individuo i .
- P_i es la probabilidad de que el evento de interés ocurra para el individuo i .
- β_0 es la intersección o término constante del modelo.
- $\beta_1, \beta_2, \dots, \beta_k$ son los coeficientes que representan la contribución de cada variable predictora ($X_1, X_2, \dots, X_k$) al logit.
- $X_1, X_2, \dots, X_k$ son las variables predictoras.

Esta ecuación logit expresa cómo la combinación lineal de las variables predictoras ponderadas por sus respectivos coeficientes (β) se transforma mediante la función logística para estimar la probabilidad del evento binario. El logaritmo natural de la razón de las probabilidades ($\ln \left(\frac{P_i}{1 - P_i} \right)$) representa el espacio logit, y al aplicar la función logística, obtenemos la probabilidad P_i . Este proceso facilita la interpretación del impacto de las variables predictoras en la probabilidad del evento de interés.

4.3.2. Modelo Probit

El Modelo Probit es una técnica estadística utilizada para analizar y modelar variables binarias o dicotómicas, como la reincidencia en el contexto de tu investigación. Al igual que el Modelo Logit, el Modelo Probit se basa en la estimación de la probabilidad de ocurrencia de un evento.

En lugar de emplear la función logística, el Modelo Probit utiliza la función de distribución normal acumulativa, conocida como función probit, para relacionar las variables predictoras con la probabilidad de ocurrencia del evento. La función probit es una función que asigna un valor entre 0 y 1 a una variable latente y se define como:

$$P_i = P(Y = 1|X) = P(u \leq 0) \quad (5)$$

Donde:

- P_i representa la probabilidad de que la variable dependiente Y sea igual a 1 (por ejemplo, la probabilidad de reincidencia) para el individuo i en función de las variables predictoras X .
- u es una variable latente que sigue una distribución normal estándar.
- Y es la variable dependiente que toma el valor 1 si el evento de interés ocurre (por ejemplo, reincidencia) y 0 en caso contrario.

La probabilidad P_i se relaciona con la variable latente u de la siguiente manera:

$$P(u \leq a + b_1X_1 + b_2X_2 + \dots + b_kX_k) = \Phi(a + b_1X_1 + b_2X_2 + \dots + b_kX_k) \quad (6)$$

Donde:

- $a, b_1, b_2, \dots, b_k$ son parámetros que deben estimarse.
- $\Phi(\cdot)$ es la función de distribución acumulativa normal, también conocida como la función probit.

La ecuación anterior puede ser simplificada cuando evaluamos la probabilidad en $u = 0$ para obtener la probabilidad base $P(Y = 1|X)$.

La estimación de los parámetros $a, b_1, b_2, \dots, b_k$ se realiza mediante métodos de máxima verosimilitud, que buscan encontrar los valores que maximizan la probabilidad de observar los datos observados bajo el modelo.

El Modelo Probit es una alternativa sólida al Modelo Logit y puede proporcionar información valiosa sobre cómo las variables predictoras influyen en la probabilidad de reincidencia. Algunos estudios previos han encontrado que el Modelo Probit es especialmente adecuado cuando se asume una distribución normal en la variable latente.

$$P_i = \Phi(\beta_0 + \beta_1X_1 + \beta_2X_2 + \dots + \beta_kX_k) \quad (7)$$

Donde:

- P_i es la probabilidad de que la variable dependiente Y sea igual a 1 (por ejemplo, la probabilidad de reincidencia) para el individuo i en función de las variables predictoras $X_1, X_2, \dots, X_k$.
- $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ son los coeficientes a estimar.
- $\Phi(\cdot)$ es la función de distribución acumulativa normal, también conocida como la función probit.

4.3.3. Modelo Elastic Net para Regresión Logística

El Modelo Elastic Net es una técnica de modelado que combina las características de dos métodos de regularización ampliamente utilizados en regresión lineal: L1 (Lasso) y L2 (Ridge). La regresión Lasso utiliza la penalización L1, que agrega el valor absoluto de los coeficientes a la función de pérdida, favoreciendo la dispersión de los coeficientes, por lo tanto, promoviendo la selección automática de variables al reducir algunos de ellos a cero. La regresión Ridge, por otro lado, utiliza la penalización L2, que agrega el cuadrado de los coeficientes a la función de pérdida, controlando la magnitud global de los coeficientes y ayudando a mitigar los efectos de la multicolinealidad.

El Modelo Elastic Net puede adaptarse para estimar variables dicotómicas, como la reincidencia carcelaria, a través de una regresión logística. La función de pérdida para la regresión logística con penalizaciones L1 y L2 es:

$$\text{Minimizar } L(\beta) = -\frac{1}{N} \sum_{i=1}^N [y_i \ln(P_i) + (1 - y_i) \ln(1 - P_i)] + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \quad (8)$$

Donde:

- N es el número de observaciones.
- y_i es la variable dependiente dicotómica para la observación i .
- P_i es la probabilidad predicha de que y_i sea igual a 1.
- β es el vector de coeficientes a estimar.
- λ_1 y λ_2 son parámetros de regularización no negativos que controlan la fuerza de las penalizaciones L1 y L2, respectivamente.
- $\|\cdot\|_2^2$ representa la norma euclidiana al cuadrado.
- $\|\cdot\|_1$ representa la norma L1.

Esta formulación adapta el Modelo Elastic Net a la regresión logística, permitiendo la estimación de variables dicotómicas como la reincidencia. La solución para el modelo se obtiene ajustando los coeficientes β que minimizan la función de pérdida. La introducción de penalizaciones L1 y L2 facilita la selección automática de variables y la regularización del modelo, especialmente útil en presencia de multicolinealidad.

4.4. Modelos de Conteo

En el análisis de datos de conteo, es común utilizar modelos estadísticos para comprender y predecir eventos discretos, como la cantidad de reincidencias delictivas en un individuo. En esta sección, exploraremos dos modelos de conteo ampliamente utilizados: la Regresión de Poisson y la Regresión Binomial Negativa.

4.4.1. Regresión de Poisson

La Regresión de Poisson es un modelo lineal generalizado que se emplea para modelar datos de conteo, donde la variable dependiente Y sigue una distribución de Poisson. Utiliza un vector de variables independientes X para predecir la tasa de ocurrencia de eventos. La expectativa y varianza de Y dado X se expresan como:

$$\lambda = E(Y|X) = e^{X'\beta} \quad (9)$$

Y

$$\log E(Y|X) = X'\beta \quad (10)$$

Si tenemos N observaciones, la función de log - verosimilitud es:

$$\mathcal{L}(\beta|X, Y) = \sum_{i=1}^N \left(Y_i X'_i \beta - e^{X'_i \beta} - \log(Y_i!) \right) \quad (11)$$

Desarrollando esta expresión, obtenemos:

$$\mathbb{E}(Y|X) = \exp(X\beta) = \exp(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k) \quad (12)$$

En los modelos de conteo del presente documento, Y representa la variable de conteo, X el vector de variables predictoras, y β los coeficientes a estimar. La tasa de ocurrencia λ se modela como una función exponencial de la combinación lineal de las variables predictoras. Para estimar los parámetros del modelo, se utiliza el método de máxima verosimilitud, buscando los valores de β que maximizan la probabilidad de observar los datos bajo el modelo.

4.4.2. Regresión Binomial Negativa

En la Regresión Binomial Negativa, la variable dependiente Y representa conteos y está relacionada con las variables predictoras X_i , que pueden ser categóricas o cuantitativas. Este modelo se ajusta mediante máxima verosimilitud y mínimos cuadrados ponderados, lo que permite la selección de variables por pasos y la prueba de la significancia de los coeficientes mediante pruebas de razón de verosimilitud.

La variable dependiente Y sigue la distribución de la Binomial Negativa y se define como:

$$P(Y = y/\mu) = \frac{\Gamma(y + \alpha^{-1})}{\Gamma(y + 1)\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu} \right)^{\alpha^{-1}} \left(\frac{\mu^{-1}}{\alpha^{-1} + \mu} \right)^y, \quad \mu > 0, \quad \alpha \geq 0$$

Donde:

$$\mathbb{E}(Y) = \mu$$

La función de log-verosimilitud para este modelo es:

$$\begin{aligned} \mathcal{L}(\alpha, \beta|X, Y) = \sum_{i=1}^N & \left[y_i \log \left(\frac{\alpha}{\mu + \alpha} \right) \right. \\ & \left. + \log \left(\frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\Gamma(\alpha^{-1})} \right) + \alpha^{-1} \log \left(\frac{\mu}{\mu + \alpha} \right) - y_i \log(\mu + \alpha) \right] \end{aligned}$$

Siendo N el número de observaciones, y_i el conteo para la observación i , y X las variables predictoras. Esta función de log-verosimilitud mide la concordancia entre el modelo y los datos observados, y el objetivo es maximizarla para obtener los valores óptimos de los parámetros α y β .

Estos modelos son herramientas poderosas para analizar datos de conteo y proporcionan información valiosa sobre cómo las variables explicativas afectan la probabilidad de eventos, como la reincidencia delictiva.

4.4.3. Binomial Negativa Inflada por Ceros

La Binomial Negativa Inflada por Ceros es un modelo estadístico que combina elementos de la distribución binomial negativa y la distribución binomial para manejar datos de conteo con una alta proporción de ceros. En este modelo, la función de distribución de probabilidad se define como:

$$F(Y_1 = y) = \begin{cases} \pi_i + (1 - \pi_i) \frac{k}{\mu_i + k}, & \text{si } y = 0 \\ (1 - \pi_i) f_{NB}(y_i), & \text{si } y = 1, 2, \dots \end{cases}$$

Donde:

- F es la función de distribución de probabilidad de la binomial negativa.
- Y_1 representa el conteo de eventos.
- π_i es la probabilidad de ocurrencia de ceros en los datos, es decir, la probabilidad de que no ocurra el evento de interés.
- μ_i es la media del modelo binomial negativo, que representa la tasa de ocurrencia del evento.
- k es un parámetro de sobredispersión que refleja la variabilidad adicional en los datos.

En este modelo, la presencia de ceros se modela como una distribución binomial con probabilidad π_i , mientras que los conteos distintos de cero siguen una distribución binomial negativa con media μ_i y parámetro de sobredispersión k .

La Binomial Negativa Inflada por Ceros es especialmente útil cuando los datos de conteo muestran una alta proporción de ceros y una sobredispersión significativa. Este enfoque permite modelar de manera efectiva esta variabilidad en los datos.

4.4.4. Modelo Hurdle Poisson (Hurdle Poisson Model)

El **Modelo Hurdle Poisson** es una técnica estadística utilizada para modelar datos de conteo con exceso de ceros en dos etapas. Fue propuesto por Mullahy (1986) y es apropiado para situaciones en las que la mayoría de las observaciones son ceros, y los ceros pueden deberse a un proceso distinto al que genera los conteos positivos.

La primera etapa del modelo consiste en modelar la probabilidad de que ocurran ceros en los datos utilizando un modelo de regresión logística. Esto se expresa matemáticamente de la siguiente manera:

$$\log \left(\frac{P(Y = 0)}{1 - P(Y = 0)} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k \quad (13)$$

Donde:

- Y es una variable aleatoria de conteo.
- $X_1, X_2, \dots, X_k$ son predictores.
- $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ son coeficientes a estimar.

En la segunda etapa, se modela el conteo condicional en los casos en que no ocurren ceros utilizando un modelo de Poisson. Esto se expresa como:

$$P(Y = y) = \begin{cases} P(Y = 0) & \text{si } y = 0 \\ (1 - P(Y = 0)) \frac{e^{-\lambda} \lambda^y}{y!} & \text{si } y = 1, 2, 3, \dots \end{cases} \quad (14)$$

Donde:

- $P(Y = y|Y > 0)$ es la probabilidad de que Y tome el valor y dado que Y es mayor que cero.
- λ es el parámetro de la distribución de Poisson, que se estima a partir de los datos.

El Modelo Hurdle Poisson permite abordar efectivamente la presencia de ceros excesivos en datos de conteo, modelando tanto la probabilidad de ceros como el conteo condicional en los casos en que no hay ceros.

4.4.5. Modelo Hurdle Binomial Negativo (Hurdle Negative Binomial Model)

El **Modelo Hurdle Binomial Negativo** es una extensión del modelo de Poisson que aborda la sobredispersión en datos de conteo y la presencia de un alto número de ceros. Este enfoque es apropiado cuando la mayoría de las observaciones son ceros y los ceros pueden deberse a un proceso distinto al que genera los conteos positivos.

La primera etapa del modelo consiste en modelar la probabilidad de que ocurran ceros en los datos utilizando un modelo de regresión logística. Esto se expresa matemáticamente de la siguiente manera:

$$\log \left(\frac{P(Y = 0)}{1 - P(Y = 0)} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k \quad (15)$$

Donde:

- Y es una variable aleatoria de conteo.
- $X_1, X_2, \dots, X_k$ son predictores.
- $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ son coeficientes a estimar.

En la segunda etapa, se modela el conteo condicional en los casos en que no ocurren ceros utilizando un modelo de Binomial Negativo. Esto se expresa como:

$$P(Y = y|Y > 0) = \binom{y + r - 1}{y} \cdot p^r \cdot (1 - p)^y, \quad y = 1, 2, 3, \dots \quad (16)$$

Donde:

- $P(Y = y|Y > 0)$ es la probabilidad de que Y tome el valor y dado que Y es mayor que cero.
- p es la probabilidad de éxito en un ensayo individual.
- r es un parámetro de dispersión que controla la variabilidad.

El Modelo Hurdle Binomial Negativo permite abordar de manera efectiva la presencia de ceros excesivos en datos de conteo y la sobredispersión al modelar tanto la probabilidad de ceros como el conteo condicional en los casos en que no hay ceros.

Este enfoque se utiliza en tu tesis para analizar y predecir la reincidencia de la población carcelaria en Colombia, teniendo en cuenta la complejidad de los datos de conteo y la presencia de ceros en la muestra.

4.5. Modelos de Recuento por Truncados en Cero

En el análisis de datos de recuento, es común encontrarse con situaciones en las que una gran proporción de observaciones tiene un valor de cero. Los Modelos de Recuento truncados en cero son una estrategia efectiva para abordar este tipo de datos, donde los ceros están presentes en la muestra debido a limitaciones en la medición o por naturaleza de los datos.

4.5.1. Poisson Truncado en Cero

El modelo de Poisson Cero Truncado es una técnica que se emplea cuando se enfrentan datos con un exceso de ceros. Forma parte de los modelos condicionales o en dos partes. En primer lugar, se modela la presencia o ausencia de ceros con un modelo de regresión logística. La probabilidad de que un valor sea cero se modela de la siguiente manera:

$$P(Y = 0) = \frac{e^{-\lambda}}{1 + e^{-\lambda}} \quad (17)$$

Donde:

- Y es la variable de respuesta de conteo.
- λ es el predictor lineal del modelo logístico.

La distribución de Poisson Truncada en Cero se define como:

$$P(Y = y) = \begin{cases} \frac{e^{-\lambda} \lambda^y}{y!} & \text{si } y > 0 \\ 0 & \text{si } y = 0 \end{cases} \quad (18)$$

La función de log-verosimilitud para el Modelo de Poisson Truncado en Cero se expresa como:

$$\mathcal{L}(\beta | \mathbf{X}, \mathbf{Y}) = \sum_{i=1}^N [Y_i \log(\lambda_i) - \lambda_i - \log(Y_i!)] \quad (19)$$

Donde:

- N es el número de observaciones.
- Y_i es la variable de respuesta para la observación i .
- $\mathbf{X}$ es el vector de variables predictoras.
- β son los coeficientes a estimar.
- λ_i es la tasa de ocurrencia estimada para la observación i , dada por $\lambda_i = e^{\mathbf{X}_i' \beta}$.

Este modelo es útil cuando se quiere modelar la probabilidad de ocurrencia de eventos contados, teniendo en cuenta la presencia de ceros y la dependencia entre la media y la ocurrencia de eventos no nulos.

4.5.2. Binomial Negativa Truncada en Cero

El modelo de Binomial Negativa Truncada en Cero es una técnica utilizada para modelar datos de conteo con exceso de ceros. Se integra dentro de los modelos de dos partes, donde la primera etapa modela la probabilidad de ceros mediante un modelo logístico, y la segunda etapa modela el conteo condicional en casos no nulos mediante una distribución de Binomial Negativa.

La probabilidad de que un valor sea cero se modela con un modelo logístico, expresado como:

$$P(Y = 0) = \frac{1}{1 + e^{-\lambda}} \quad (20)$$

Donde:

- Y es la variable de respuesta de conteo.
- λ es el predictor lineal del modelo logístico.

La distribución de Binomial Negativa Truncada en Cero se define como:

$$P(Y = y) = \begin{cases} \binom{y+r-1}{y} p^r (1-p)^y & \text{si } y > 0 \\ 0 & \text{si } y = 0 \end{cases} \quad (21)$$

La función de log-verosimilitud para el Modelo de Binomial Negativa Truncada en Cero se expresa como:

$$\mathcal{L}(\beta | \mathbf{X}, \mathbf{Y}) = \sum_{i=1}^N [Y_i \log(p) + r \log(r) + (Y_i + r) \log(\lambda_i) - \log(Y_i!)] \quad (22)$$

Donde:

- N es el número de observaciones.
- Y_i es la variable de respuesta para la observación i .
- $\mathbf{X}$ es el vector de variables predictoras.
- β son los coeficientes a estimar.
- λ_i es la tasa de ocurrencia estimada para la observación i , dada por $\lambda_i = e^{\mathbf{X}_i' \beta}$.
- p y r son parámetros de la distribución de Binomial Negativa.

