IMPLEMENTACIÓN DE UN MODELO DE ANALÍTICA DE DATOS QUE EXPLICA LA REINCIDENCIA DELICTIVA DE PERSONAS CONDENADAS BAJO LA VIGILANCIA DEL INPEC

Ariel Mauricio Estrada Henao

Universidad de Manizales
Facultad de Ciencias e Ingeniería
Maestría en Gestión Estratégica de la Información
Manizales, Año 2022

IMPLEMENTACIÓN DE UN MODELO DE ANALÍTICA DE DATOS QUE EXPLICA LA REINCIDENCIA DELICTIVA DE PERSONAS CONDENADAS BAJO LA VIGILANCIA DEL INPEC

Ariel Mauricio Estrada Henao

Informe final de trabajo de grado presentado como requisito parcial para optar al título de Magíster en Gestión Estratégica de la información

Director:
Juan Alejandro Trujillo P.

Línea de Investigación:
Gestión del conocimiento

Universidad de Manizales
Facultad de Ciencias e Ingeniería
Maestría en Gestión Estratégica de la Información
Manizales, Año 2022

Para Mauricio y Amparo, por ser incondicionales.

A Luisa y Bambú, por acrecentar el sentido de la vida.

Resumen

Este estudio tiene como objetivo el entendimiento de la reincidencia delictiva partiendo de una serie de características o variables de las personas que fueron condenadas. Entender la reincidencia delictiva es muy relevante dado que ella agrava las cifras de sobrepoblación de los establecimientos de reclusión en Colombia, la cual se traduce en hacinamiento e impacto negativo del proceso de resocialización de las personas condenadas. Para desarrollar los objetivos del proyecto se utilizó la metodología de proyectos de minería de datos llamada crisp-dm, con ella se lograron entender y preparar los datos, modelar los diferentes algoritmos usados en la investigación y finalmente realizar análisis y evaluación de los resultados. Se pudo concluir que existen variables que impactan en mayor o menor medida la reincidencia delictiva, lo cual permitió clasificar correctamente tanto a personas reincidentes como no reincidentes con un acierto del 76%. Entender la reincidencia delictiva y cuales son aquellas variables que más inciden en ella podría aportar beneficios sociales para el estado mediante el acompañamiento gubernamental de los perfiles más vulnerables. También se podrían obtener beneficios económicos mediante la reducción de los índices de reincidencia logrando con ello, la disminución de la responsabilidad económica que representan las personas privadas de la libertad para el estado.

Palabras clave: Reincidencia Delictiva, Condena, Delito, Hacinamiento, Reclusión, Modelo Clasificatorio, Analítica de Datos, Machine Learning, crisp-dm.

Contenido

Pág.

Lista de Figuras

Pág.

Lista de tablas

Pág.

Lista de Símbolos y abreviaturas

Abreviaturas

Abreviatura	Término
INPEC	Instituto Nacional Penitenciario y Carcelario
ERON	Establecimientos de Reclusión de Orden Nacional
CRISP-DM	Cross-Industry Standard Process for Data Mining

Introducción

Colombia es un país con altos índices de desigualdad, violencia y criminalidad. Según Corporación Excelencia en la Justicia (2021), los índices de criminalidad vienen en aumento constante desde el año 2011. Por ende, las políticas públicas del país deben ir orientadas a mejorar las condiciones de vida de sus ciudadanos. Los recursos físicos y económicos que se utilicen para ello tienen el compromiso de aportar a la construcción de un mejor país. En materia Carcelaria y Penitenciaria los Establecimientos de Reclusión del Orden Nacional ERON a cargo del Instituto Nacional Penitenciario y Carcelario INPEC presentan problemas en diferentes aspectos. La sobrepoblación penitenciaria y carcelaria traducida en hacinamiento, es uno de los problemas más graves que tienen los establecimientos de reclusión (Acosta, 2021).

El índice de hacinamiento de los establecimientos de reclusión de orden nacional en los últimos siete años ha sido en promedio de 41% (INPEC, 2020), lo que ha generado que las condiciones mínimas de habitabilidad, convivencia, estudio, trabajo, enseñanza se vean directamente afectadas. Los problemas propios que conllevan la sobrepoblación afectan el proceso de resocialización de las personas condenadas, el cual es uno de los fines de la privación de la libertad.

La reincidencia delictiva es una muestra del trabajo que aún tiene por hacer el proceso de resocialización dentro de los establecimientos de reclusión. Pero dicho proceso no se puede llevar a cabo con éxito sin unas condiciones mínimas garantizadas dentro de las Penitenciarías.