Este modelo es útil cuando se desea modelar la probabilidad de ocurrencia de eventos contados, teniendo en cuenta la presencia de ceros y la dependencia entre la media y la ocurrencia de eventos no nulos.

5. Objetivos

5.1. Objetivo general

Desarrollar modelos estadísticos que permitan estimar la probabilidad y el número potencial de reincidencias entre la población carcelaria en Colombia, con el propósito de identificar a futuros reincidentes basándose en datos proporcionados por el Instituto Nacional Penitenciario y Carcelario (INPEC).

5.2. Objetivos específicos

- Calcular la tasa de reincidencia entre los reclusos del INPEC y analizar cómo varía en función de su perfil.
- Estimar la frecuencia de reincidencia para individuos con perfiles específicos.
- Evidenciar si los programas de reinserción u otros aspectos sociales tienen impacto significativo sobre la reincidencia.
- Determinar el método más adecuado para estimar los efectos de una posible reincidencia.
- Identificar el modelo predictivo más adecuado para determinar los potenciales perfiles reincidentes
- Desarrollar un Score de Reincidencia que permita identificar perfiles propensos a la reincidencia delictiva.

6. Metodología

En esta sección, se proporciona una descripción detallada de la metodología que se empleará en el estudio de la reincidencia carcelaria en Colombia. El análisis se fundamenta en el conjunto de datos suministrados por el Instituto Nacional Penitenciario y Carcelario (INPEC), con una fecha de corte que corresponde a noviembre de 2019. Estos datos conforman la piedra angular de nuestro enfoque de modelado y análisis, aportando información esencial sobre la población carcelaria del país y sus patrones de reincidencia.

Un punto crucial de nuestra metodología es la selección de la población objeto de estudio. En este sentido, se toma la decisión de excluir a los individuos sindicados y enfocar exclusivamente la atención en los condenados. Esta elección se sustenta en la consideración de que la definición de reincidencia implica la comisión de un delito por parte de alguien que ya ha sido previamente condenado. Por ende, al restringir el análisis a los individuos condenados, se logra una delimitación más precisa y coherente con el concepto de reincidencia.

Esta exclusión de los sindicados tiene un impacto en la población estudiada, lo que resulta en una reducción significativa del tamaño de la muestra. Sin embargo, esta restricción se ajusta de manera más congruente a los objetivos del estudio al enfocarse en individuos que ya han pasado por un proceso legal que culminó en una condena. Por lo tanto, se tiene la expectativa de que este trabajo contribuya a una comprensión más precisa de los patrones de reincidencia en el contexto de la población penitenciaria colombiana.

El siguiente cuadro representa el diccionario de datos que describe en detalle cada variable incluida en la base proporcionada por el Instituto Nacional Penitenciario y Carcelario (INPEC). Esta base de datos, que representa la población a cargo del INPEC a noviembre de 2019, cuenta con un total de 14 variables que describen diversas características de 111,038 individuos reclusos en diferentes establecimientos carcelarios adscritos y vigilados por el INPEC. Este diccionario es fundamental para comprender y contextualizar la información recopilada en la base de datos y se utiliza como referencia constante a lo largo del estudio. Cada variable, desde la Edad del individuo hasta la Regional del INPEC a la que está adscrito, se define de manera precisa, lo que facilita la interpretación y el análisis de los datos recopilados. Este nivel de detalle es esencial para garantizar la coherencia y la precisión en la comprensión de los patrones de reincidencia en la población penitenciaria colombiana, y sienta las bases para llevar a cabo un análisis riguroso y significativo. La tabla proporciona una descripción clara y concisa de cada variable, lo que resulta en un recurso valioso para investigadores y profesionales que trabajan en el ámbito de la justicia penal y la reincidencia delictiva.

A continuación, se procederá al ajuste de diversos modelos estadísticos con el propósito de estimar el perfil de que un condenado sea reincidente. Estos modelos se basarán en un conjunto de variables explicativas o regresoras, y abarcarán las siguientes categorías:

TABLA 1: DESCRIPCIÓN DE VARIABLES

Variablen	Edad en años del individuo
Edad	Edad en años del individuo
Fecha de ingreso	Fecha en la que ingresó al establecimiento carcelario
Ciudad de domicilio	Lugar donde vivía
Estado civil	Soltero, casado, en unión libre o viudo
Género	Género al que pertenece el individuo
Tiempo de condena	Tiempo en meses que fue condenado el individuo
Situación judicial	Estado judicial a la fecha
Nivel educativo	Nivel educativo alcanzado por el individuo
Actividad ocupacional	Actividad realizada dentro del establecimiento carcelario
Delitos	Delitos por los que es condenado
No. Reincidencias	Cantidad de reincidencias del individuo
Establecimiento carcelario	Nombre del establecimiento
Número de hijos	Cantidad de hijos del individuo
Regional	Regional del INPEC donde está interno

- Modelos de regresión binaria, que explicarán los factores determinantes para que un individuo sea reincidente o no. Estos modelos cuantificarán el impacto de cada variable en la probabilidad de reincidencia.
- Modelos de conteo cero inflados, como el modelo Hurdle, que combina los modelos Poisson y Binomial negativo para estimar la cantidad de veces que un individuo reincidiría.
- Modelos binarios, los cuales evaluarán la probabilidad de reincidencia. Debido a que las tasas de reincidencia varían en diferentes establecimientos, se aplicarán modelos no supervisados para segmentar la población carcelaria y estimar la probabilidad de reincidencia en cada segmento.

Finalmente, se realizará un proceso de backtesting o validación cruzada utilizando datos con fecha de corte en noviembre de 2020. Esta etapa permitirá verificar la precisión y la calidad de cada modelo, utilizando la población carcelaria del año 2019 para entrenar los modelos y la población del año 2020 para evaluar su rendimiento.

A continuación, se presenta una explicación de los criterios de información utilizados para la selección y evaluación de los modelos.

6.1. Criterios de información

Los criterios de información son medidas utilizadas en la estadística para comparar y seleccionar modelos estadísticos. Estos criterios se basan en principios de teoría de la información y penalizan la complejidad de los modelos, teniendo en cuenta tanto su capacidad para ajustarse a los datos como su capacidad de generalización.

La idea fundamental detrás de los criterios de información es encontrar un equilibrio entre la bondad de ajuste del modelo y su complejidad. Un modelo que se ajusta bien a los datos pero es muy complejo puede estar sobreajustado, lo que significa que se adapta demasiado a los detalles específicos del conjunto de datos de entrenamiento y puede tener un rendimiento deficiente en nuevos datos. Por otro lado, un modelo muy simple puede no capturar adecuadamente la estructura subyacente de los datos.

Los criterios de información más comunes son el AIC (Akaike's Information Criterion), el BIC (Bayesian Information Criterion) y el Hannan-Quinn. Estos criterios proporcionan una forma sistemática de evaluar

y comparar diferentes modelos, permitiendo seleccionar el modelo que mejor se ajuste a los datos y que sea más parsimonioso.

AIC (Akaike's Information Criterion): El AIC es una medida que penaliza la complejidad del modelo al evaluar su capacidad para describir y predecir los datos. Cuanto menor sea el valor del AIC, mejor será el modelo en términos de ajuste y parsimonia. El AIC se calcula como

$$AIC = -2 \cdot \log - \text{verosimilitud} + 2 \cdot \text{número de parámetros} \quad (23)$$

El término de penalización $2 \cdot \text{número de parámetros}$ ayuda a evitar la sobreajuste del modelo.

$$AIC_i = -2 \log L_i + 2p_i \quad (24)$$

BIC (Bayesian Information Criterion): El BIC es similar al AIC y también penaliza la complejidad del modelo. Sin embargo, el BIC impone una penalización más fuerte en la complejidad que el AIC. El BIC se calcula como

$$BIC = -2 \cdot \log - \text{verosimilitud} + \log(n) \cdot \text{número de parámetros} \quad (25)$$

donde n es el tamaño de la muestra. El término adicional $\log(n) \cdot \text{número de parámetros}$ penaliza más los modelos con un mayor número de parámetros en comparación con el AIC

$$BIC_i = -2 \log L_i + p_i \log(n) \quad (26)$$

Hannan-Quinn: El criterio de información de Hannan-Quinn es otra medida que penaliza la complejidad del modelo. Es similar al AIC y al BIC, pero tiene una penalización adicional para modelos con un mayor número de parámetros. Se calcula como

$$\text{Hannan} - \text{Quinn} = -2 \cdot \log - \text{verosimilitud} + 2 \cdot \log(\log(n)) \cdot \text{número de parámetros}. \quad (27)$$

$$HQ_i = -2 \log L_i + 2p_i \log(n) \quad (28)$$

6.2. Selección y Evaluación del Modelo Score

KS (Kolmogorov-Smirnov Statistic):

El Statistic de Kolmogorov-Smirnov (KS) es una métrica utilizada para evaluar la capacidad de discriminación de los modelos predictivos en el contexto de esta investigación (Doe et al., 2008). Esta medida cuantifica la máxima discrepancia entre dos distribuciones clave: la distribución de individuos que han reincidido y la distribución de aquellos que no lo han hecho. En términos sencillos, el KS busca determinar cuán efectivamente un modelo es capaz de diferenciar entre los dos grupos de interés. Un valor de KS cercano a 1 indicará un alto poder discriminativo, lo que sugiere que el modelo puede distinguir de manera precisa entre los reincidentes y no reincidentes. Por otro lado, un valor cercano a 0 señalará una capacidad limitada del modelo para llevar a cabo esta distinción.

Gini Index:

El índice de Gini es otra métrica relevante utilizada en el análisis de los modelos predictivos (Smith - Johnson, 2004). Este índice es fundamental para evaluar la calidad de las predicciones realizadas por un modelo. Su función es medir la capacidad del modelo para clasificar correctamente a los individuos que han reincidido, en comparación con aquellos que no lo han hecho. El índice de Gini varía en una escala de 0 a 1, donde un valor de 1 indica una predicción perfecta, es decir, el modelo es capaz de separar de manera ideal a los reincidentes de los no reincidentes. Por otro lado, un valor de 0 implica una predicción aleatoria.

ROC (Receiver Operating Characteristic) Curve:

La Curva ROC es una herramienta gráfica de gran utilidad para evaluar el desempeño de los modelos predictivos en la distinción de individuos reincidentes y no reincidentes (Brown - Smith, 2016). Esta curva muestra la relación entre la tasa de verdaderos positivos (sensibilidad) y la tasa de falsos positivos (complemento de especificidad) para diferentes umbrales de clasificación utilizados por el modelo. En otras palabras, cuantifica cómo de bien el modelo es capaz de distinguir entre los dos grupos de interés. Cuanto más se acerque la curva ROC al área bajo la curva (AUC) a 1, mayor será la capacidad predictiva del modelo, indicando una mejor habilidad para diferenciar a los reincidentes de los no reincidentes.

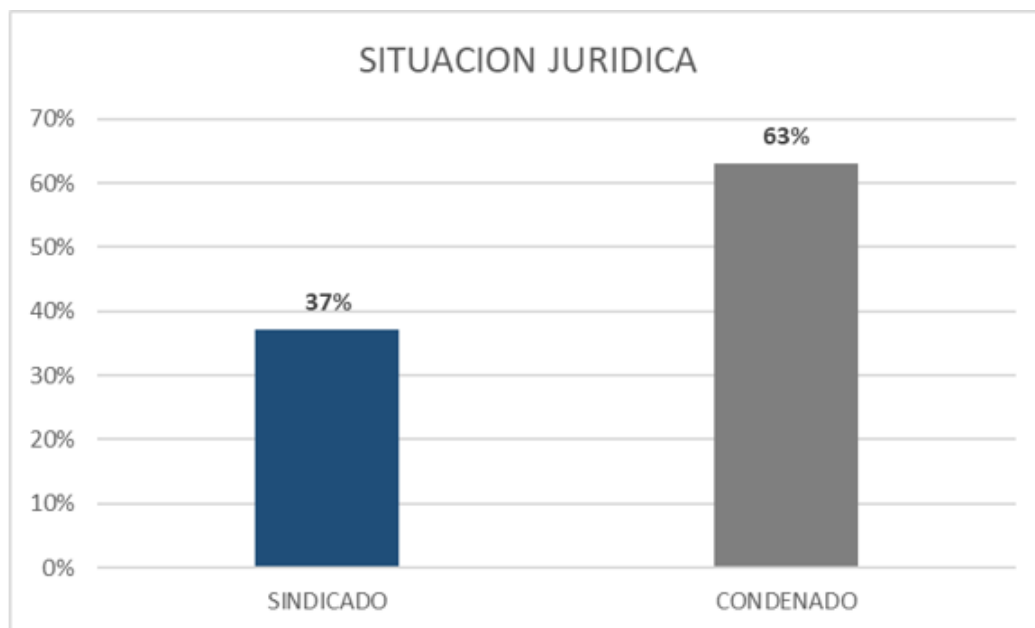
En este trabajo de grado, estas métricas jugarán un papel fundamental en la evaluación crítica de los modelos probabilísticos desarrollados. Permitirán a los investigadores determinar cuál de los modelos es más efectivo para identificar a los individuos con perfiles propensos a la reincidencia delictiva en la población carcelaria de Colombia. Estas métricas ofrecerán una visión objetiva del rendimiento de los modelos, facilitando la toma de decisiones informadas en función de la calidad de las predicciones.

Tras el ajuste de los modelos probabilísticos, se procedió a la creación de un modelo Score para estimar los perfiles propensos a la reincidencia delictiva en la población carcelaria. Para evaluar y comparar la calidad de estos modelos Score, se emplean métricas clave como el Statistic de Kolmogorov-Smirnov (KS), el índice de Gini y la Curva ROC (Receiver Operating Characteristic). Estas métricas nos proporcionan una visión objetiva del rendimiento de los modelos y ayudan a tomar decisiones informadas en función de la calidad de nuestras predicciones.

6.3. Análisis de Datos

Antes de entrar en el análisis detallado, es importante aclarar la población de estudio y las exclusiones aplicadas en este análisis. La base de datos proporcionada por el INPEC contiene información sobre un total de 192,019 individuos, representando la población a cargo del INPEC hasta noviembre de 2019. Sin embargo, no todos estos individuos tienen una situación jurídica definida. Algunos de ellos se identifican como 'sindicalizados', lo que indica que su situación legal está pendiente. Para fines de este ejercicio de análisis, hemos decidido excluir a estos individuos de nuestra muestra, ya que no podemos definir con certeza su reincidencia en ausencia de una condena previa.

FIGURA 1: GRAFICO DE LA SITUACION JURIDICA



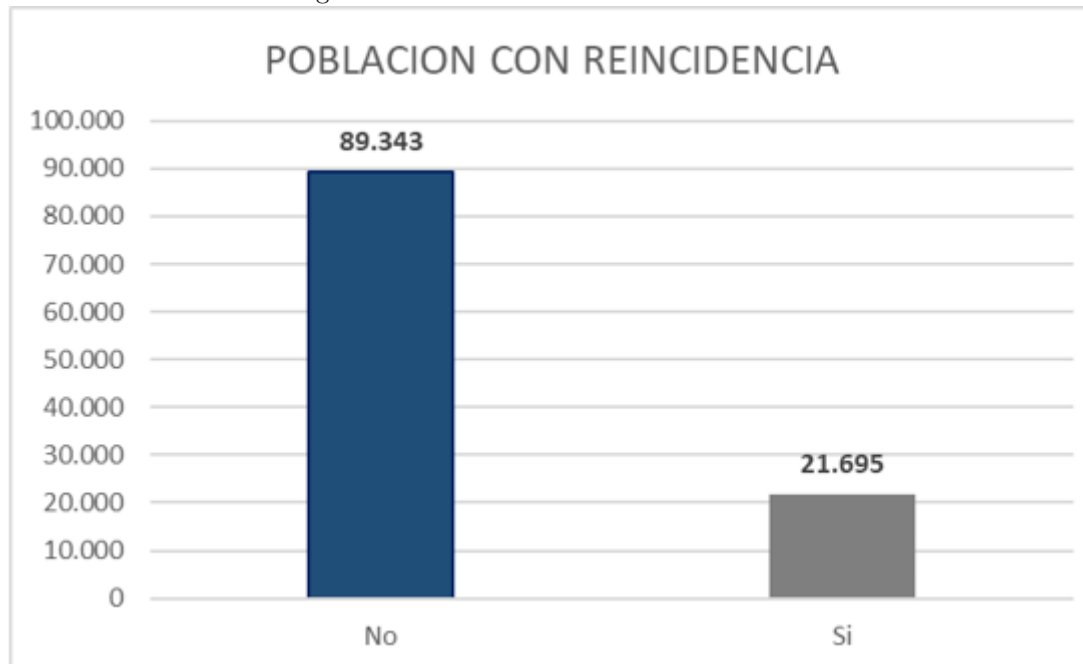
Fuente: INPEC Nov 2019

En la Figura 1, se observa que, de la población carcelaria total bajo la supervisión del INPEC, solo el 63 % de los individuos tiene su condena definida, mientras que el 37 % restante se encuentra en la categoría de sindicados.

De acuerdo con esto la base se disminuye significativamente, así mismo se realiza un análisis de calidad de datos, donde se excluyen individuos que no tengan al menos tres variables y/o presenten datos atípicos o inconsistencia en los registros.

Una vez definida la calidad de la información y teniendo solo población condena, se trabaja con una base de 111.038 individuos, con información estructurada y necesaria para modelar y hacer inferencia de esta. Se puede observar que la población reincidente tiene poca participación en la población total recluida en periodos anteriores a 2013 (teniendo en cuenta el año de ingreso de la población actual), pero es evidente que es creciente a través del tiempo.

Figura 2: POBLACION CON REINCIDENCIA

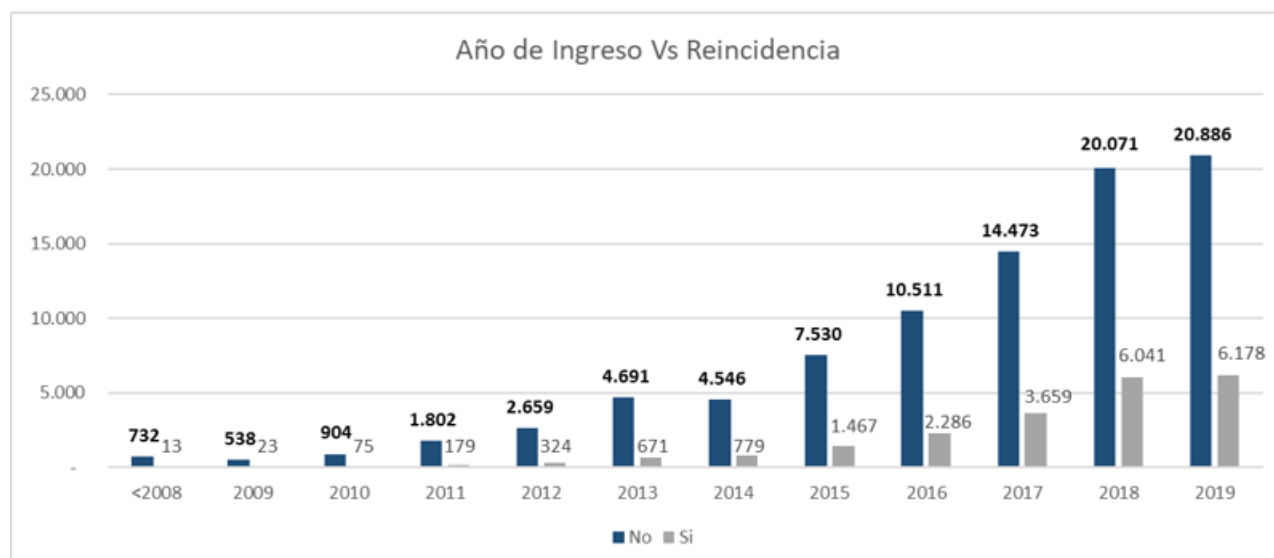


Fuente: INPEC Nov 2019

En la figura 2 se puede evidenciar que tan solo el 20 % de los individuos a estudiar han sido reincidentes y la población no reincidente presenta una gran mayoría (80 %)

Dado que la población presenta una participación importante de individuos reincidentes, ya que un 20 % es una cifra considerable y teniendo en cuenta que uno de los objetivos de INPEC y el sistema carcelario colombiano es resocializar a los individuos con el fin de que los individuos vuelvan a la sociedad y no reincidan en actividades delictivas.

FIGURA 3: GRAFICO DE BARRAS DE AÑO DE INGRESO VS REINCIDENCIA



Fuente: INPEC Nov 2019

En la Figura se Evidencia la evolución a través del tiempo del ingreso de la población carcelaria discriminado por cantidad de población reincidente, viendo así el importante crecimiento de la población a medida que pasa el tiempo, pero también el aumento que se percibe en la población reincidente

Se evidencia que la mayor parte de la población a ingresado en los últimos años, y que la reincidencia viene aumentando significativamente cada año, esto puede deberse a que la población del país va creciendo y así mismo su población carcelaria.

A continuación, analizaremos diferentes características de la población carcelaria, de las cuales el INPEC compartió para realizar el análisis.

TABLA 2: GENERO VS REINCIDENCIA.

Genero	No	Si	Total	Total reinci Genero
Femenino	9.062	1.393	10.455	13 %
Masculino	80.281	20.302	100.583	20 %
Total general	89.343	21.695	111.038	20 %

Fuente: INPEC Nov 2019

El 13 % de las mujeres tienden a reincidir, no obstante, el 20 % de los hombres reinciden, por lo que vemos que al ser hombre aumenta la probabilidad de reincidencia levemente. Y analizando la generalidad de la población tan solo el 20 % de la población que actualmente cumple una condena es reincidente.

Al segmentar la población por género, se evidencia que la cantidad de reclusos de género masculino es casi diez veces más grande que la población femenina, y así mismo la participación de individuos de género masculino en calidad de reincidencia es mucho mayor ya que los individuos de género femenino tan solo tienen un porcentaje de reincidencia del 13 %.

TABLA 3: ESTADO CIVIL VS REINCIDENCIA.

Estado Civil	Canti. No	Canti. Si	Tot. Poblacion	Porcen. Total	Porcen. Reincidente
UNION LIBRE	103	12.686	12.789	11,5 %	58,5 %
SOLTERO(A)	106	7.746	7.852	7,1 %	35,7 %
CASADO(A)	18	861	879	0,8 %	4,0 %
SEPARADO(A)	5	282	287	0,3 %	1,3 %
VIUDO(A)	3	86	89	0,1 %	0,4 %
sin Dato	89.106	15	89.121	80,3 %	0,1 %
DIVORCIADO(A)	-	12	12	0,0 %	0,1 %
PENDIENTE POR	1	6	7	0,0 %	0,0 %
RELIGIOSO(A)	1	1	2	0,0 %	0,0 %
Total	89.343	21.695	111.038	100 %	100 %

Fuente: INPEC Nov 2019

Se podría afirmar que el 80 % de los individuos no informan de su estado civil, y con el restante de los individuos que informa se evidencia que en su mayoría son reincidentes, lo que da fuerza a la variable para mantenerla.

En la tabla 3 se puede observar que la población reincidente se concentra en individuos con estado civil de unión libre o solteros, por lo cual, en el análisis de regresión se validará si esta característica aporta en forma relevante a un perfil de los individuos reclusos reacios a la reincorporación a la sociedad, ya que en solo estas dos condiciones se concentra el 94.2 % de la población reincidente.

TABLA 4: ESTADO INGRESO VS REINCIDENCIA.

ESTADO INGRESO	Canti. No	Canti. Si	Tot. Poblacion	Porcen. Total	Porcen. Reincidente
INTRAMUROS	60.032	17.552	77.584	69,9 %	80,9 %
PRISION DOMICILIARIA	25.458	-	25.458	22,9 %	0,0 %
VIGILANCIA ELECTRONICA	3.153	605	3.758	3,4 %	2,8 %
DETENCION DOMICILIARIA	357	-	357	0,3 %	0,0 %
ESPERA TRASLADO	343	-	343	0,3 %	0,0 %
DOMICILIARIA	-	3.538	3.538	3,2 %	16,3 %
Total	89.343	21.695	111.038	100 %	100 %

Fuente: INPEC Nov 2019

En la tabla 4 se evidencia que aproximadamente el 30 % de los individuos a cargo del INPEC se encuentran fuera de una medida intramural, a su vez es importante resaltar que el 81 % de la población reincidente tiene medida intramural.

El estado de los internos ya que no todos los individuos están 24 horas en centros de reclusión, muchos de ellos están bajo prisión domiciliaria, aproximadamente el 30 % de la población carcelaria a cargo del INPEC.

En la Tabla 5 al segmentar por rangos de años de condena y comparar con la reincidencia, se identifica que los individuos con menor rango de condena son más reincidentes, así mismo se evidencia que la población carcelaria a cargo del INPEC tiene mayor participación en condenas menores o iguales a 10 años.

Se resalta la participación reincidente en poblaciones con condenas altas, ya que, podría decirse, son menos propensas a reincidir, por otro lado, las poblaciones con condenas de uno a cinco años son significativamente más predispuestas a reincidir, dado que el 48 % de la población es población reincidente, y esta población represente el 40 % de la población total condenada del INPEC.

TABLA 5: RANGO AÑOS DE CONDENA VS REINCIDENCIA.

Rango Años de Condena	Canti. No	Canti. Si	Tot. Poblacion	Porcen. Total	Porcen. Reincidente
<1	470	240	710	1 %	1 %
1-5	33.732	10.461	44.193	40 %	48 %
6-10	22.742	5.938	28.680	26 %	27 %
11-20	21.392	3.444	24.836	22 %	16 %
21-35	8.053	1.229	9.282	8 %	6 %
>36	2.954	383	3.337	3 %	2 %
Total	89.343	21.695	111.038	100 %	100 %

Fuente: INPEC Nov 2019

TABLA 6: NIVEL EDUCATIVO VS REINCIDENCIA

NIVEL EDUCATIVO	Canti. No	Canti. Si	Tot. Poblacion	Porcen. Total	Porcen. Reincidente
BASICA MEDIA Y VOCACIONAL	50.961	13.118	64.079	57,7 %	60,5 %
BASICA PRIMARIA	29.759	7.371	37.130	33,4 %	34,0 %
EDUCACION SUPERIOR	4.253	371	4.624	4,2 %	1,7 %
ILETRADOS	4.370	835	5.205	4,7 %	3,8 %
Total	89.343	21.695	111.038	100,0 %	100,0 %

Fuente: INPEC Nov 2019

En la tabla 6 al clasificar la población carcelaria de acuerdo con su nivel educativo, se evidencia que los iletrados tienen menor participación de reincidencia que los reclusos con educación en básica primaria y media vocacional, así mismo se observa que los reclusos con educación superior son los que menos participación en reincidencia tienen.

La mayoría de la población presenta estudios de educación Media, en otras palabras, culminaron sus estudios en el colegio, así mismo esta población presenta una reincidencia del 60 %. Por otro lado, los iletrados tan solo presentan una tasa del 3.8 %, una tasa muy inferior.

TABLA 7: ACTIVIDAD VS REINCIDENCIA

	No	Si	Total	Porcen. total	Porcen. Reincidente
ENSEÑANZA	1.255	134	1.389	1,3 %	0,6 %
ESTUDIO	26.245	8.168	34.413	31,0 %	37,6 %
sin dato	32.659	5.851	38.510	34,7 %	27,0 %
TRABAJO	29.184	7.542	36.726	33,1 %	34,8 %
Total	89.343	26.695	111.038	100,0 %	100,0 %

Fuente: INPEC Nov 2019

En la tabla 7 al clasificar la población carcelaria de acuerdo con su actividad, se evidencia que los de actividad de enseñanza tienen menor participación de residencia que los reclusos con estudio y trabajo, así mismo se observa que la mayor participación de reincidencia son los que tienen una actividad de estudio.

La mayoría de la población tiene una actividad en el transcurso de su condena ya que el 65 % de ellos

trabaja, estudia o educa. En la tabla existe la categoría de “sin Dato” lo que refiere a que no realizan ninguna actividad o cumplen su condena fuera de un centro intramural. Así mismo es importante resaltar que los individuos que no tienen una tarea en la cárcel han sido menos reincidentes que el resto de los sujetos.

TABLA 8: CONDUCTA VS REINCIDENCIA.

Conducta Vs Reincidencia	Canti. No	Canti. Si	Tot Poblacion	Porcen. Total	Porcen. Reincidente
Buena	-	4.281	4.281	3,9 %	19,7 %
Ejemplar	-	11.989	11.989	10,8 %	55,3 %
Mala	-	555	555	0,5 %	2,6 %
Pendiente	-	249	249	0,2 %	1,1 %
Regular	-	440	440	0,4 %	2,0 %
sin Dato	89.343	4.181	93524	84,2 %	19,3 %
Total	89.343	21.695	111.083	100,0 %	100,0 %

Fuente: INPEC Nov 2019

En la tabla 8, se clasifica la población carcelaria de acuerdo con su conducta durante la retención, se puede evidenciar, que los reclusos con una mayor participación en la reincidencia son aquellos que tienen una conducta ejemplar y también se tiene observa una menor participación en la reincidencia de los reclusos con conducta pendiente.

Se puede observar encuentra que sólo los individuos reincidentes presentan una calificación de conducta, dado que esta valoración se otorga una vez termine su condena, así mismo es importante recalcar que los individuos a los que se les otorgo una calificación “Ejemplar” son los individuos que más reincidieron, seguidos por los que se les valoró con una conducta “Buena” que es la segunda mejor calificación que puede tener un interno.

TABLA 9: REGIONALL VS REINCIDENCIA.

	No	Si	Tot. Población	Porcen. Total	Porcen. Reincidente
CENTRAL	31.903	8.283	40.186	36 %	21 %
NOROESTE	11.637	2.787	14.424	13 %	19 %
NORTE	9.355	1.765	11.120	10 %	16 %
OCCIDENTE	18.087	3.886	21.973	20 %	18 %
ORIENTE	8.186	2.141	10.327	9 %	21 %
VIEJO CALDAS	10.175	2.833	13.008	12 %	22 %
Total	89.343	21.695	111.038	100 %	20 %

Fuente: INPEC Nov 2019

En la tabla 9, se clasifican la población carcelaria de acuerdo con su región, donde se puede evidenciar que la región central tiene la mayor participación (38 %) de reincidencia, mientras que la menor es la de los reclusos de la región del norte (8 %).

Se evidencia que la regional central, por acumular la mayor parte de la población penitenciaria, concentra la mayor parte de los individuos que reinciden, así mismo es importante analizar poblaciones como norte (8 %) u Oriente (10 %) que son poblaciones con poca participación en la reincidencia carcelaria.

6.4. Categorización de Variables

En esta sección, nos centraremos en la categorización de variables esenciales para el análisis de la reincidencia carcelaria en Colombia. Dos de las variables clave abordadas son los delitos cometidos y el estado civil de los individuos en estudio. La categorización de estas variables desempeña un papel fundamental al permitir un análisis más efectivo y al abordar desafíos potenciales, como la presencia de numerosos delitos o categorías de estado civil con una población reducida. A través de esta categorización, se busca optimizar la aplicabilidad de los modelos estadísticos, lo que a su vez facilitará un enfoque más preciso y significativo en el estudio de la reincidencia carcelaria en el contexto colombiano.

En la población analizada se puede observar una gran variedad de delitos cometidos por los internos, la mayoría de ellos están condenado por más de un delito, y muchos de ellos están correlacionados y son causa de mismo acto, ya que un solo acto delincuenciales puede infringir la ley de diferentes formas, dependiendo de la naturaleza de la acción y las leyes aplicables en un determinado sistema legal. Algunos actos pueden infringir múltiples disposiciones legales y por lo tanto, dar lugar a la imputación de varios delitos.

Teniendo en cuenta lo anterior podemos caracterizar cada delito y así mismo encontrar en efecto en reincidencia para cada individuo, esta caracterización se realiza con base en el Código Penal Colombiano, ya que El Código Penal Colombiano define y tipifica una amplia gama de delitos, que van desde delitos contra la vida, la integridad personal y la libertad, hasta delitos contra el patrimonio, la seguridad pública, la administración pública, el orden económico y social, entre otros. También establece las penas y las medidas de seguridad aplicables a cada delito.

Los delitos contemplados en el Código Penal Colombiano se organizan en diferentes títulos. El código está dividido en varios libros, cada uno de los cuales abarca una temática específica. Cada libro se compone de títulos, y estos títulos agrupan los delitos relacionados.

TABLA 10: TABLA DE DELITOS SEGÚN EL CÓDIGO PENAL COLOMBIANO

Título	Delitos Relacionados
Título 1	Delitos contra la vida y la integridad personal
Título 2	Delitos contra las personas y bienes protegidos por el derecho
Título 3	Delitos contra la libertad individual y otras garantías
Título 4	Delitos contra la libertad, integridad y formaciones sexuales
Título 5	Delitos contra la integridad moral
Título 6	Delitos contra la familia
Título 7	Delitos contra el patrimonio económico
Título 8	De los delitos contra los derechos de autor
Título 9	Delitos contra la fe pública
Título 10	Delitos contra el orden económico social
Título 11	De los delitos contra los recursos naturales y el medio ambiente
Título 12	Delitos contra la seguridad pública
Título 13	De los delitos contra la salud pública
Título 14	Delitos contra mecanismos de participación democrática
Título 15	Delitos contra la administración pública
Título 16	Delitos contra la eficaz y recta impartición de justicia
Título 17	Delitos contra la existencia y seguridad del estado
Título 18	De los delitos contra el régimen constitucional legal
Título 19	Disposiciones generales

Fuente: Código Penal Colombiano (Ley 599 de 2000)

En la tabla 10, se pueden observar los títulos con sus delitos relacionados

Teniendo en cuenta lo anterior validamos la agrupación de cada uno de los delitos.

TABLA 11: TABLA DE DELITOS VS REINCIDENCIA.