Disminuir la reincidencia delictiva sería un aporte para rebajar los índices de hacinamiento carcelario y penitenciario, lo que a su vez permitiría ofrecer unas condiciones más dignas y favorables en el proceso de resocialización de una persona condenada. Para lograrlo se deben conocer cuáles son los factores que más impactan la reincidencia delictiva desde una perspectiva basada en los datos; que le permitan a futuro a las autoridades e instituciones competentes implementar acciones relacionadas con este objetivo.

La investigación se trabajó con una base de datos aportada por el Instituto Nacional Penitenciario y Carcelario INPEC que se solicitó a través de la plataforma de datos abiertos de Colombia www.datos.gov.co. La información aportada por el INPEC contiene más de medio millón de registros anonimizados que datan desde el año 1970 a la fecha. En los registros hay información relacionada con aspectos personales, jurídicos y locativos de las personas condenadas.

Para la investigación se implementó la metodología CRISP DM con la cual se logró comprender la problemática, entender los datos contenidos en la base de datos, preparar los datos según las necesidades de la investigacion, posterior a ello realizar modelados con diferentes técnicas y finalmente evaluar los resultados. La implementación no es una etapa concebida para esta investigación.

La investigación es de carácter académico y no pretende estigmatizar, ni segregar grupos poblacionales que históricamente hayan sido relacionados con la criminalidad. Procura únicamente desde un punto de vista científico, y basándose en los datos aportados por el INPEC conocer cuáles son las variables que más pueden influir para que una persona reincida en una conducta criminal.

Aspiramos que los resultados de la investigación sirvan como insumo futuro y aporten a la discusión en el momento de elaborar políticas penitenciarias y carcelarias, programas de intervención social y todas aquellas iniciativas en función de disminuir la criminalidad del país.

1. Planteamiento del problema de investigación y justificación

1.1 Descripción del área problemática

Según el decreto presidencial 1242 de 1993[1] el Instituto Penitenciario y Carcelario INPEC es la institución del gobierno encargada de garantizar la ejecución de penas, vigilancia, custodia, atención social y tratamiento de las personas privadas de la libertad. El cumplimiento de las condenas se lleva a cabo mediante sus diferentes modalidades.

En lo que respecta al cumplimiento de condena intramuros existía al momento de la realización de esta investigacion una sobrepoblación de 43.342 personas recluidas, que se traduce en un hacinamiento del 53,7%, lo cual se recrudece con la reincidencia delictiva que para febrero del año 2020 corresponde a 24.349 personas (INPEC, 2020).

Teniendo en cuenta que una de las obligaciones del gobierno colombiano con las personas privadas de la libertad es garantizar condiciones de reclusión dignas (Echeverry, 2017) el hacinamiento carcelario y penitenciario se convierte en un problema de políticas públicas. Uno de los objetivos de la privación de la libertad es la resocialización de las personas condenadas, la cual se logra a través del trabajo, estudio y enseñanza; actividades que además les permiten a los reclusos y las reclusas reducciones en su condena (Hernández, 2018). La reincidencia delictiva es un indicador que muestra un fallido intento del proceso resocializador. Las personas que regresan a un centro penitenciario, engrosan por su parte las cifras de hacinamiento.

La sobrepoblación penitenciara de personas recluidas de forma intramuros presenta, además del costo social ya descrito, un costo económico considerable para el estado. Gastos de funcionamiento como Personal, Adquisición de Bienes, Servicios, Logística, Salud, Alimentación, Seguridad, Impuestos, además de las inversiones hechas en Infraestructura, Tecnología, Programas de Atención a la Población Reclusa, entre otros; representan un costo económico anual de $31.179.764 por cada persona recluida (INPEC, 2022).

[1] https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=66836

En materia de políticas carcelarias y penitenciarias la Corte Constitucional mediante la sentencia T-153 de 1998 manifestó la gravedad de la situación decretando un Estado de Cosas Inconstitucional (ECI) en las prisiones colombianas que luego estuvo acompañado por la sentencia T-388 de 2013 y T-762 de 2015, señalando la vulneración masiva de derechos fundamentales de los reclusos y reclusas en Colombia, determinando además que el hacinamiento carcelario es el principal protagonista de la crisis penitenciaria (Huertas Díaz et al., 2014).

En un país con un sistema penitenciario colapsado, las personas reincidentes aumentan las cifras de sobrepoblación penitenciaria cada año. Esto se enmarca en un problema de orden social dado que el hacinamiento penitenciario interfiere con la resocialización de la población privada de la libertad y en un problema de orden económico entendiendo que, a mayor número de personas condenadas dentro de los centros penitenciarios y entre más altos índices de reincidencia delictiva, mayores serán los recursos que deba destinar el estado colombiano para tal fin.