Título Vs Reincidencia	Canti. No	Canti. Si	Tot. Población	Porcen. Total	Porcen. Reincidente
TITULO 1	23.025	4.229	27.254	23 %	18 %
TITULO 2	425	51	476	0 %	0 %
TITULO 3	6.194	1.297	7.491	6 %	6 %
TITULO 4	11.373	757	12.130	10 %	3 %
TITULO 5	13	5	18	0 %	0 %
TITULO 6	2.319	446	2.765	2 %	2 %
TITULO 7	20.670	9.699	30.369	26 %	41 %
TITULO 8	11	2	13	0 %	0 %
TITULO 9	1.642	436	2.078	2 %	2 %
TITULO 10	2.121	479	2.600	2 %	2 %
TITULO 11	149	9	158	0 %	0 %
TITULO 12	19.329	4.672	24.001	21 %	20 %
TITULO 13	1.322	292	1.614	1 %	1 %
TITULO 14	8	1	9	0 %	0 %
TITULO 15	1.914	358	2.272	2 %	2 %
TITULO 16	1.911	743	2.654	2 %	3 %
TITULO 17	2	-	2	0 %	0 %
TITULO 18	394	36	430	0 %	0 %
TITULO 19	-	-	-	0 %	0 %
Total				100,0 %	100,0 %

Fuente: Código Penal Colombiano (Ley 599 de 2000)

En la tabla 11, se puede observar que los condenados en promedio son más veces imputados por el título número 7, el cual hace referencia a los delitos contra el patrimonio económico, cabe resaltar que un individuo puede ser juzgado por más de un delito, por lo que dentro de este título pueden existir varios delitos relacionados. Por lo que en la tabla no se calcula un total dado que superaría la cantidad de la población total.

Algunos de los títulos presentan una participación extremadamente baja en la población, como es el caso de los títulos 2, 5, 8, 11, 17, 18 y 19. Debido a esta limitada representación, hemos optado por agrupar estos títulos con aquellos que comparten temáticas similares, tomando como referencia el Código Penal Colombiano. Del mismo modo, hemos aplicado un enfoque de agrupación al estado civil de los individuos, debido a la baja participación en algunas categorías.

El análisis de la población carcelaria revela una diversidad de estados civiles, aunque algunos de ellos muestran una participación significativamente baja en la muestra, como se detalla en la Tabla 12. Categorías como 'divorciado(a)', 'pendiente por religioso(a)', 'sin dato', 'unión libre' y 'viudo(a)' muestran un número reducido de individuos en comparación con otras categorías más comunes, como 'casado(a)' y 'soltero(a)'. Dado que el análisis estadístico puede verse afectado por categorías con una baja población, se ha optado por agrupar estos estados civiles en categorías más amplias. Por ejemplo, 'divorciado(a)' y 'viudo(a)' se han agrupado en la categoría 'Separados', 'pendiente por religioso(a)' y 'sin dato' en la categoría 'Otros', y 'unión libre' junto con 'casado(a)' en la categoría 'Pareja'. Esta agrupación permitirá un análisis más robusto y significativo de los efectos del estado civil en la reincidencia carcelaria en el contexto colombiano.

TABLA 12: TABLA DE AGRUPACION ESTADO CIVIL

Estado Civil	Cantidad	Agrupacion
CASADO(A)	879	Pareja
DIVORCIADO(A)	12	Separados
PENDIENTE POR	7	Otros
RELIGIOSO(A)	2	Otros
SEPARADO(A)	287	Separados
sin Dato	89.121	Otros
SOLTERO(A)	7.852	Soltero
UNION LIBRE	12.789	Pareja
VIUDO(A)	89	Separados

Fuente: INPEC Nov 2019

En la tabla 12, se distribuye la población penitenciaria de acuerdo a su estado civil, debiéndose agrupar de la siguiente forma: En pareja; aquellos que se encuentran casados y en unión libre. Separados; los que se clasifican en divorciados, separados y viudos. En el grupo de otros van; Pendientes, religioso y sin dato. Finalmente, el grupo de solteros permanece igual. Lo anterior se debe a que en el momento de realizar el muestreo y el entrenamiento de diferentes modelos truncados puede que la ausencia de alguno de estas categorías afecte la estimación.

6.5. Análisis de Asociación y Correlación: Correlación de V de Cramer.

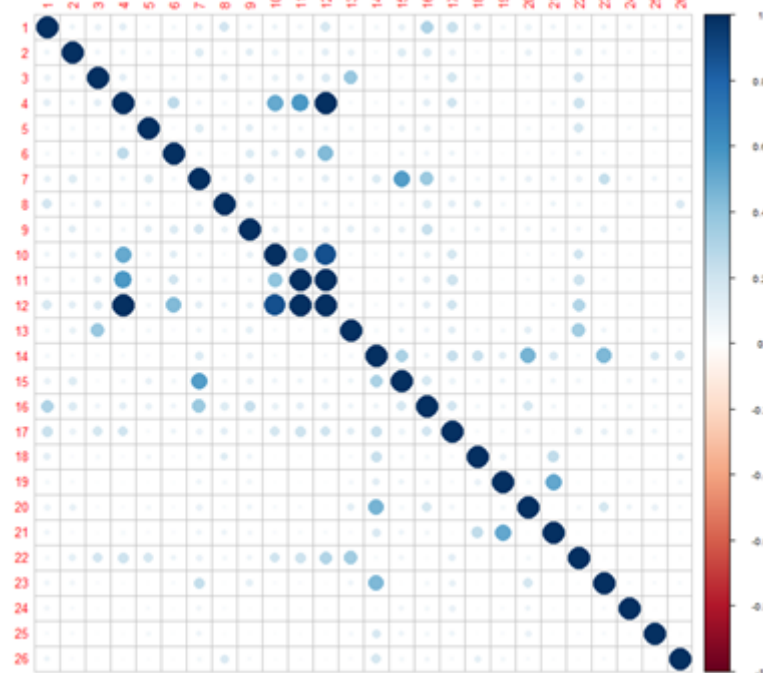
En las secciones anteriores, se estableció una sólida base metodológica para el análisis de las relaciones entre variables clave en el estudio. En esta sección, se llevará a cabo un análisis de datos, centrándose en la aplicación de la correlación de V de Cramer. Esta técnica resulta fundamental para explorar y cuantificar la fuerza de las relaciones entre variables categóricas, lo que enriquecerá la comprensión de la dinámica de la reincidencia carcelaria en Colombia. Mediante el empleo de la correlación de V de Cramer, se identificarán patrones significativos y asociaciones que aportarán valiosos hallazgos y conclusiones en este estudio.

TABLA 13: NOMBRES DE LAS VARIABLES

1	EDAD	2	Año ingreso	3	Domicilio	4	ESTADO CIVIL	5	GENERO
6	ESTADO INGRESO	7	Categori.años_condena	8	Nivel educativo	9	ACTIVIDAD OCUPACIONAL	10	ULTIMA CALIFICACION
11	No.REINCIDENCIAS	12	REINCIDENCIAS	13	Regional	14	Cantidad de delitos	15	Titulo1
16	Titulo4	17	Titulo7	18	Titulo9	19	Titulo10	20	Titulo12
21	Titulo16	22	ESTABLECIMIENTO	23	Titulo2y3	24	Titulo5_6y8	25	Titulo11_13y14
26	Titulo15_17y18								

En la tabla 13 se enumeran e identifican las variables que se utilizan para realizar el análisis de asociación. Se evidencia la presencia de 26 variables, con las cuales se aplica la correlación de V Cramer entre ellas, con el fin de evidenciar las asociaciones presentes y evitar multicolinealidad en los ejercicios de regresión.

FIGURA 4: GRAFICA COEFICIENTE DE CORRELACIÓN DE V Cramer



Fuente: INPEC Nov 2019

En la figura 4 se observa en la matriz de correlación utilizando el coeficiente de correlación de V Cramer, Este gráfico permite analizar la relación lineal entre pares de variables en un conjunto de datos. A mayor intensidad del color ya sea Azul (Correlación directa), o Rojo (Correlación inversa), mayor es la correlación, los colores con poca intensidad denotan una correlación baja entre las variables numeradas horizontal y verticalmente, por eso la diagonal es completamente azul y con un círculo que llena completamente el cuadro correspondiente, pues se refiere a la correlación de cada variable consigo misma. El tamaño del círculo también determina una medida de la correlación entre las variables relacionadas.

En la Figura 4 se identifica que las variables Reincidencia (12) y estado civil (4) están altamente correlacionadas al 99 %, debido a que la mayoría de los individuos no presentan dato de estado civil y así mismo por la naturaleza de los datos (donde usualmente no hay reincidentes), la correlación estima una alta asociación entre estado civil y Reincidencia. Por otro lado, vemos una alta correlación también con la variable reincidencia (12) y Ultima calificación (10), del 92 %, esto es debido a como vimos en la tabla 8, solo los individuos con reincidencia tienen calificación dado que los que están cumpliendo condena por primera vez no tienen una última calificación. Y por ultimo las variables Numero de Reincidencia (11) y reincidencia (12) donde su correlación es natural, dado que una implica la otra. Del resto de las variables podemos ver aparente normalidad y correlaciones moderadas.

Para llevar a cabo el análisis, se utilizó el entorno de desarrollo RStudio en su versión 4.2. Se emplearon diversas librerías para realizar operaciones específicas. La lectura y manipulación de datos se realizó con el uso de las librerías `readxl` y `dplyr`, mientras que la exploración de la correlación de variables categóricas se llevó a cabo con `vcd` y el comando `assocstats`. Para balancear las muestras, se aplicó la librería `DMwR`, y se utilizaron comandos como `AIC()`, `BIC()`, y `HQIC()` de la librería `ICglm`. Asimismo, se implementaron técnicas de Elastic Net utilizando `glmnet` y se realizaron análisis de binomiales negativas y técnicas de hurdle con las librerías `MASS` y `pscl`, respectivamente. Estos componentes tecnológicos se combinaron para efectuar un análisis exhaustivo y avanzado de los datos, permitiendo la aplicación de modelos predictivos y técnicas estadísticas especializadas.

Estas elecciones tecnológicas se realizaron con el objetivo de garantizar un análisis preciso, reproducible

y robusto, aprovechando las funcionalidades específicas de las librerías y comandos mencionados.

7. Análisis de Resultados

Esta sección se dedica al análisis de los resultados derivados de los modelos previamente mencionados. Se comenzará presentando los resultados de los modelos probabilísticos, los cuales son insumo para calcular un score de reincidencia para cada individuo. Posteriormente, se continuará con los modelos de conteo, cuyo objetivo es estimar la cantidad de veces que un individuo podría reincidir delictivamente.

7.1. Modelos de Probabilidad de Reincidencia

En esta subsección, se presentan los resultados de los modelos probabilísticos que buscan calcular el score de reincidencia para cada individuo en la población carcelaria de Colombia. Estos modelos, como el Modelo Logit, el Modelo Probit y el Modelo Elastic Net, desempeñan un papel fundamental en nuestro análisis al proporcionar insights sobre las relaciones entre diversas variables predictoras y la probabilidad de reincidencia. El Modelo Logit, en particular, se utiliza como punto de partida, seguido por el Modelo Probit y el Modelo Elastic Net para explorar diferentes enfoques y evaluar sus resultados en comparación con el Modelo Logit. Estos modelos probabilísticos sientan las bases para una comprensión más profunda de las dinámicas que influyen en la reincidencia carcelaria en la población estudiada.

7.1.1. Modelo Logit.

En esta sección se avanza hacia el modelado de la probabilidad de reincidencia, un paso fundamental en el análisis. Se utilizó un modelo Logit como punto de partida para examinar cómo diferentes variables influyen en la probabilidad de reincidencia de los individuos en la población carcelaria en Colombia. Este primer modelo Logit proporciona información valiosa sobre las relaciones entre las variables predictoras y la reincidencia. Más adelante en esta sección, se exploró un modelo Probit y Modelo Elastic Net, ampliando la perspectiva y evaluando sus resultados en comparación con el modelo Logit.

Los tres modelos incluidos en la tabla son el Modelo Saturado (con todas las variables relacionadas en la tabla 1, pag. 17), el Modelo Nulo (solo con el intercepto), y el Modelo Optimizado (el de mejor devianza y calculado con el método paso a paso).

TABLA 14: MODELO LOGIT

Modelos	AIC	BIC	Hannan-Quinn
Saturado	9.831,93	10.276,45	9.968,36
Nulo	76.605,54	76.614,80	76.608,38
Optimizado	9.825,67	10.214,33	9.944,76
Balanceado Optimizado	25.873,12	26.322,98	26.009,28

Fuente: INPEC Nov 2019

La Tabla 14 presenta los resultados de los diversos criterios de información, indicando que el modelo Optimizado muestra ligeramente mejores desempeños en comparación con el modelo Saturado, y significativamente mejores que los modelos Optimizados y Balanceado Optimizado. Estos criterios reflejan la calidad del ajuste de cada modelo, destacando la superioridad relativa del modelo Optimizado en términos de AIC, BIC y Hannan-Quinn.

El modelo Optimizado se obtuvo mediante un proceso de optimización basado en el análisis de deviance. Se realizó una comparación con el modelo Saturado y el modelo Nulo. En el análisis de deviance, se

emplearon modelos Logit para evaluar la adecuación del modelo Optimizado en relación con el modelo Nulo, que considera únicamente la intersección, y el modelo Saturado, que incluye todas las variables disponibles.

Involucra la creación de dos modelos, uno que abarca todas las variables disponibles, y otro, que solo considera la intersección. Luego, se utiliza la función `step()` para realizar un análisis hacia adelante y hacia atrás, con el objetivo de identificar el modelo más ajustado. El resultado final es el modelo Optimizado, que se compara con los demás en términos de deviance.

Es importante destacar que estos modelos son comparables porque se ajustan a la misma población y comparten la misma función de enlace (Logit), lo que facilita una evaluación coherente y significativa de su desempeño.

Para validar los modelos y garantizar su eficacia en distintas poblaciones, se llevaron a cabo diversas pruebas, incluyendo la aplicación del modelo a la base de datos que no se utilizó para su entrenamiento. Esto tenía el propósito de asegurar que el prototipo no solo fuera efectivo en la muestra de entrenamiento original, sino que también mantuviera su poder predictivo en otras poblaciones, sin una disminución significativa en su capacidad de pronosticar la reincidencia.

Validando Modelo optimizado Tipo Score para el modelo Logit

Tabla 15: Modelo Optimizado (Validacion de Score Poblacion de Entrenamiento)

ROC	KS	GINI
0,888	0,606	0,78

Decil	Total	Reincidentes	No Reincidentes	PorcentajeTotal	TasaRein	ODDS	NoRein.Acum	ReinAcum	KS	Roc
1	7.773	5.915	1.858	10%	76%	0,31	0,03	0,39	0,36	0,02
2	7.772	3.498	4.274	10%	45%	1,22	0,10	0,62	0,52	0,03
3	7.772	2.409	5.363	10%	31%	2,23	0,18	0,78	0,60	0,03
4	7.773	1.605	6.168	10%	21%	3,84	0,28	0,89	0,61	0,02
5	7.773	966	6.807	10%	12%	7,05	0,39	0,95	0,56	0,01
6	7.772	498	7.274	10%	6%	14,61	0,51	0,99	0,48	0,00
7	7.773	217	7.556	10%	3%	34,82	0,63	1,00	0,37	0,00
8	7.772	15	7.757	10%	0%	517,13	0,75	1,00	0,25	0,00
9	7.773	0	7.773	10%	0%	Inf	0,88	1,00	0,12	0,00
10	7.773	0	7.773	10%	0%	Inf	1,00	1,00	0,00	0,00

0 %	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %	100 %
0	444	635	764	845	905	948	982	1000	1000	1000

Fuente: INPEC Nov 2019

En la tabla 15 se presenta una validación del score, para eso se segmenta la población con la que entreno el modelo optimizado por deciles, dejando en los primeros deciles los individuos que obtuvieron un score más bajo y en los últimos deciles a los scores más altos, evidenciando que los scores más bajos acumulan la mayor cantidad de posibles reincidentes. Donde se evidencia que el primer decil, correspondiente a los scores más bajos, tiene una tasa de reincidencia del 76 %, mientras que en los deciles 8,9 y 10 no presentan tasa de reincidencia alguna.

TABLA 16: MODELO OPTIMIZADO (VALIDACION DE SCORE POBLACION DE PRUEBA)

ROC	KS	GINI
0.888	0,608	0.776

Decil	Total	Reincidentes	No Reincidentes	PorcentajeTotal	TasaRein	ODDS	NoRein.Acum	ReinAcum	KS	Roc
1	3.332	2.577	755	10%	77%	0,29	0,03	0,39	0,37	0,02
2	3.330	1.495	1.835	10%	45%	1,23	0,10	0,62	0,53	0,03
3	3.331	1.047	2.284	10%	31%	2,18	0,18	0,78	0,60	0,03
4	3.332	694	2.638	10%	21%	3,80	0,28	0,89	0,61	0,02
5	3.331	418	2.913	10%	13%	6,97	0,39	0,95	0,56	0,01
6	3.331	213	3.118	10%	6%	14,64	0,51	0,99	0,48	0,00
7	3.331	96	3.235	10%	3%	33,70	0,63	1,00	0,37	0,00
8	3.331	5	3.326	10%	0%	665,20	0,75	1,00	0,25	0,00
9	3.331	0	3.331	10%	0%	Inf	0,88	1,00	0,12	0,00
10	3.332	0	3.332	10%	0%	Inf	1,00	1,00	0,00	0,00

0 %	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %	100 %
0	433	632	762	847	905	948	983	1000	1000	1000

Fuente: INPEC Nov 2019

En la tabla 16 se presenta una validación del score, para eso se segmenta la población con la que se prueba el modelo optimizador por deciles, ordenando y dejando en los primeros deciles los individuos que obtuvieron un score más bajo y en los últimos deciles a los scores más altos, evidenciando que los scores más bajos acumulan la mayor cantidad de posibles reincidentes.