1.2 Formulación del problema

¿A través de un modelo basado en analítica de datos es posible explicar la reincidencia delictiva de una persona condenada que haya estado bajo la vigilancia del Instituto Nacional Penitenciario y Carcelario INPEC?

1.3 Justificación

Colombia es un país que se constituye en un estado unitario, social y democrático de derecho. Cuya Estructura de Estado está compuesta por tres Ramas del Poder Público: Legislativa, Ejecutiva y Judicial; los órganos autónomos e independientes, la organización electoral, los organismos de control y el Sistema Integral de Verdad, Justicia, Reparación y No Repetición (Función Pública, 2019). Dichos componentes contribuyen al cumplimiento de las funciones del estado. Las tres ramas del poder público tienen un rol fundamental en el concepto de justicia del estado colombiano ya que como lo dijo la Corte Constitucional; la política criminal y penitenciaria es el resultado coordinado del trabajo de las ramas del poder público (Echeverry, 2017).

La Rama Legislativa es la encargada de formular las leyes, ejercer control sobre el Gobierno y reformar la Constitución; a la Rama Judicial le corresponde administrar justicia principalmente mediante sentencias, fallos, o autos; y la Rama Ejecutiva por su parte es la encargada de llevar a cabo las actividades administrativas disponibles a los intereses generales de la comunidad, dentro de ellas, garantizar a través de su Ministerio

de Justicia y del Derecho la ejecución de las sentencias condenatorias emitidas por los jueces de la república; misión que le corresponde llevar a cabo al Instituto Nacional Penitenciario y Carcelario INPEC.

El problema de hacinamiento no solo se manifiesta en la ausencia de condiciones dignas de infraestructura para el pago de las condenas. El proceso de resocialización de las personas condenadas; se ve afectado al no existir las condiciones locativas y de capital humano suficiente que garanticen la posibilidad de trabajo, estudio y enseñanza para todas y todos aquellos interesados en redimir su pena con esas alternativas. Al no poder participar en los programas orientados a la resocialización, la persona condenada puede ser persuadida a utilizar su tiempo libre en actividades delictivas dentro de los centros penitenciarios (Jiménez, 2017).

Aunado a lo anterior la reincidencia de conductas delictivas es otro factor que habla de los problemas que atraviesa el país en materia penitenciaria; según INPEC (2022) la población reincidente en los ERON a febrero del año 2022 correspondía al 20,9% del total de las 107.377 personas condenadas, esto corresponde a 22.444 personas reincidentes, que representa para el estado un gasto anual de 700 mil millones de pesos anuales aproximadamente. Partiendo del hecho de que las personas reincidentes, en teoría, ya habían atravesado un proceso resocializador, es un costo económico demasiado alto.

En el trabajo de Higuera & Gómez (2019) describen que la política penitenciaria tiene el importante reto de encontrar formas mediante las cuales la judicatura o el gobierno acepte o establezca medidas opcionales al encarcelamiento que contribuyan a reducir el hacinamiento, garantizándose el cumplimiento de los fines de la imposición de la condena. Se deben realizar esfuerzos en este sentido encaminados a mejorar la situación penitenciaria del país. Las autoridades encargadas deben procurar la formulación de una política criminal orientada a la reducción de la población penitenciaria, destacando la importancia de promover el uso de mecanismos de vigilancia diferentes al de la privación de la libertad en intramuros (Romero Rodríguez, 2017).

Desde la psicología, el derecho y otras disciplinas se han hecho estudios para definir aquellas variables que inciden en la reincidencia criminal, destacándose en el estudio de (Cuervo et al., 2017) el género, la edad, la personalidad entre otras, como variables predictoras en jóvenes. El consumo de drogas, el abandono de escolarización temprana, familias frágiles por ausencia de padres y la incorporación a economías informales fueron asuntos que resaltaron al momento de entender la reincidencia criminal desde una mirada netamente cualitativa (Ariza et al., 2020). Existen iniciativas desde diferentes áreas del conocimiento por definir las variables que mejor expliquen la

reincidencia delictiva, desde la ciencia de datos existe mucho camino por recorrer en este sentido. Entender la reincidencia delictiva partiendo netamente de los datos históricos relacionados, aportaría mucho a la discusión.

Disminuir el hacinamiento penitenciario mediante el pago de condenas usando modalidades diferentes a la privación de la libertad en intramuros, mejoraría las condiciones actuales en los establecimientos penitenciarios, lo cual permitiría llevar a cabo un mejor proceso resocializador. Así se lograría un impacto positivo desde lo social para las personas condenadas y desde lo económico para el estado colombiano. Esto partiendo de la lógica que entre mejor sea el proceso de resocialización serán menores los índices de reincidencia y los costos económicos asociados a ella.

Para lo anterior se necesita contar con herramientas científicas que permitan tomar decisiones basadas en los datos. Conocer la posibilidad de reincidencia delictiva de una persona condenada podría ser información muy relevante a la hora de definir nuevas políticas penitenciarias en las que se dé prioridad, por ejemplo, a la prisión domiciliaria. Saber cuál es la posibilidad de que una persona delinca nuevamente no solo aportaría en este sentido, también le permitiría al estado implementar estrategias en materia de desarrollo social focalizadas a las personas con mayor vulnerabilidad hacia la reincidencia; mediante campañas persuasivas o programas sociales orientados a dichas personas, que permitan disminuir el riesgo de que delincan nuevamente.

2. Objetivos

2.1 Objetivo general

Implementar un modelo basado en Analítica de Datos que explique la reincidencia delictiva de una persona que haya estado bajo la vigilancia del Instituto Nacional Penitenciario y Carcelario INPEC.

2.2 Objetivos específicos

• Realizar análisis exploratorio de la base de datos “PERSONAS CONDENADAS INPEC”

• Implementar un modelo clasificatorio que permita explicar la reincidencia delictiva en Colombia de las personas que hayan estado bajo la vigilancia del Instituto Nacional Penitenciario y Carcelario INPEC.

• Describir estrategias que puedan ayudar a disminuir la reincidencia delictiva en Colombia de las personas que se encuentren o hayan estado bajo la vigilancia del Instituto Nacional Penitenciario y Carcelario INPEC.

3. Referente Contextual

La investigación se realizó usando los registros de todos los establecimientos de reclusión de orden nacional del INPEC, los cuales están compuestos por las regionales Central, Viejo Caldas, Noreste, Norte, Oriente y Occidental. El detalle más amplio de cada regional se muestra en la Tabla 1.

Nº	NOMBRE DE LA REGIONAL	DEPARTAMENTO DÓNDE TIENE PRESENCIA
1	Central	Amazonas Boyacá Caquetá Bogotá D.C. Cundinamarca Huila Meta Tolima Casanare
3	Viejo Caldas	Caldas Quindío Tolima Risaralda Boyacá
2	Noroeste	Antioquia Chocó
4	Norte	Atlántico Bolívar Cesar Córdoba La Guajira Magdalena San Andrés Sucre
5	Oriente	Arauca Cesar Norte de Santander Santander
6	Occidental	Cauca Nariño Valle del Cauca

Tabla 1

Nº	NOMBRE DE LA REGIONAL	DEPARTAMENTO DÓNDE TIENE PRESENCIA
1	Central	Amazonas Boyacá Caquetá Bogotá D.C. Cundinamarca Huila Meta Tolima Casanare
3	Viejo Caldas	Caldas Quindío Tolima Risaralda Boyacá
2	Noroeste	Antioquia Chocó
4	Norte	Atlántico Bolívar Cesar Córdoba La Guajira Magdalena San Andrés Sucre
5	Oriente	Arauca Cesar Norte de Santander Santander
6	Occidental	Cauca Nariño Valle del Cauca

Tabla 1. Infraestructura de las regionales INPEC

El periodo de tiempo investigado corresponde entre el 15 de marzo de 1970 hasta el 2 de octubre del 2020; registros en la base de datos de primera y última fecha de ingreso de persona condenada.

3.1 Marco referencial

Los gobiernos y las instituciones que los conforman se encuentran en constante búsqueda de la mejora continua, la optimización de sus recursos y el impacto positivo hacía las comunidades. A lo largo del mundo se buscan mejoras en el que hacer de las instituciones, apoyándose en la tecnología a través del uso de la información.

La ciencia de datos ha emprendido un camino multidisciplinario a lo largo del mundo, aportando herramientas para resolver retos en muchas áreas del conocimiento. Los asuntos relacionados con justicia y castigo no han sido ajenos a ello. En la investigación realizada por (Chen et al., 2020) se utilizó un conjunto de datos del Tribunal Popular Supremo de China con el fin de predecir la pena que debe cumplir una persona basándose en los cargos que se le han imputado; en su investigación plantean algunas preocupaciones éticas tales como la discriminación por raza, clase social, edad, etc. que puede ser heredada por los modelos; además del problema que representan los nuevos delitos para los modelos de aprendizaje supervisados que trabajan con datos históricos. En este mismo sentido una predicción de condena basada en la descripción textual de los hechos es posible mediante el uso de modelos como el de regresión lineal y redes neuronales (S. Li et al., 2020).