TABLA 17: MODELO BALANCEADO (VALIDACIÓN DE SCORE POBLACIÓN ENTRENAMIENTO)

ROC	KS	AR
0,86	0.55	0.72

Decil	Total	Reincidentes	No Reincidentes	PorcentajeTotal	TasaRein	ODDS	NoRein.Acum	ReinAcum	KS	Roc
1	7.773	5.098	2.675	10%	66%	0,53	0,04	0,34	0,29	0,04
2	7.772	3.376	4.396	10%	43%	1,30	0,11	0,56	0,45	0,04
3	7.773	2.535	5.238	10%	33%	2,07	0,20	0,73	0,53	0,03
4	7.772	1.807	5.965	10%	23%	3,30	0,29	0,85	0,55	0,02
5	7.773	1.160	6.613	10%	15%	5,70	0,40	0,92	0,53	0,01
6	7.772	858	6.914	10%	11%	8,06	0,51	0,98	0,47	0,01
7	7.773	308	7.465	10%	4%	24,24	0,63	1,00	0,37	0,00
8	5.458	8	5.450	7%	0%	681,25	0,72	1,00	0,29	0,00
9	2.852	0	2.852	4%	0%	Inf	0,76	1,00	0,24	0,00
10	15.008	0	15.008	19%	0%	Inf	1,00	1,00	0,00	0,00

0 %	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %	100 %
0	433	632	762	847	905	948	983	1000	1000	1000

Fuente: INPEC Nov 2019

En la tabla 17 se presenta una validación del score, para eso se segmenta la población con la que entreno el modelo balanceado por deciles, ordenando y dejando en los primeros deciles los individuos que obtuvieron un score más bajo y en los últimos deciles a los scores más altos, evidenciando que los scores más bajos acumulan la mayor cantidad de posibles reincidentes.

TABLA 18: MODELO BALANCEADO (VALIDACIÓN DE SCORE POBLACIÓN PRUEBA)

ROC	KS	AR
0,861	0.561	0.722

Decil	Total	Reincidentes	No Reincidentes	PorcentajeTotal	TasaRein	ODDS	NoRein.Acum	ReinAcum	KS	Roc
1	3.332	2.255	1.077	10%	68%	0,48	0,04	0,35	0,30	0,03
2	3.330	1.460	1.870	10%	44%	1,28	0,11	0,57	0,46	0,04
3	3.332	1.075	2.257	10%	32%	2,10	0,19	0,73	0,54	0,03
4	3.331	780	2.551	10%	23%	3,27	0,29	0,85	0,56	0,02
5	3.331	506	2.825	10%	15%	5,58	0,40	0,93	0,53	0,01
6	3.330	327	3.003	10%	10%	9,18	0,51	0,98	0,47	0,01
7	3.332	139	3.193	10%	4%	22,97	0,63	1,00	0,37	0,00
8	2.278	3	2.275	7%	0%	758,33	0,71	1,00	0,29	0,00
9	1.255	0	1.255	4%	0%	Inf	0,76	1,00	0,24	0,00
10	6.461	0	6.461	19%	0%	Inf	1,00	1,00	0,00	0,00

0 %	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %	100 %
0	300	510	649	751	832	894	954	1000	1000	1000

Fuente: INPEC Nov 2019

En la Tabla 18 se presenta una validación del score. Para ello, se segmenta la población con la que se prueba el modelo balanceado por deciles, dejando en los primeros deciles a los individuos que obtuvieron un índice más bajo y en los últimos deciles a los índices más altos, evidenciando que los índices más bajos acumulan la mayor cantidad de posibles reincidentes.

7.1.2. Modelo Probit.

Continuando con el análisis, se presenta un enfoque de modelado alternativo utilizando un modelo Probit. Al igual que el modelo Logit, el modelo Probit ayudará a explorar las relaciones entre las variables predictoras y la probabilidad de reincidencia de los individuos en el sistema carcelario en Colombia. La inclusión de un modelo Probit en el estudio permitirá comparar y contrastar los resultados obtenidos con el modelo Logit, enriqueciendo así la comprensión de los factores que influyen en la reincidencia

TABLA 19: MODELO PROBIT.

Modelos	AIC	BIC	Hannan-Quinn
Saturado	9.958,02	10.402,55	10.094,47
Nulo	76.682,19	76.691,45	76.685,03
Optimizado	9.950,64	10.339,6	10.070,03
Balanceado	25.844,63	26.294,11	25.980,41

Fuente: INPEC Nov 2019

En la Tabla 19, muestra los resultados de los criterios de información para el Modelo Probit. Se destaca que el modelo Optimizado presenta AIC, BIC y Hannan-Quinn inferiores en comparación con el Saturado y el Balanceado Optimizado. Esto sugiere que el modelo Optimizado es más eficiente en la explicación de la variabilidad en los datos, indicando su idoneidad como un modelo robusto en la evaluación de la reincidencia carcelaria. Estos resultados respaldan la elección del modelo Optimizado como la opción más adecuada para el análisis predictivo y la identificación de perfiles reincidentes en la población penitenciaria.

El Modelo Probit Optimizado, derivado de un proceso de optimización basado en el análisis de deviance, fue comparado con los modelos Saturado y Nulo. Mediante el uso de modelos Logit en el análisis de deviance, se evaluó la idoneidad del Modelo Probit Optimizado frente al Modelo Nulo (considerando solo la intersección) y el Modelo Saturado (que abarca todas las variables disponibles). La selección del Modelo Probit Optimizado se basó en un análisis hacia adelante y hacia atrás utilizando la función step(). Estos modelos son comparables al ajustarse a la misma población y compartir la misma función de enlace (Probit). Para validar su eficacia en diferentes poblaciones, se realizaron pruebas, incluyendo la aplicación del modelo a una base de datos no utilizada en el entrenamiento. Este enfoque garantiza que el Modelo Probit Optimizado sea efectivo en distintos contextos sin una pérdida significativa de su capacidad predictiva para la reincidencia carcelaria.

Validando Modelo optimizado Tipo Score para el modelo Probit

TABLA 20: MODELO OPTIMIZADO (VALIDACIÓN DE SCORE POBLACIÓN ENTRENAMIENTO)

ROC	KS	GINI
0.788	0.44	0,58

Decil	Total	Reincidentes	No Reincidentes	PorcentajeTotal	TasaRein	ODDS	NoRein.Acum	ReinAcum	KS	Roc
1	7.773	4.390	3.383	10%	57%	0,77	0,05	0,29	0,24	0,05
2	7.772	2.964	4.808	10%	38%	1,62	0,13	0,49	0,36	0,05
3	7.773	2.281	5.492	10%	29%	2,41	0,22	0,64	0,42	0,04
4	7.771	1.786	5.985	10%	23%	3,35	0,31	0,75	0,44	0,03
5	7.774	1.220	6.554	10%	16%	5,37	0,42	0,83	0,42	0,02
6	7.772	946	6.826	10%	12%	7,22	0,53	0,90	0,37	0,02
7	7.773	717	7.056	10%	9%	9,84	0,64	0,94	0,30	0,01
8	7.772	515	7.257	10%	7%	14,09	0,76	0,98	0,22	0,01
9	7.773	295	7.478	10%	4%	25,35	0,88	1,00	0,12	0,00
10	7.773	36	7.737	10%	1%	214,92	1,00	1,00	0,00	0,00

0 %	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %	100 %
92	550	667	747	808	854	891	922	949	982	1000

Fuente: INPEC Nov 2019

En la tabla 20· se presenta una validación del score, para eso se segmenta la población con la que entreno el modelo optimizado por deciles, dejando en los primeros deciles los individuos que obtuvieron un score más bajo y en los últimos deciles a los scores más altos, evidenciando que los scores más bajos acumulan la mayor cantidad de posibles reincidentes.

TABLA 21: MODELO OPTIMIZADO (VALIDACIÓN DE SCORE POBLACIÓN PRUEBA)

ROC	KS	GINI
0.785	0.423	0.57

Decil	Total	Reincidentes	No Reincidentes	PorcentajeTotal	TasaRein	ODDS	NoRein.Acum	ReinAcum	KS	Roc
1	3.332	1.911	1.421	10%	58%	0,74	0,05	0,29	0,24	0,05
2	3.331	1.280	2.051	10%	39%	1,60	0,13	0,49	0,36	0,05
3	3.331	928	2.403	10%	28%	2,59	0,22	0,63	0,41	0,04
4	3.331	725	2.606	10%	22%	3,59	0,32	0,74	0,42	0,03
5	3.331	590	2.741	10%	18%	4,65	0,42	0,83	0,41	0,02
6	3.331	434	2.897	10%	13%	6,68	0,53	0,90	0,37	0,02
7	3.331	309	3.022	10%	9%	9,78	0,64	0,94	0,30	0,01
8	3.331	209	3.122	10%	6%	14,94	0,76	0,98	0,22	0,01
9	3.331	145	3.186	10%	4%	21,97	0,88	1,00	0,12	0,00
10	3.332	14	3.318	10%	0%	237,00	1,00	1,00	0,00	0,00

0 %	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %	100 %
82	549	666	749	809	855	893	922	950	982	1000

Fuente: INPEC Nov 2019

En la tabla 21 se presenta una validación del score, para eso se segmenta la población con la que se prueba el modelo optimizado por deciles, dejando en los primeros deciles los individuos que obtuvieron un score más bajo y en los últimos deciles a los scores más altos, evidenciando que los scores más bajos acumulan la mayor cantidad de posibles reincidentes.

7.1.3. Modelo Elastic Net.

Se procedió a implementar el modelo Elastic Net y se llevaron a cabo múltiples pruebas variando el valor de alpha en el proceso. Sin embargo, en todas las pruebas realizadas, se observó un patrón constante de sobreajuste al conjunto de prueba. Esto se tradujo en la creación de modelos altamente sobreajustados que mostraron dificultades para generalizar los resultados en datos no vistos. Este hallazgo plantea importantes consideraciones sobre la adecuación de Elastic Net en el contexto de nuestro análisis y destaca la necesidad de abordar el sobreajuste de manera efectiva en futuros enfoques y modelos.

7.2. Modelos de Conteo:

En las siguientes secciones, se abordarán diversas técnicas y modelos de conteo diseñados para mejorar la precisión de las predicciones.

7.2.1. Regresión de Poisson:

Tabla 22: Comparación de Modelos de Regresión de Poisson

Modelos	AIC	BIC	Hannan-Quinn	MAE	ECM
Saturado	35.845,36	36.354,71	36.001,71	11,66	222,12
Nulo	115.422,90	115.432,20	115.425,80	1,51	2,8070
Optimizado	35.782,7	35.893,83	35.816,81	11,66	222,13

Fuente: INPEC Nov 2019

En la Tabla 22, se exhiben los variados criterios de información empleados para evaluar las diferentes iteraciones del modelo de regresión de Poisson. Sorprendentemente, el Modelo Nulo sobresale al proporcionar un ajuste apropiado a datos no observados, evidenciado por sus valores ligeramente inferiores en MAE y un ECM similar en la base de prueba. Estos resultados sugieren que el modelo de Poisson puede no ser la opción más idónea para modelar esta base, dada la marcada presencia de ceros en la variable objetivo.

El Modelo Saturado incluye todas las variables predictoras disponibles y muestra un desempeño mixto según los criterios de información. Si bien tiene valores más bajos de AIC, BIC y Hannan-Quinn, lo que sugiere un mejor ajuste en comparación con otros modelos, su capacidad de predicción en la base de prueba, como se evidencia en MAE y ECM, es menos satisfactoria.

El Modelo Nulo, que consiste en un intercepto, tiene valores más altos en los criterios de información, pero sorprendentemente muestra un desempeño superior en la validación cruzada en la base de prueba.

El Modelo Optimizado, seleccionado mediante un proceso de Devianza, tiene un rendimiento competitivo en términos de AIC, BIC y Hannan-Quinn, pero experimenta dificultades en la validación cruzada, con valores de MAE y ECM comparables a los del Modelo Saturado.

Esta discrepancia en el rendimiento entre los modelos se puede atribuir en parte a la gran cantidad de ceros presentes en la base de datos. La sobredispersión resultante en los modelos puede influir en su capacidad para predecir con precisión, lo que se refleja en la validación cruzada.

En vista de estos hallazgos, la elección del modelo óptimo para fines predictivos podría depender de la priorización de la precisión de las predicciones sobre la complejidad del modelo. A pesar de su simplicidad, el Modelo Nulo muestra un rendimiento superior en la validación cruzada y podría ser preferible para aplicaciones donde la precisión es fundamental.

7.2.2. Regresión de Poisson:

Tabla 23: Comparación de Modelos de Regresión de Poisson

Modelos	AIC	BIC	Hannan-Quinn	MAE	ECM
Saturado	35.845,36	36.354,71	36.001,71	11,66	222,12
Nulo	115.422,90	115.432,20	115.425,80	1,51	2,8070
Optimizado	35.782,7	35.893,83	35.816,81	11,66	222,13

Fuente: INPEC Nov 2019

Los resultados de los modelos de Regresión Binomial Negativa se resumen en la Tabla 23, que incluye métricas como el AIC, BIC y Hannan-Quinn, junto con el Error Absoluto Medio (MAE) y el Error Cuadrático Medio (ECM), calculados mediante una validación cruzada en la base de prueba donde los modelos muestran sobreajustes debido a su MAE y ECM tan elevados.

El Modelo Saturado, que incorpora todas las variables disponibles, presenta valores más bajos en AIC, BIC y Hannan-Quinn en comparación con otros modelos, indicando un mejor ajuste aparente. Sin embargo, su desempeño en la validación cruzada, como se refleja en MAE y ECM, no es tan sólido.

El Modelo Nulo, que consiste solo en un intercepto, muestra valores más altos en los criterios de información, pero sorprendentemente supera a los otros modelos en la validación cruzada, con valores de MAE y ECM más bajos.

El Modelo Optimizado, seleccionado a través de un proceso de optimización, presenta valores competitivos en términos de AIC, BIC y Hannan-Quinn. Sin embargo, como en el caso de la Regresión de Poisson, su desempeño en la validación cruzada es comparable al del Modelo Saturado.

Al igual que en el análisis de la Regresión de Poisson, esta discrepancia en el rendimiento de los modelos de Regresión Binomial Negativa puede deberse a la presencia de ceros inflados en la base de datos, lo que conlleva a la sobredispersión y puede impactar la capacidad de predicción en la validación cruzada.

En consecuencia, al elegir el modelo óptimo para fines de predicción, puede ser esencial ponderar la precisión de las predicciones frente a la complejidad del modelo. A pesar de su simplicidad, el Modelo Nulo ofrece un rendimiento superior en la validación cruzada y podría ser preferible en aplicaciones donde la precisión es crucial.

En resumen, la evaluación de los modelos de Regresión Binomial Negativa pone de manifiesto la importancia de considerar la estructura particular de los datos al seleccionar el modelo óptimo para predecir la reincidencia en la población carcelaria.

7.3. Modelos de recuento inflados por ceros:

En este apartado, se explorarán modelos de recuento inflados por ceros que desempeñan un papel crucial en la predicción de reincidencias delictivas. Estos modelos permiten abordar de manera efectiva la presencia de ceros en los datos y proporcionan una herramienta valiosa para comprender mejor la reincidencia en el sistema penal.

7.3.1. Poisson Inflado con Ceros (PIC):

Dentro de la familia de modelos de recuento inflados por ceros, el modelo Poisson Inflado con Ceros (PIC) se destaca por su capacidad para lidiar con situaciones en las que los ceros son más frecuentes de lo esperado.

TABLA 24: MODELO DE POISSON INFLADO CON CEROS (PIC)

Modelos	AIC	BIC	Hannan-Quinn	MAE	ECM
Saturado	80.052.91	80.682.66	80.246.22	0.331	0.365
Nulo	106.407.2	106.425.7	106.412.9	0.478	0.507
Optimizado	40.163.76	40.645.33	40.311.58	0.12	0.13

Fuente: INPEC Nov 2019

En la tabla 24 sugieren que el Modelo Optimizado se destaca en términos de AIC, BIC y Hannan-Quinn, lo que indica un mejor ajuste aparente y un menor grado de sobreajuste en comparación con otros modelos. Además, en la validación cruzada, el Modelo Optimizado demuestra un rendimiento sólido, con valores más bajos de MAE y ECM, lo que implica una capacidad de predicción más precisa.

En esta tabla, se presentan los resultados de diferentes modelos de Poisson Inflado con Ceros (PIC), junto con sus criterios de información, errores medios absolutos (MAE) y errores cuadráticos medios (ECM). Los valores más bajos en AIC, BIC y Hannan-Quinn sugieren que el Modelo Optimizado es el que presenta un mejor ajuste aparente y una menor probabilidad de sobreajuste. Además, en la validación cruzada, el Modelo Optimizado demuestra un rendimiento sólido con MAE y ECM más bajos, lo que indica una capacidad de predicción más precisa.