La reincidencia delictiva es un asunto que causa interés a nivel global. El estudio realizado por (Fernández Monteiro, 2018) que buscaba predecir los factores de riesgo en menores infractores de España desde un enfoque de psicología criminal y haciendo uso de modelos de regresión logística concluyó que factores como el historial delictivo previo y actual, y el consumo de sustancias se correlacionan fuertemente con la reincidencia delictiva.

La seguridad es una premisa de gobiernos a nivel mundial. El uso de algoritmos de clasificación se ha implementado incluso para analizar y clasificar tweets relacionados con crimen que permitan una rápida respuesta de las autoridades (Tiwari et al., 2020). Así mismo la lucha contra delitos relacionados con drogas ha impulsado investigaciones usando modelos clasificatorios. La utilización de Arboles de Decisión ha permitido encontrar patrones espaciales y temporales relacionados con la detección y predicción de grupos delincuenciales que están involucrados con ciertos tipos de drogas (Xia et al., 2021).

No solo los modelos de procesamiento de datos estructurados y texto se han puesto en función de realizar mejoras en las políticas públicas de seguridad ciudadana. El procesamiento de imágenes se puede utilizar para el análisis de sentimiento orientado a la predicción de delitos como el robo de vehículos, carterismo, peleas, fraudes y disputas (Z. Li et al., 2021). Se pueden reducir las tasas de delincuencia gracias al uso de algoritmos de machine learning basándose en distintos tipos de delitos, los lugares donde se cometieron y la época en que fueron perpetrados (Vijayalakshmi, 2020).

De acuerdo con (Wang & Ma, 2021) mediante el uso de minería de datos a través de algoritmos como el Suport Vector Machine o el Random Forest se puede obtener información relevante que permita trabajar en la prevención de delitos contra la salud pública. De igual manera en la investigación llevada a cabo por (Silva et al., 2020) indican

que es posible hacer uso de técnicas de minería de datos como el Decision Tree para establecer patrones delictivos en lesiones causadas por accidentes, envenenamiento, asalto, traumas, obteniendo buena precisión en el modelo, especialmente para las muertes por homicidio.

Predecir la ocurrencia de conductas criminales en grupos focalizados como pacientes psiquiátricos es posible mediante el uso de variables clínicas y demográficas de las que se dispone antes del evento criminal; incluso con ellas se puede conocer la probabilidad de que cometan delitos no violentos o sexuales (Watts et al., 2021). En materia de delitos sexuales el trabajo de (Lussier et al., 2019) haciendo uso de modelos de clasificación, concluye que la edad del delincuente es una variable de riesgo importante para la predicción de la conducta y menciona además, la etapa de la carrera criminal como un factor de importancia a la hora del análisis.

Las herramientas que proporciona el uso de los datos para la generación de modelos clasificatorios y predictivos no solo aportan para conocer cuales variables son más importantes a la hora de prevenir una conducta criminal como la agresión sexual, con ellas además se puede conocer que factores pueden influir para que el delito sea más o menos grave. Así lo relatan (Reid & Beauregard, 2020) en su investigación, donde indican que factores como peleas entre agresor y víctima 48 horas previas al delito, posesión de armas por el agresor, la cercanía entre víctima y agresor son variables que empeoran o disminuyen la gravedad del delito.

El uso de machine learning tiene un amplio espectro de aplicabilidad que puede ir desde lo puntual como investigar grupos sociales focalizados por cierto tipo de conducta criminal como a lo global mediante el uso en políticas públicas. Kourtit et al (2021) en su trabajo realizan un análisis comparativo con datos relacionados con desempeño socioeconómico de 30 ciudades; logrando una categorización por clústeres basada en variables cuantitativas, concluyendo además que al parecer las políticas de seguridad ciudadana son importantes a la hora de predecir la variabilidad en el desarrollo urbano en general de una ciudad.

Si bien las técnicas relacionadas con machine learning son fundamentales a la hora de tomar decisiones basadas en datos, requieren en muchos casos de la interpretación de personas expertas en la materia. La implementación de modelos predictivos y clasificatorios, en campos donde la toma de decisiones va acompañada de gran responsabilidad, como lo es la vida o la muerte en la medicina, la libertad o condena en la justicia, requiere de la inclusión de personas con gran conocimiento de los temas (Qazi & Wong, 2019).

Se han realizado diferentes estudios relacionados con la reincidencia delictiva y si bien no hay una opinión unánime sobre cuál es el mejor método para predecirla, hay mucho potencial y trabajo por desarrollar en métodos como el gradient boosting[2]y el random forest[3] (Tollenaar & Van Der Heijden, 2019). Como lo indican (Ghasemi et al., 2021) los modelos de Machine Learning pueden arrojar mejores resultados en indicadores como el Área Bajo la Curva respecto a metodologías tradicionales de predicción criminal; mejorando la capacidad de clasificar a las personas en función de su probabilidad de reincidencia.

La implementación de procesos de machine learning en ocasiones acarrea consigo dificultades que pueden variar según cada escenario; tal como lo relata (Canhoto, 2020), las leyes, la forma para el intercambio de datos entre organizaciones y los costos, son elementos que pueden complejizar un ejercicio colaborativo de machine learning entre organizaciones. A nivel técnico también se deben guardar precauciones en lo metodológico con el fin de evitar sesgos con la discriminación algorítmica que se puede dar fácilmente entre grupos cómo nacional vs extranjero o joven vs viejo (Karimi-Haghighi & Castillo, 2021).

3.2 Marco teórico

3.2.1 Ciencia de Datos

La ciencia de datos es un área del conocimiento que combina diferentes disciplinas y cuyo principal objetivo es convertir los datos en valor real; estos pueden ser estructurados o no estructurados, en gran o poca cantidad, estáticos o en reproducción en línea y su valor puede ser dado a manera de predicciones, decisiones automatizadas, modelos entrenados o mediante visualización de datos que entreguen información de importancia (Van der Aalst, 2016). Los principios base de la ciencia de datos son la extracción de información y conocimiento a partir de datos teniendo así gran relación con el concepto de la minería de datos: la obtención real de conocimiento de los datos a través de tecnologías que incorporan estos principios (Provost & Fawcett, 2013).

2 Traducido del inglés: aumento de gradiente.
3 Traducido del inglés: bosque aleatorio.

Como apreciamos la ciencia de datos es entonces una disciplina que mezcla la experticia en varios temas de conocimiento y la cual ha tomado cada vez más importancia a nivel empresarial, gubernamental, académico y de otros intereses; en la Figura 1 que se muestra a continuación representa una definición ilustrada:

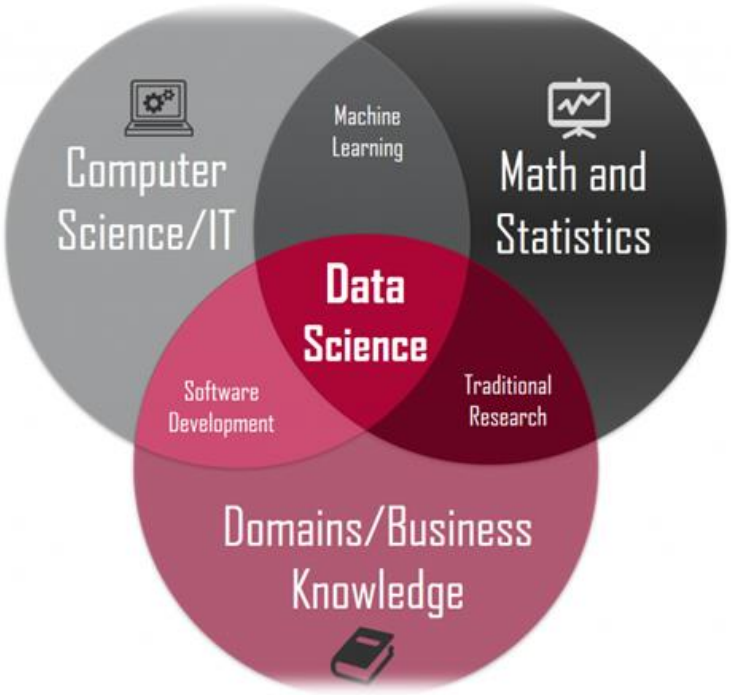

Figura 1. Skills Needed for Data Science. Fuente: (Medium, 2019)

Observamos la necesidad de habilidades en matemática y estadística (superior derecha) necesaria para el entendimiento de los datos, la realización de análisis descriptivos y el cálculo de probabilidades y regresiones. Habilidades requeridas en ciencias de la computación (superior izquierdo) para la manipulación de base de datos y consultas (SQL y otras), procesamiento de datos, herramientas de programación y visualización de datos y desarrollo e implementación de códigos en lenguajes de programación. Por último, dominio del conocimiento/ conocimiento del negocio (centro inferior) que hace referencia a la experticia necesaria en el campo o industria especifica que se está interviniendo; que ayudara al momento de entender los datos, plantear hipótesis y evaluar los resultados.

3.2.2 Metodología CRISP-DM

CRISP-DM (*Cross-Industry Standard Process for Data Mining*) es una metodología usada y probada en la industria con el fin de guiar el proceso de minería de

datos, la cual incluye descripciones de las fases comunes de un proyecto, las tareas que se deben desarrollar en cada fase y como se relacionan entre ellas; a manera general la metodología CRISP-DM provee una representación general del ciclo de vida de la minería de datos (IBM, 2015).