En conclusión, el Modelo Optimizado de Poisson Inflado con Ceros (PIC) se destaca como la elección más adecuada para predecir la reincidencia delictiva en esta población carcelaria. Su capacidad para abordar los ceros inflados y proporcionar predicciones precisas lo convierte en una herramienta valiosa para comprender y gestionar la reincidencia en el sistema penal.

Estos hallazgos respaldan la elección del Modelo Optimizado para cumplir con los objetivos de predicción de este estudio.

7.3.2. Binomial negativa inflado por ceros:

TABLA 25: MODELO DE BINOMIAL NEGATIVA INFLADO POR CEROS

Modelos	AIC	BIC	Hannan-Quinn	MAE	ECM
Saturado	80.054,91	80.693,92	80.251,06	0,331	0,365
Nulo	106.155,6	106183.3	106164.1	0,478	0,507
Optimizado	40.165,76	40.656,59	40.316,42	0,12	0,133

Fuente: INPEC Nov 2019

La tabla 25 muestra las métricas clave de evaluación para el Modelo de Binomial Negativa Inflada por Ceros. Este modelo es analizado en comparación con el "Modelo Saturado",^{el} "Modelo Nulo" el "Modelo Optimizado".

Modelo Saturado: se caracteriza por tener valores más altos de AIC, BIC y Hannan-Quinn en comparación con los otros modelos. Estos indicadores señalan un ajuste deficiente a los datos, a pesar de tener un MAE de 0.331 y un ECM de 0.365 en la base de prueba. La diferencia entre los valores de AIC, BIC y Hannan-Quinn y el desempeño del modelo en la base de prueba sugieren que el "Modelo Saturado" no es el ajuste más adecuado, ya que no logra equilibrar el sobreajuste.

Modelo Nulo: presenta valores más bajos en AIC, BIC y Hannan-Quinn en comparación con el "Modelo Saturado". Sin embargo, en la base de prueba, muestra un MAE significativamente más alto de 0.478 y un ECM de 0.507. A pesar de los indicadores de calidad de ajuste inferiores, los valores de AIC, BIC y Hannan-Quinn sugieren que el "Modelo Nulo" no es el mejor ajuste para los datos.

Modelo Optimizado: se destaca entre los tres modelos al exhibir valores más bajos de AIC, BIC y Hannan-Quinn. A pesar de que el MAE y el ECM en la base de prueba son idénticos al "Modelo Nulo", la relación entre los indicadores de calidad de ajuste y el rendimiento en la población de prueba sugiere que el "Modelo Optimizado"^{es} una opción sólida para estimar el conteo de reincidencias delictivas.

En resumen, aunque los valores de AIC, BIC y Hannan-Quinn de los modelos a menudo se utilizan para indicar un buen ajuste, es vital recordar que los modelos tienden a mostrar un fuerte sobreajuste, lo que afecta la capacidad de generalización en datos no vistos. Por lo tanto, la elección del modelo debe equilibrar el rendimiento en la base de entrenamiento y la precisión en la base de prueba, lo que requerirá una consideración cuidadosa en futuros análisis y pronósticos relacionados con la reincidencia delictiva.

7.4. Modelos de recuento por cero truncado:

En el amplio espectro de modelos de conteo utilizados en análisis estadísticos, la sección 8.3 se enfoca en dos enfoques específicos para el manejo de conteos excluyentes de cero. Estos modelos son esenciales cuando los valores de cero son inherentemente excluidos o poco probables en la contabilización de eventos.

El primer modelo, conocido como el modelo de Poisson Cero Truncado (Poisson Zero-Truncated), aborda situaciones donde los ceros no son representativos en los datos y no deberían ser considerados en el modelo. En la sección 8.3.1, se examina en profundidad el modelo de Poisson Cero Truncado, explorando sus aplicaciones y características específicas.

En el segundo modelo, denominado el modelo de Binomial Negativa Cero Truncado (Zero-Truncated Negative Binomial), se encuentra una solución para situaciones similares donde los valores de cero son excluidos debido a restricciones en la contabilización. La sección 8.3.2 se dedica a analizar este modelo, proporcionando información detallada sobre su utilidad y propiedades distintivas.

7.4.1. Poisson cero truncado:

Tabla 26: Modelo de Poisson cero truncado

Modelos	AIC	BIC	Hannan-Quinn	MAE	ECM
Saturado	28.555,96	28.969,48	28.784,61	0,761	1,08
Nulo	29.724,99	29.732,62	28.253,97	0,4018	0,669
Optimizado	28.352,09	28.542,73	28.384,61	0,761	1,08

Fuente: INPEC Nov 2019

En la tabla 26, se observan diferencias en los criterios de información, lo que podría dar la impresión de que el "Modelo de Poisson Cero Truncado" es la mejor opción. No obstante, al analizar su desempeño en la población de prueba, se detecta un patrón de sobreajuste, donde predice correctamente solo en la base de entrenamiento.

Modelo Saturado: El "Modelo de Poisson Cero Truncado" saturado presenta valores de AIC, BIC y Hannan-Quinn más bajos en comparación con los valores del "Modelo Nulo". Sin embargo, muestra un MAE de 0,761 y un ECM de 1,08 en la base de prueba, lo que sugiere que podría estar sobreajustando los datos.

Modelo Nulo: El "Modelo Nulo" tiene valores significativamente más altos en AIC, BIC y Hannan-Quinn en comparación con el "Modelo Saturado". Además, muestra un MAE de 0,4018 y un ECM de 0,669 en la base de prueba. Esto indica que el "Modelo Nulo" podría no ser un buen ajuste para los datos.

Modelo Optimizado: El "Modelo Optimizado" muestra valores de AIC, BIC y Hannan-Quinn similares al "Modelo Saturado". Tiene un MAE y un ECM idénticos al "Modelo Saturado", lo que sugiere que podría no ser una mejora significativa sobre el "Modelo Saturado".

7.4.2. Binomial negativa cero truncado:

Tabla 27: Modelo binomial negativa cero truncado

Modelos	AIC	BIC	Hannan-Quinn	MAE	ECM
Saturado	28.212,77	28.555,93	28.263,16	1,388	2,430
Nulo	29.473,36	29.488,61	29.473,89	0,728	0.857
Optimizado	28.220,06	28.456,46	28.198,82	1,369	2,377

Fuente: INPEC Nov 2019

En la tabla 27, se observan diferencias en los criterios de información, lo que podría llevar a pensar que el "Modelo de Binomial Negativa Cero Truncado" es una opción a considerar. Sin embargo, es fundamental destacar que al analizar su desempeño en la población de prueba, se detecta un patrón de sobreajuste, donde predice correctamente solo en la base de entrenamiento.

Modelo Saturado: Los valores específicos del "Modelo Saturado" no fueron proporcionados, lo que limita nuestra capacidad para realizar conclusiones sobre este modelo en particular.

Modelo Nulo: El "Modelo Nulo" muestra valores más altos de AIC, BIC y Hannan-Quinn en comparación con el "Modelo de Binomial Negativa Cero Truncado". Además, tiene un MAE de 0,728 y un ECM de 0,857 en la base de prueba. Estos indicadores sugieren que el "Modelo Nulo" podría estar sobreajustando los datos en cierta medida.

Modelo Optimizado: Los valores específicos del "Modelo Optimizado" no fueron proporcionados, lo que limita nuestras conclusiones sobre este modelo.

7.5. Modelos Hurdle

7.5.1. Modelo Hurdle Poisson (Hurdle Poisson Model)

Tabla 28: Modelo Hurdle Poisson (Hurdle Poisson Model)

Modelos	AIC	BIC	Hannan-Quinn	MAE	ECM
Saturado	99.763,88	100.301	99.928,76	0,437	0,463
Nulo	106.407,2	106.425,7	106.412,9	0,478	0,507
Optimizado	28.555,29	29.240,6	28.765,65	0.121	0.131

Fuente: INPEC Nov 2019

En la tabla 28, se pueden observar notables disparidades entre los modelos analizados en cuanto a sus métricas de rendimiento. Destacando entre ellos, el Modelo Optimizado emerge como la elección preferida, al exhibir indicadores de precisión más sobresalientes en contraste con el Modelo Saturado.

.^{El} "Modelo Saturado" presenta un MAE de 0,437 y un ECM de 0,463 en la base de prueba, lo que indica que se ajusta de manera razonable a nuevos datos, aunque no necesariamente de la manera más óptima. Por otro lado, el "Modelo Nulo" exhibe valores más altos en los criterios de AIC, BIC y Hannan-Quinn, lo que sugiere que no se ajusta adecuadamente a los datos. Además, en la base de prueba, se obtiene un MAE de 0,478 y un ECM de 0,507, confirmando su rendimiento subóptimo. Es destacable que el "Modelo Optimizado" destaca al mostrar valores sustancialmente mejores en los criterios de AIC, BIC y Hannan-Quinn en comparación con los otros dos modelos. Además, en la base de prueba, se distingue al presentar un MAE de solamente 0,121 y un ECM de 0,131. Estos valores significativamente inferiores en comparación con los otros dos modelos sugieren una alternativa más sencilla y eficaz, lo que podría resultar en un rendimiento efectivo.

7.5.2. Modelo Binomial Negativo (Negative Binomial Model)

Tabla 29: Modelo Binomial Negativo (Negative Binomial Model)

Modelos	AIC	BIC	Hannan-Quinn	MAE	ECM
Saturado	97754.62	98356.58	97939.4	0.425	0.453
Nulo	106.155,6	106.183,3	106.164,1	0,478	0,507
Optimizado	28.503,67	29.198,25	28.716,88	0,121	0,131

Fuente: INPEC Nov 2019

En la tabla 29, se pueden observar notables disparidades entre los modelos analizados en cuanto a sus métricas de rendimiento. Destacando entre ellos, el Modelo Optimizado emerge como la elección preferida, al exhibir indicadores de precisión más sobresalientes en contraste con el Modelo Saturado.

El "Modelo Saturado" muestra un MAE de 0,437 y un ECM de 0,463 en la base de prueba, lo que sugiere un ajuste razonable a nuevos datos, aunque no necesariamente óptimo.

Por otro lado, el "Modelo Nulo" presenta valores más altos en AIC, BIC y Hannan-Quinn, lo que indica un ajuste menos adecuado a los datos. Además, muestra un MAE de 0,478 y un ECM de 0,507 en la base de prueba, lo que confirma su desempeño subóptimo.

Es notable que el "Modelo Optimizado" sobresale al mostrar valores sustancialmente mejores en los criterios de AIC, BIC y Hannan-Quinn en comparación con los otros dos modelos. Además, en la base de

prueba, se destaca al exhibir un MAE de tan solo 0,121 y un ECM de 0,131. Estos valores significativamente inferiores en comparación con los otros dos modelos sugieren una alternativa más simple y eficaz, lo que podría traducirse en un rendimiento igualmente efectivo.

8. Conclusiones

La exploración de los modelos probabilísticos condujo al descubrimiento de resultados notables, generando así una evaluación detallada y una comprensión más profunda de su efectividad y desafíos de cada una de las técnicas aplicadas. Las tablas 14 y 19 evidenciaron que, en apariencia, los modelos saturado Logit y Probit presentaban indicadores superiores en términos de criterios de información. Sin embargo, un análisis más detallado desveló un desafío clave: el sobreajuste. A pesar de su prometedor desempeño en la base de entrenamiento, estos modelos mostraron una limitada capacidad para predecir con precisión los resultados en poblaciones de prueba. Tendían a asignar probabilidades extremas a los reincidentes y no reincidentes, careciendo de puntos medios en la población.

Para abordar esta problemática, se empleó un proceso de selección de variables. Un modelo optimizado fue desarrollado a través de un proceso de deviance (mediante el código step), lo que permitió evaluar diferentes combinaciones de variables explicativas. Sin embargo, incluso el modelo optimizado reveló desafíos al predecir la reincidencia en poblaciones de prueba.

Finalmente, se llegó a una conclusión clave: el mejor modelo fue el entrenado con las variables seleccionadas del modelo optimizado. Este enfoque, respaldado por la técnica de sobremuestreo de minorías sintéticas, resultó ser la estrategia más efectiva para modelar una base de datos donde la mayoría de los casos no reincide, lo que genera una distribución cero inflada. La combinación de la selección de variables y el sobremuestreo permitió mejorar significativamente la capacidad predictiva del modelo.

Estos hallazgos son cruciales para informar la toma de decisiones en el estudio de la reincidencia carcelaria en Colombia. Ofrecen una comprensión más profunda de los retos y las oportunidades en este campo, y destacan la importancia de técnicas avanzadas en la modelación de datos con distribuciones desafiantes.

Una vez finalizada la construcción de los modelos probabilísticos y calculadas las probabilidades de reincidencia para cada individuo, se procedió a desarrollar un Score de Reincidencia. En este proceso, se evaluaron diferentes modelos, y se identificó que el mejor desempeño se logró con el modelo entrenado utilizando la muestra balanceada y las variables del modelo optimizado a través del proceso de deviance. Este modelo sobresalió en términos de métricas de evaluación como KS (Kolmogorov-Smirnov), AR (AUC-ROC), y la curva ROC en comparación con los demás modelos, tanto en la validación interna con la base de entrenamiento como en la validación externa con la base de prueba.

El Score de Reincidencia se organizó en deciles, de manera que los individuos con un puntaje más bajo tienen una mayor probabilidad de reincidir, mientras que aquellos con un puntaje más alto tienen una probabilidad menor. Esta segmentación permitirá a las autoridades penitenciarias identificar y priorizar a los individuos con un mayor riesgo de reincidencia y, por lo tanto, asignar recursos y programas de reintegración de manera más eficiente.

La implementación de este Score de Reincidencia contribuirá en gran medida a la labor de las autoridades competentes y a la sociedad en su conjunto, al ofrecer una herramienta precisa y basada en evidencia para la toma de decisiones en el sistema de justicia penal y en la gestión de la reincidencia carcelaria en Colombia. Estos resultados respaldan la efectividad de abordar el desbalance de clases y optimizar los modelos, lo que representa un avance significativo en la comprensión y prevención de la reincidencia criminal en el país.

Una vez seleccionado el modelo que mejor se adapta a los datos, es fundamental resaltar las variables que ejercen una influencia significativa en la probabilidad de reincidencia. A pesar de que la optimización predictiva ha sido el enfoque principal de este trabajo, el modelo logístico ha identificado factores cruciales vinculados al ingreso y la ocupación. Específicamente, se ha observado que el estado de condena

domiciliaria y la participación en actividades educativas y laborales se correlacionan con un aumento en las probabilidades de reincidencia, como se observa en la sección 6.3.1 Análisis descriptivo, del presente trabajo. Este hallazgo respalda la noción previamente discutida de que las cárceles pueden convertirse en entornos que fomentan la violencia, vulneran los derechos humanos y facilitan la formación de redes criminales (CELIV, 2015). Aunque la identificación precisa de estas variables no era el objetivo primario de la investigación, estas observaciones proporcionan perspectivas valiosas sobre la compleja dinámica que rodea la reincidencia carcelaria.

En este contexto, el análisis específico de programas de resocialización reveló que aquellos que enseñan o educan a otros reclusos presentan tasas significativamente más bajas de reincidencia, señalando la necesidad de realizar un análisis más detallado de la forma en que se implementan estos programas. Asimismo, el nivel educativo emergió como un factor destacado, evidenciando que la educación superior se asocia con menores probabilidades de reincidencia, mientras que la educación básica primaria se asocia con mayores probabilidades. Estos resultados ofrecen una perspectiva valiosa sobre la importancia de la educación en el proceso de resocialización y subrayan la necesidad de profundizar en cómo se diseñan y se llevan a cabo los programas educativos en el ámbito carcelario. Esto podría implicar una revisión crítica de las estrategias y enfoques educativos para maximizar su impacto positivo en la reducción de la reincidencia carcelaria.

en los modelos de regresión Poisson y Binomial Negativa se presenta un sobre ajuste debido a la naturaleza de los datos evidenciado por el análisis de validación cruzada (MAE y ECM) el cual son valores muy altos y en los dos casos siendo el modelo nulo el modelo con menor MAE y ECM.

Este estudio buscó evaluar el impacto de los programas de resocialización en la reincidencia carcelaria. Los resultados muestran tasas de reincidencia del 34.8 % para quienes trabajan en la cárcel, 37.6 % para los que estudian, y solo 0.6 % para quienes enseñan o educan a otros reclusos. En contraste, aquellos que no participan en programas tienen una tasa de reincidencia del 27 %, como se puede observar en la tabla 7

Estos hallazgos sugieren que los programas de resocialización podrían no estar cumpliendo plenamente su función. Específicamente, la baja tasa de reincidencia entre quienes enseñan o educan a otros reclusos indica que se deben revisar y mejorar estos programas para abordar eficazmente la reincidencia delictiva en Colombia.

A lo largo de este proceso, se evaluaron varios enfoques, y los Modelos de Regresión Cero-Inflada se destacaron como la elección más efectiva. Esta elección se basó en su capacidad para manejar la complejidad de los datos de reincidencia, lo que es esencial para lograr predicciones precisas.

Dentro de la categoría de Modelos de Regresión Cero-Inflada, el Modelo de Hurdle de Distribución Binomial Negativa se destacó como el más adecuado en todas las pruebas realizadas. Esta elección se basó en su habilidad para capturar con precisión las sutilezas de la reincidencia, lo que conlleva a predicciones más fiables.