La metodología CRISP-DM se compone de seis etapas: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación e implementación. En la Figura 2 se puede observar la metodología de manera gráfica:

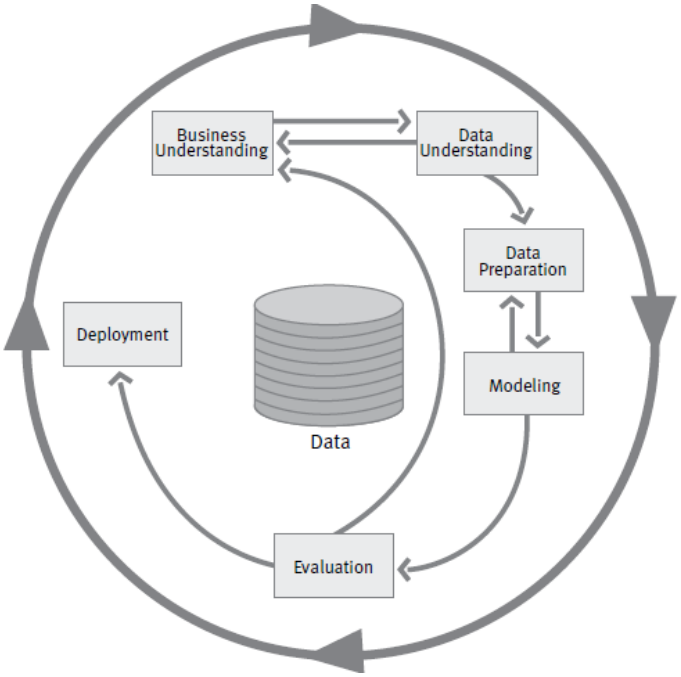

Figura 2. Phases of the CRISP-DM. Fuente: (Pete et al., 2000)

Como se puede observar en la Figura 2, si bien las etapas de la metodología siguen una secuencia ordenada, en algún punto del desarrollo se puede volver a una etapa anterior, esto con el fin de realizar, cambios, validaciones, postulados, mejoras y entender de mejor manera el estudio que se esté realizando.

Entendimiento del negocio

En esta fase se comprenden los objetivos y requisitos institucionales o empresariales del proyecto, y se establecen criterios de éxito del proyecto de minería de datos (de Graaf, 2019). Entender el problema que se requiere solucionar es fundamental

para la interpretación correcta de los resultados y garantiza que estos sean provechosos; esto acompañado de la realización de un plan de proyecto.

Entendimiento de los datos

Este paso comienza con la recolección inicial de los datos, ubicar su origen y definir su destino, estos pueden provenir de diferentes o únicas fuentes y extensiones; para su integración puede hacerse uso de procesos de ETL (Extract, Transform, Load) con el fin de conformar el Data Set[4] que se utilizará en el trabajo.

Para el entendimiento de los datos se debe hacer un acercamiento inicial que permita familiarizarnos y conocer cuáles son las variables y dimensiones que los componen (Gholamzadeh Nabati & Thoben, 2016). La exploración del Data Set es una actividad muy importante y se logra mediante el uso de estadística descriptiva que permita un mejor entendimiento de los datos; para ello se construyen tablas de frecuencias y gráficos de distribuciones, se observan medidas de dispersión como la varianza y medidas de tendencia central como la media, mediana y moda. Finalmente se hace una verificación de la calidad y completitud de los datos, se identifican valores atípicos de manera gráfica y se define cuales variables tienen valores nulos o faltantes.

Preparación de los datos

Esta fase se construye en función de lo que posteriormente se quiere realizar en el modelado, ya que dependiendo de los modelos o técnicas a utilizar los datos deberán tener una configuración específica, es la fase que consume mayor cantidad de tiempo.

Se inicia preparando los datos para que puedan ser usados en las técnicas de Data Mining que se utilicen posteriormente, sea mediante visualización, búsqueda de relaciones entre variables y otras que se requieran; esta fase incluya además la selección de datos para modelado, limpieza e imputación de datos, generación de variables nuevas, procesos de integración cuando existen diferentes fuentes de datos y cambios de formato (Arancibia, 2009).

[4] Traducido del inglés: Set de Datos.