La selección de las variables predictoras se llevó a cabo cuidadosamente, teniendo en cuenta el criterio de deviance, lo que fortaleció aún más la elección del Modelo de Hurdle de Distribución Binomial Negativa. La aplicación de este modelo en el estudio proporciona una herramienta efectiva para la evaluación de factores que inciden en la reincidencia carcelaria, y promete contribuir significativamente a la mejora de los programas de resocialización en Colombia.

9. Bibliografía

1. Alcaide Delgado, Mario. (2015). Modelo de regresión binomial negativa. [Tesis de Maestría, Sevilla, España]. Archivo digital.
<https://idus.us.es/bitstream/handle/11441/40815/Alcaide20Delgado20Mario20TFG.pdf?sequence=1>

2. Bales D, William. (2015). Assessing the impact of imprisonment on recidivism.
https://www.researchgate.net/publication/257677075_Assessing_the_impact_of_imprisonment_on_recidivism
3. Caicedo, J., (2014). Reincidencia Carcelaria en Colombia: Un análisis de duración. [Tesis de Maestría, Universidad Javeriana, Bogotá, Colombia].
4. Dammert Lucia. (2015). Sistemas carcelarios en américa latina: Avances y desafíos. ICEPALA.
https://www.researchgate.net/publication/320216123_Sistemas_carcelarios_en_America_Latina_Avances_y_Desafios
5. Distribución Poisson, L., Gómez, M., & Benlloch, M. (s/f). La distribución Poisson. Upv.es. Recuperado el 26 de junio de 2023, de
<https://riunet.upv.es/bitstream/handle/10251/7937/Distribucion%20Poisson.pdf>
6. Elastic Net.
https://diposit.ub.edu/dspace/bitstream/2445/179200/1/TFM-CAF_LckenGimnez_2021.pdf
7. EOM (2019). Un mundo de prisioneros. Mayores poblaciones carcelarias. El Orden Mundial
<https://elordenmundial.com/mapas/paises-mayor-numero-de-presos-en-la-carcel/>
8. Figueroa Arboccó Julianna T. (2005). La fecundidad y su relación con variables socioeconómicas, demográficas y educativas aplicando el Modelo de Regresión Poisson. [Tesis de Grado, Universidad Nacional Mayor de San Marcos, Lima, Perú]. Archivo digital.
https://sisbib.unmsm.edu.pe/bibvirtualdata/monografias/basic/figueroa_ag/cap3.pdf
9. Fernando, J., & Bernal, B. (s/f). Cepal.org. Recuperado el 26 de junio de 2023, de https://oig.cepal.org/sites/default/files/2000_codigopenal_colombia.pdf
10. Garzón Juan C., Llorente María V. & Suárez Manuela. (2018). ¿Qué hacer con la reincidencia delincencial? El problema y sus posibles soluciones. Fundación Ideas para la Paz (FIP).
<http://www.ideaspaz.org/publications/posts/1658>
11. Gómez, M., Benlloch, M., & Punto De Partida. (s/f). La distribución binomial. Upv.es. Recuperado el 26 de junio de 2023, de <https://riunet.upv.es/bitstream/handle/10251/7936/Distribucion%20binomial.pdf?se>
12. Gómez-Déniz, E., & Sarabia, y. J. M. (s/f). LA DISTRIBUCIÓN BINOMIAL-EXPONENCIAL TRUNCADA CON APLICACIONES EN EL SECTOR DEL SEGURO DE AUTOMÓVILES. Core.ac.uk. Recuperado el 26 de junio de 2023, de <https://core.ac.uk/download/pdf/147469385.pdf>
13. Hilbe, M, J. (2014). Modelado de datos de conteo. Cambridge University Press, Nueva York.
<https://www.cambridge.org/core/books/modeling-countdata/BFEB3985905CA70523D9F98DA8E64D08>.
14. Jiménez Hernández Norberto. (2017). La resocialización como fin de la pena – Una frustración en el sistema penitenciario y carcelario colombiano. Revista SciELO - Scientific Electronic Library Online. 1(1). 539-559.
<http://www.scielo.br/pdf/ccrh/v30n81/0103-4979-ccrh-30-81-0539.pdf>
15. K Applegate Brandon., Surette Ray. & J McCarthy Bernard. (1979). Detention and desistance from crime: Evaluating the influence of a new generation jail on recidivism.
<https://www.sciencedirect.com/science/article/pii/S0047235299000240>
16. Logit y Probit.
https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib0515/Libro.pdf

17. La distribución binomial negativa.
https://www.ugr.es/~proman/ProbI/2016_2017/PDF/DistribucionesDiscretas.pdf
18. Pontificia Universidad Javeriana. (2015). Reincidencia en el delito. CELIV.
<http://www.celiv.untref.edu.ar/delito-en-cifras-abril2014.html>.
19. Ramírez González Andrés M. (2017). Modelo de regresión de poisson inflacionado en ceros, en el modelamiento de la deserción estudiantil en la universidad del Quindío. [Tesis de Maestría, Universidad tecnológica de Pereira, Colombia].
20. Riaño Vargas Juan M. (2016). Reincidencia penitenciaria: Factores que influyen en la reincidencia de la población desmovilizada de las AUC. [Tesis doctoral, Universidad de los Andes, Bogotá, Colombia].
21. Salinas-Rodríguez, A., Manrique-Espinoza, B. & Sosa-Rubí, S. G. (2009). Análisis estadístico para datos de conteo: aplicaciones para el uso de los servicios de salud. Revista SciELO - Scientific Electronic Library Online, 51(5), 397–406.
22. Velasco Vázquez María L. (2008). Un Modelo de Regresión Poisson Inflado con Ceros para Analizar datos de un Experimento de Fungicidas en Jitomate. [Tesis de Maestría, Universidad Veracruzana, Xalapa, México].
23. Mullahy, J. (1986). Specification and testing of some modified count data models. Journal of Econometrics, 33(3), 341-365.
24. Greene, W. H. (2003). *Econometric Analysis*. Prentice Hall.
25. Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press.
26. Greene, W. H. (2003). *Econometric Analysis*. Prentice Hall.
27. Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

10. Anexo

10.1. Anexo A: Detalles Adicionales de Modelos

Modelo Logit

A continuación se presentan los detalles adicionales del Modelo Logit utilizado en el estudio.

Tabla 30: Coeficientes del Modelo Logit Saturado

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-202.38	618.62	-0.33	0.74
EDAD	0.02	0.00	18.75	0.00
‘Año ingreso’	0.08	0.01	13.10	0.00
DomicilioCat2	-0.58	0.03	-17.11	0.00
DomicilioCat3	-0.99	0.03	-32.08	0.00
GENEROMasculino	0.41	0.05	8.52	0.00
‘ESTADO INGRESO’DOMICILIARIA	39.50	639.22	0.06	0.95
‘ESTADO INGRESO’ESPERA TRASLADO	-0.21	878.97	-0.00	1.00
‘ESTADO INGRESO’INTRAMUROS	17.83	605.92	0.03	0.98
‘ESTADO INGRESO’PRISION DOMICILIARIA	-0.26	610.16	-0.00	1.00
‘ESTADO INGRESO’VIGILANCIA ELECTRONICA	17.68	605.92	0.03	0.98
Categori_Años_condena11 a 15	-0.41	0.04	-9.37	0.00
Categori_Años_condena16 a 20	-0.54	0.05	-10.68	0.00
Categori_Años_condena21 a 25	-0.43	0.07	-6.30	0.00
Categori_Años_condena26 a 30	-0.77	0.09	-8.65	0.00
Categori_Años_condena31 a 36	-0.94	0.10	-9.63	0.00
Categori_Años_condena6 a 10	-0.21	0.03	-7.08	0.00
Categori_Años_condenamayor a 36	-1.02	0.09	-11.66	0.00
‘NIVEL EDUCATIVO’BASICA PRIMARIA	0.16	0.03	6.03	0.00
‘NIVEL EDUCATIVO’EDUCACION SUPERIOR	-1.11	0.09	-12.98	0.00
‘NIVEL EDUCATIVO’ILETRADOS	-0.03	0.06	-0.46	0.64
‘ACTIVIDAD OCUPACIONAL’ESTUDIO	0.46	0.12	3.72	0.00
‘ACTIVIDAD OCUPACIONAL’sin Dato	-1.04	0.13	-8.30	0.00
‘ACTIVIDAD OCUPACIONAL’TRABAJO	0.42	0.12	3.42	0.00
RegionalNOROESTE	-0.24	0.04	-6.29	0.00
RegionalNORTE	0.38	0.05	7.79	0.00
RegionalOCCIDENTE	0.21	0.04	6.06	0.00
RegionalORIENTE	0.12	0.04	2.74	0.01
RegionalVIEJO CALDAS	-0.02	0.04	-0.54	0.59
‘Cantidad delitos’	0.20	0.02	10.32	0.00
Titulo11	-0.20	0.04	-5.15	0.00
Titulo41	-1.41	0.06	-25.02	0.00
Titulo71	0.73	0.03	25.62	0.00
Titulo91	-0.32	0.10	-3.23	0.00
Titulo101	-0.75	0.10	-7.47	0.00
Titulo121	-0.27	0.04	-7.53	0.00
Titulo161	0.75	0.09	8.29	0.00
ESTABLECIMIENTOCat2	17.54	124.08	0.14	0.89
ESTABLECIMIENTOCat3	18.70	124.08	0.15	0.88
Titulo2y31	-0.50	0.05	-9.46	0.00
Titulo5_6y81	-0.25	0.08	-3.32	0.00
Titulo11_13y141	-0.37	0.10	-3.79	0.00
Titulo15_17y181	-0.37	0.09	-4.37	0.00

Tabla 31: Coeficientes del Modelo Logit Nulo

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.42	0.01	-156.65	0.00

Modelo Probit: A continuación se presentan los detalles adicionales del Modelo

Tabla 32: Coeficientes del Modelo Logit Optimizado

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-202.38	618.62	-0.33	0.74
‘ESTADO INGRESO’DOMICILIARIA	39.50	639.22	0.06	0.95
‘ESTADO INGRESO’ESPERA TRASLADO	-0.21	878.97	-0.00	1.00
‘ESTADO INGRESO’INTRAMUROS	17.83	605.92	0.03	0.98
‘ESTADO INGRESO’PRISION DOMICILIARIA	-0.26	610.16	-0.00	1.00
‘ESTADO INGRESO’VIGILANCIA ELECTRONICA	17.68	605.92	0.03	0.98
ESTABLECIMIENTOCat2	17.54	124.08	0.14	0.89
ESTABLECIMIENTOCat3	18.70	124.08	0.15	0.88
‘ACTIVIDAD OCUPACIONAL’ESTUDIO	0.46	0.12	3.72	0.00
‘ACTIVIDAD OCUPACIONAL’sin Dato	-1.04	0.13	-8.30	0.00
‘ACTIVIDAD OCUPACIONAL’TRABAJO	0.42	0.12	3.42	0.00
Titulo71	0.73	0.03	25.62	0.00
DomicilioCat2	-0.58	0.03	-17.11	0.00
DomicilioCat3	-0.99	0.03	-32.08	0.00
Titulo41	-1.41	0.06	-25.02	0.00
Categori_Años_condena11 a 15	-0.41	0.04	-9.37	0.00
Categori_Años_condena16 a 20	-0.54	0.05	-10.68	0.00
Categori_Años_condena21 a 25	-0.43	0.07	-6.30	0.00
Categori_Años_condena26 a 30	-0.77	0.09	-8.65	0.00
Categori_Años_condena31 a 36	-0.94	0.10	-9.63	0.00
Categori_Años_condena6 a 10	-0.21	0.03	-7.08	0.00
Categori_Años_condenamayor a 36	-1.02	0.09	-11.66	0.00
EDAD	0.02	0.00	18.75	0.00
‘NIVEL EDUCATIVO’BASICA PRIMARIA	0.16	0.03	6.03	0.00
‘NIVEL EDUCATIVO’EDUCACION SUPERIOR	-1.11	0.09	-12.98	0.00
‘NIVEL EDUCATIVO’ILETRADOS	-0.03	0.06	-0.46	0.64
RegionalNOROESTE	-0.24	0.04	-6.29	0.00
RegionalNORTE	0.38	0.05	7.79	0.00
RegionalOCCIDENTE	0.21	0.04	6.06	0.00
RegionalORIENTE	0.12	0.04	2.74	0.01
RegionalVIEJO CALDAS	-0.02	0.04	-0.54	0.59
‘Año ingreso’	0.08	0.01	13.10	0.00
GENEROMasculino	0.41	0.05	8.52	0.00
Titulo161	0.75	0.09	8.29	0.00
Titulo101	-0.75	0.10	-7.47	0.00
Titulo2y31	-0.50	0.05	-9.46	0.00
‘Cantidad delitos’	0.20	0.02	10.32	0.00
Titulo121	-0.27	0.04	-7.53	0.00
Titulo11	-0.20	0.04	-5.15	0.00
Titulo15_17y181	-0.37	0.09	-4.37	0.00
Titulo11_13y141	-0.37	0.10	-3.79	0.00
Titulo5_6y81	-0.25	0.08	-3.32	0.00
Titulo91	-0.32	0.10	-3.23	0.00

Tabla 33: Coeficientes del Modelo Probit Saturado

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-111.29	134.10	-0.83	0.41
EDAD	0.01	0.00	18.18	0.00
‘Año ingreso’	0.05	0.00	14.07	0.00
DomicilioCat2	-0.34	0.02	-17.24	0.00
DomicilioCat3	-0.57	0.02	-32.11	0.00
GENEROMasculino	0.23	0.03	8.87	0.00
‘ESTADO INGRESO’DOMICILIARIA	13.19	138.82	0.10	0.92
‘ESTADO INGRESO’ESPERA TRASLADO	-0.03	191.69	-0.00	1.00
‘ESTADO INGRESO’INTRAMUROS	5.59	131.12	0.04	0.97
‘ESTADO INGRESO’PRISION DOMICILIARIA	-0.16	132.04	-0.00	1.00
‘ESTADO INGRESO’VIGILANCIA ELECTRONICA	5.56	131.12	0.04	0.97
Categori_Años_condena11 a 15	-0.24	0.03	-9.53	0.00
Categori_Años_condena16 a 20	-0.32	0.03	-11.20	0.00
Categori_Años_condena21 a 25	-0.26	0.04	-6.54	0.00
Categori_Años_condena26 a 30	-0.44	0.05	-8.80	0.00
Categori_Años_condena31 a 36	-0.53	0.05	-9.76	0.00
Categori_Años_condena6 a 10	-0.12	0.02	-7.10	0.00
Categori_Años_condenamayor a 36	-0.57	0.05	-11.71	0.00
‘NIVEL EDUCATIVO’BASICA PRIMARIA	0.09	0.01	5.92	0.00
‘NIVEL EDUCATIVO’EDUCACION SUPERIOR	-0.60	0.05	-13.09	0.00
‘NIVEL EDUCATIVO’ILETRADOS	-0.02	0.03	-0.45	0.65
‘ACTIVIDAD OCUPACIONAL’ESTUDIO	0.25	0.07	3.69	0.00
‘ACTIVIDAD OCUPACIONAL’sin Dato	-0.61	0.07	-8.74	0.00
‘ACTIVIDAD OCUPACIONAL’TRABAJO	0.23	0.07	3.43	0.00
RegionalNOROESTE	-0.14	0.02	-6.33	0.00
RegionalNORTE	0.21	0.03	7.76	0.00
RegionalOCCIDENTE	0.11	0.02	5.33	0.00
RegionalORIENTE	0.06	0.03	2.30	0.02
RegionalVIEJO CALDAS	-0.01	0.02	-0.65	0.52
‘Cantidad delitos’	0.11	0.01	10.18	0.00
Titulo11	-0.11	0.02	-4.90	0.00
Titulo41	-0.77	0.03	-25.40	0.00
Titulo71	0.43	0.02	25.75	0.00
Titulo91	-0.18	0.06	-3.16	0.00
Titulo101	-0.43	0.06	-7.47	0.00
Titulo121	-0.15	0.02	-7.38	0.00
Titulo161	0.43	0.05	8.25	0.00
ESTABLECIMIENTOCat2	5.40	27.26	0.20	0.84
ESTABLECIMIENTOCat3	6.08	27.26	0.22	0.82
Titulo2y31	-0.27	0.03	-9.08	0.00
Titulo5_6y81	-0.13	0.04	-2.97	0.00
Titulo11_13y141	-0.21	0.06	-3.75	0.00
Titulo15_17y181	-0.20	0.05	-4.15	0.00

Tabla 34: Coeficientes del Modelo Nulo Probit

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.86	0.01	-166.82	0.00

Tabla 35: Coeficientes del Modelo Optimizado Probit

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-106.97	134.50	-0.80	0.43
‘ESTADO INGRESO’DOMICILIARIA	13.17	139.21	0.09	0.92
‘ESTADO INGRESO’ESPERA TRASLADO	-0.03	192.30	-0.00	1.00
‘ESTADO INGRESO’INTRAMUROS	5.58	131.51	0.04	0.97
‘ESTADO INGRESO’PRISION DOMICILIARIA	-0.16	132.43	-0.00	1.00
‘ESTADO INGRESO’VIGILANCIA ELECTRONICA	5.55	131.51	0.04	0.97
‘Año ingreso’	0.05	0.00	13.54	0.00
ESTABLECIMIENTOCat2	5.38	27.32	0.20	0.84
ESTABLECIMIENTOCat3	6.06	27.32	0.22	0.82
Titulo71	0.51	0.01	34.38	0.00
DomicilioCat2	-0.33	0.02	-16.88	0.00
DomicilioCat3	-0.57	0.02	-32.00	0.00
RegionalNOROESTE	-0.14	0.02	-6.43	0.00
RegionalNORTE	0.21	0.03	7.78	0.00
RegionalOCCIDENTE	0.12	0.02	5.78	0.00
RegionalORIENTE	0.06	0.03	2.39	0.02
RegionalVIEJO CALDAS	-0.01	0.02	-0.28	0.78
Titulo41	-0.71	0.03	-24.97	0.00
EDAD	0.01	0.00	18.15	0.00
‘NIVEL EDUCATIVO’BASICA PRIMARIA	0.09	0.01	5.83	0.00
‘NIVEL EDUCATIVO’EDUCACION SUPERIOR	-0.60	0.05	-13.09	0.00
‘NIVEL EDUCATIVO’ILETRADOS	-0.02	0.03	-0.65	0.51
Titulo2y31	-0.15	0.03	-5.74	0.00
‘ACTIVIDAD OCUPACIONAL’ESTUDIO	0.25	0.07	3.69	0.00
‘ACTIVIDAD OCUPACIONAL’sin Dato	-0.61	0.07	-8.78	0.00
‘ACTIVIDAD OCUPACIONAL’TRABAJO	0.23	0.07	3.42	0.00
GENEROMasculino	0.23	0.03	8.70	0.00
Categori_Años_condena11 a 15	-0.22	0.02	-9.02	0.00
Categori_Años_condena16 a 20	-0.30	0.03	-11.89	0.00
Categori_Años_condena21 a 25	-0.22	0.04	-6.13	0.00
Categori_Años_condena26 a 30	-0.41	0.05	-8.68	0.00
Categori_Años_condena31 a 36	-0.50	0.05	-9.71	0.00
Categori_Años_condena6 a 10	-0.12	0.02	-6.85	0.00
Categori_Años_condenamayor a 36	-0.52	0.04	-11.65	0.00
Titulo161	0.51	0.05	10.04	0.00
Titulo101	-0.38	0.06	-6.72	0.00
Titulo15_17y181	-0.09	0.05	-1.89	0.06

Tabla 36: Coeficientes del Modelo Probit Balanceado Optimizado

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-132.13	194.31	-0.68	0.50
‘ESTADO INGRESO’DOMICILIARIA	13.55	192.72	0.07	0.94
‘ESTADO INGRESO’ESPERA TRASLADO	-0.00	284.02	-0.00	1.00
‘ESTADO INGRESO’INTRAMUROS	6.64	190.68	0.03	0.97
‘ESTADO INGRESO’PRISION DOMICILIARIA	-1.52	191.87	-0.01	0.99
‘ESTADO INGRESO’VIGILANCIA ELECTRONICA	6.84	190.68	0.04	0.97
‘Cantidad delitos’	0.25	0.01	37.91	0.00
Titulo161	0.90	0.02	38.75	0.00
‘Año ingreso’	0.06	0.00	20.16	0.00
ESTABLECIMIENTOCat2	7.75	36.93	0.21	0.83
ESTABLECIMIENTOCat3	8.35	36.93	0.23	0.82
Titulo91	0.83	0.03	32.00	0.00
Titulo101	0.83	0.02	36.01	0.00
Titulo71	0.48	0.01	38.91	0.00
DomicilioCat2	-0.35	0.02	-22.53	0.00
DomicilioCat3	-0.61	0.01	-42.60	0.00
EDAD	0.01	0.00	19.06	0.00
Categori_Años_condena11 a 15	-0.05	0.02	-2.68	0.01
Categori_Años_condena16 a 20	-0.30	0.02	-13.15	0.00
Categori_Años_condena21 a 25	-0.31	0.03	-9.65	0.00
Categori_Años_condena26 a 30	-0.44	0.04	-10.86	0.00
Categori_Años_condena31 a 36	-0.39	0.04	-9.49	0.00
Categori_Años_condena6 a 10	-0.06	0.01	-4.24	0.00
Categori_Años_condenamayor a 36	-0.45	0.04	-12.06	0.00
Titulo11_13y141	0.45	0.03	13.90	0.00
‘ACTIVIDAD OCUPACIONAL’ESTUDIO	0.37	0.06	6.48	0.00
‘ACTIVIDAD OCUPACIONAL’sin Dato	-0.02	0.06	-0.27	0.78
‘ACTIVIDAD OCUPACIONAL’TRABAJO	0.34	0.06	5.98	0.00
RegionalNOROESTE	-0.27	0.02	-14.89	0.00
RegionalNORTE	0.13	0.02	5.71	0.00
RegionalOCCIDENTE	0.01	0.02	0.75	0.45
RegionalORIENTE	-0.09	0.02	-4.17	0.00
RegionalVIEJO CALDAS	-0.04	0.02	-2.17	0.03
‘NIVEL EDUCATIVO’BASICA PRIMARIA	0.16	0.01	12.80	0.00
‘NIVEL EDUCATIVO’EDUCACION SUPERIOR	-0.23	0.03	-8.22	0.00
‘NIVEL EDUCATIVO’ILETRADOS	0.10	0.03	3.77	0.00
Titulo41	-0.49	0.02	-21.16	0.00
GENEROMasculino	-0.13	0.02	-7.35	0.00
Titulo2y31	0.12	0.02	6.47	0.00
Titulo15_17y181	0.26	0.03	9.23	0.00
Titulo121	-0.01	0.01	-1.02	0.31
Titulo11	-0.05	0.02	-3.09	0.00

Tabla 37: Coeficientes del Modelo Poisson Saturado

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-121.69	2655.43	-0.05	0.96
EDAD	0.00	0.00	6.07	0.00
‘Año ingreso’	0.04	0.00	9.58	0.00
DomicilioCat2	0.02	0.02	1.21	0.23
DomicilioCat3	0.04	0.02	2.01	0.04
‘ESTADO CIVIL’pareja	8.38	0.29	28.96	0.00
‘ESTADO CIVIL’Separados	8.26	0.29	28.09	0.00
‘ESTADO CIVIL’Soltero	8.33	0.29	28.77	0.00
GENEROMasculino	0.05	0.03	1.86	0.06
‘ESTADO INGRESO’DOMICILIARIA	13.96	2621.37	0.01	1.00
‘ESTADO INGRESO’ESPERA TRASLADO	-0.00	3792.82	-0.00	1.00
‘ESTADO INGRESO’INTRAMUROS	14.01	2621.37	0.01	1.00
‘ESTADO INGRESO’PRISION DOMICILIARIA	-1.75	2634.13	-0.00	1.00
‘ESTADO INGRESO’VIGILANCIA ELECTRONICA	13.91	2621.37	0.01	1.00
Categori_Años_condena11 a 15	-0.11	0.03	-4.13	0.00
Categori_Años_condena16 a 20	-0.15	0.03	-4.58	0.00
Categori_Años_condena21 a 25	-0.19	0.05	-4.03	0.00
Categori_Años_condena26 a 30	-0.26	0.06	-4.22	0.00
Categori_Años_condena31 a 36	-0.31	0.07	-4.32	0.00
Categori_Años_condena6 a 10	-0.07	0.02	-4.00	0.00
Categori_Años_condenamayor a 36	-0.30	0.06	-4.72	0.00
‘NIVEL EDUCATIVO’BASICA PRIMARIA	0.02	0.01	1.36	0.17
‘NIVEL EDUCATIVO’EDUCACION SUPERIOR	-0.07	0.05	-1.25	0.21
‘NIVEL EDUCATIVO’ILETRADOS	-0.01	0.04	-0.39	0.70
‘ACTIVIDAD OCUPACIONAL’ESTUDIO	0.02	0.09	0.22	0.82
‘ACTIVIDAD OCUPACIONAL’sin Dato	0.03	0.09	0.36	0.72
‘ACTIVIDAD OCUPACIONAL’TRABAJO	0.02	0.09	0.26	0.79
‘ULTIMA CALIFICACION CONDUCTA’Ejemplar	-0.01	0.02	-0.39	0.70
‘ULTIMA CALIFICACION CONDUCTA’Mala	0.05	0.04	1.14	0.26
‘ULTIMA CALIFICACION CONDUCTA’PENDIENTE	0.04	0.06	0.56	0.57
‘ULTIMA CALIFICACION CONDUCTA’Regular	0.04	0.05	0.83	0.40
‘ULTIMA CALIFICACION CONDUCTA’sin Dato	-0.10	0.02	-4.34	0.00
RegionalNOROESTE	0.03	0.02	1.29	0.20
RegionalNORTE	0.07	0.03	2.56	0.01
RegionalOCCIDENTE	0.02	0.02	0.98	0.32
RegionalORIENTE	0.07	0.03	2.55	0.01
RegionalVIEJO CALDAS	0.01	0.02	0.57	0.57
‘Cantidad delitos’	-0.02	0.01	-1.67	0.09
Titulo11	0.01	0.02	0.28	0.78
Titulo41	-0.13	0.04	-3.13	0.00
Titulo71	0.19	0.02	12.25	0.00
Titulo91	-0.04	0.05	-0.76	0.45
Titulo101	0.02	0.06	0.31	0.76
Titulo121	-0.06	0.02	-2.63	0.01
Titulo161	0.04	0.05	0.87	0.38
ESTABLECIMIENTOCat2	16.50	423.90	0.04	0.97
ESTABLECIMIENTOCat3	16.51	423.90	0.04	0.97
Titulo2y31	-0.03	0.03	-0.96	0.34
Titulo5_6y81	0.05	0.05	1.19	0.24
Titulo11_13y141	-0.03	0.06	-0.52	0.61
Titulo15_17y181	-0.03	0.05	-0.59	0.56

Tabla 38: Coeficientes del Modelo Poisson Nulo

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.22	0.01	-184.69	0.00

Tabla 39: Coeficientes del Modelo Poisson Optimizado

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-121.43	2655.83	-0.05	0.96
‘ESTADO CIVIL’pareja	8.38	0.29	28.97	0.00
‘ESTADO CIVIL’Separados	8.26	0.29	28.10	0.00
‘ESTADO CIVIL’Soltero	8.33	0.29	28.78	0.00
Categori_Años_condena11 a 15	-0.11	0.03	-4.30	0.00
Categori_Años_condena16 a 20	-0.15	0.03	-4.81	0.00
Categori_Años_condena21 a 25	-0.19	0.05	-4.19	0.00
Categori_Años_condena26 a 30	-0.26	0.06	-4.32	0.00
Categori_Años_condena31 a 36	-0.32	0.07	-4.45	0.00
Categori_Años_condena6 a 10	-0.07	0.02	-4.13	0.00
Categori_Años_condenamayor a 36	-0.31	0.06	-4.97	0.00
Titulo71	0.19	0.01	13.12	0.00
‘ESTADO INGRESO’DOMICILIARIA	13.97	2621.79	0.01	1.00
‘ESTADO INGRESO’ESPERA TRASLADO	-0.00	3793.34	-0.00	1.00
‘ESTADO INGRESO’INTRAMUROS	14.00	2621.79	0.01	1.00
‘ESTADO INGRESO’PRISION DOMICILIARIA	-1.75	2634.55	-0.00	1.00
‘ESTADO INGRESO’VIGILANCIA ELECTRONICA	13.92	2621.79	0.01	1.00
‘Año ingreso’	0.04	0.00	9.63	0.00
ESTABLECIMIENTOCat2	16.51	423.79	0.04	0.97
ESTABLECIMIENTOCat3	16.52	423.79	0.04	0.97
EDAD	0.00	0.00	6.20	0.00
Titulo121	-0.06	0.02	-2.98	0.00
‘ULTIMA CALIFICACION CONDUCTA’Ejemplar	-0.01	0.02	-0.40	0.69
‘ULTIMA CALIFICACION CONDUCTA’Mala	0.05	0.04	1.14	0.26
‘ULTIMA CALIFICACION CONDUCTA’PENDIENTE	0.04	0.06	0.56	0.58
‘ULTIMA CALIFICACION CONDUCTA’Regular	0.04	0.05	0.88	0.38
‘ULTIMA CALIFICACION CONDUCTA’sin Dato	-0.09	0.02	-4.32	0.00
RegionalNOROESTE	0.03	0.02	1.22	0.22
RegionalNORTE	0.07	0.03	2.51	0.01
RegionalOCCIDENTE	0.02	0.02	0.95	0.34
RegionalORIENTE	0.07	0.03	2.58	0.01
RegionalVIEJO CALDAS	0.01	0.02	0.56	0.58
Titulo41	-0.13	0.04	-3.19	0.00
‘Cantidad delitos’	-0.02	0.01	-2.47	0.01
GENEROMasculino	0.06	0.03	1.96	0.05
DomicilioCat2	0.03	0.02	1.34	0.18
DomicilioCat3	0.04	0.02	2.17	0.03

Tabla 40: Coeficientes del Modelo Binomial Negativa Saturado

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-131.99	19851.48	-0.01	0.99
EDAD	0.01	0.00	7.06	0.00
‘Año ingreso’	0.04	0.00	9.83	0.00
DomicilioCat2	0.04	0.02	2.08	0.04
DomicilioCat3	0.03	0.02	1.90	0.06
‘ESTADO CIVIL’pareja	7.67	0.21	37.40	0.00
‘ESTADO CIVIL’Separados	7.52	0.21	35.56	0.00
‘ESTADO CIVIL’Soltero	7.62	0.21	37.10	0.00
GENEROMasculino	0.06	0.03	2.06	0.04
‘ESTADO INGRESO’DOMICILIARIA	18.67	19589.26	0.00	1.00
‘ESTADO INGRESO’ESPERA TRASLADO	0.00	28009.17	0.00	1.00
‘ESTADO INGRESO’INTRAMUROS	18.71	19589.26	0.00	1.00
‘ESTADO INGRESO’PRISION DOMICILIARIA	-1.22	19689.04	-0.00	1.00
‘ESTADO INGRESO’VIGILANCIA ELECTRONICA	18.63	19589.26	0.00	1.00
Categori_Años_condena11 a 15	-0.11	0.03	-4.00	0.00
Categori_Años_condena16 a 20	-0.13	0.03	-3.80	0.00
Categori_Años_condena21 a 25	-0.19	0.05	-3.94	0.00
Categori_Años_condena26 a 30	-0.24	0.06	-3.79	0.00
Categori_Años_condena31 a 36	-0.27	0.07	-3.78	0.00
Categori_Años_condena6 a 10	-0.07	0.02	-4.16	0.00
Categori_Años_condenamayor a 36	-0.28	0.06	-4.34	0.00
‘NIVEL EDUCATIVO’BASICA PRIMARIA	0.02	0.01	1.11	0.27
‘NIVEL EDUCATIVO’EDUCACION SUPERIOR	-0.06	0.05	-1.10	0.27
‘NIVEL EDUCATIVO’ILETRADOS	-0.02	0.04	-0.57	0.57
‘ACTIVIDAD OCUPACIONAL’ESTUDIO	0.05	0.09	0.49	0.62
‘ACTIVIDAD OCUPACIONAL’sin Dato	0.07	0.09	0.77	0.44
‘ACTIVIDAD OCUPACIONAL’TRABAJO	0.05	0.09	0.54	0.59
‘ULTIMA CALIFICACION CONDUCTA’Ejemplar	-0.01	0.02	-0.36	0.72
‘ULTIMA CALIFICACION CONDUCTA’Mala	0.04	0.04	0.88	0.38
‘ULTIMA CALIFICACION CONDUCTA’PENDIENTE	0.01	0.06	0.14	0.89
‘ULTIMA CALIFICACION CONDUCTA’Regular	0.01	0.05	0.22	0.82
‘ULTIMA CALIFICACION CONDUCTA’sin Dato	-0.12	0.02	-5.27	0.00
RegionalNOROESTE	0.02	0.02	0.77	0.44
RegionalNORTE	0.06	0.03	1.95	0.05
RegionalOCCIDENTE	0.00	0.02	0.01	0.99
RegionalORIENTE	0.05	0.03	1.82	0.07
RegionalVIEJO CALDAS	-0.01	0.02	-0.54	0.59
‘Cantidad delitos’	-0.03	0.01	-2.27	0.02
Titulo11	0.01	0.02	0.25	0.80
Titulo41	-0.11	0.04	-2.69	0.01
Titulo71	0.20	0.02	12.22	0.00
Titulo91	-0.02	0.05	-0.35	0.72
Titulo101	0.00	0.06	0.04	0.97
Titulo121	-0.04	0.02	-1.95	0.05
Titulo161	0.07	0.05	1.43	0.15
ESTABLECIMIENTOCat2	20.45	3215.87	0.01	0.99
ESTABLECIMIENTOCat3	20.47	3215.87	0.01	0.99
Titulo2y31	-0.05	0.03	-1.44	0.15
Titulo5_6y81	0.03	0.05	0.59	0.55
Titulo11_13y141	-0.02	0.06	-0.25	0.80
Titulo15_17y181	-0.05	0.05	-0.97	0.33

Tabla 41: Coeficientes del Modelo Binomial Negativa Nulo

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.22	0.01	-138.23	0.00

Tabla 42: Coeficientes del Modelo inflado en ceros Poisson Nulo

	Estimate	Std. Error	z value	Pr(> z)
X.Intercept.	-0.10	0.01	-9.71	0.00
X.Intercept..1	0.72	0.01	49.15	0.00