El formateo de datos garantiza la utilización de los datos de acuerdo a las necesidades del modelado, técnicas como la desratización, label encoding y one hot encoding se utilizan comúnmente en esta actividad. La construcción de datos y la reducción de dimensionalidad son las últimas actividades de esta fase. Partiendo de hipótesis planteadas se construyen nuevas variables a partir de los datos existentes, las cuales buscan aportar información de mucha relevancia para la fase de modelado. La reducción de dimensionalidad permite mejorar la comprensión del fenómeno estudiado mediante la eliminación de variables que no aportan a la investigación tales como aquellas con tantas categorías como número de registros, variables constantes y variables de dudosa procedencia), existiendo también herramientas estadísticas que ayudan para tal fin.

Modelado

Como lo relata (Pete et al., 2000) existen diferentes algoritmos de machine learning para resolver problemas específicos. Por ello se usan diferentes técnicas (modelos) adecuadas y los parámetros de cada una se ajustan buscando las mejores combinaciones posibles; algunos modelos exigen formatos o características puntuales, por lo que, a menudo es necesario regresar a la fase de preparación de datos. Por último, se registra el rendimiento de los modelos implementados.

La escogencia de la técnica de modelado debe fundamentarse en: dominio de la técnica, uso apropiado, disposición de datos adecuados y tiempo disponible para modelado (Arancibia, 2009). El uso apropiado del modelo es muy importante, ya que por ejemplo si la investigación está relacionada con clasificar existen modelos específicos para esa labor tales como los arboles de decisión, el bosque aleatorio y el k-nearest neighbour.

Crear un plan de prueba, construir y evaluar el modelo son las actividades restantes de esta fase. Las características de los modelos dependen de los parámetros con los que se construyen, por ello se busca a través de la modificación iterativa de parámetros obtener los mejores resultados posibles en el modelado. Finalmente se evalúan los modelos en contexto con el problema que se está solucionando y actividad en la cual la experticia del campo se hace fundamental.

Evaluación

En esta fase, como lo indica (Schröer et al., 2021) los resultados de los modelos deben compararse con los objetivos del negocio definidos en la primera fase; estos deben ser interpretados en función del problema que se quería solucionar. Adicional a ello los modelos deben ser evaluados en función de las métricas de mayor interés para la investigación; existiendo un conjunto amplio de métricas y herramientas para la evaluación, dentro de las cuales se destacan comúnmente la matriz de confusión, la exactitud, la precisión, la sensibilidad (f1scrore) y el área bajo la curva.

Implementación

Por último, durante la fase de desarrollo, los conocimientos adquiridos con los modelos implementados se debaten con el usuario y se implementan para dar solución al problema inicialmente planteado (Gholamzadeh Nabati & Thoben, 2016). Así mismo lo indican (Wirth & Jochen, 2000) en su investigación, quienes mencionan que el conocimiento adquirido deberá estructurarse y socializarse de tal forma que el cliente pueda utilizarlo. Esta fase puede ser tan sencilla como generar un informe o tan complicada que requiera la implementación de un proceso de minería de datos repetible.

3.2.3 Machine Learning

Machine Learning tiene como objetivo fundamental la extracción de conocimiento a partir de los datos utilizando campos del conocimiento como la estadística, la inteligencia artificial y la informática; es una herramienta más común de los que creemos y está presente en la vida cotidiana desde recomendaciones de películas para ver, comida para ordenar, hasta reconocimiento facial de amigos en fotografías (Müller & Guido, 2016).

Tipos de Analítica

Mediante el machine learning pueden hacerse varios tipos de análisis. Como indica (Swamynathan, 2019) el análisis descriptivo se encarga de mostrarnos lo que ya ocurrió, su utilidad se basa en ayudarnos a comprender comportamientos pasados; el análisis predictivo tiene como finalidad hacer predicciones o calcular probabilidades de

eventos futuros partiendo de patrones históricos; por último el análisis prescriptivo cuyo principal objetivo es hallar la mejor ruta de acción para una situación determinada que le permita a las personas encargadas de la toma de decisiones anticiparse a posibles resultados previo la toma de las decisiones reales, este análisis se relaciona con los mencionados anteriormente.

Aprendizaje supervisado

El aprendizaje supervisado es un método usado principalmente en problemas de clasificación donde la variable que se quiere clasificar es binaria como Si/No, Enfermo/Sano etc. Este método funciona con dos variables diferentes, se usan de manera general datos de entrenamiento que funcionan como entrada (x) y etiquetas como salida (y), de esta manera el algoritmo de aprendizaje aprende una función de mapeo desde la entrada (x) a la salida (y); su objetivo principal es predecir la etiqueta de salida (y) para los nuevos datos ingresados en el modelo (Prabhu et al., 2019). En este tipo de aprendizaje se utilizan típicamente modelos como el árbol de decisiones, bosques aleatorios y máquina de soporte vectorial.

Aprendizaje no supervisado

Según (Igual & Seguí, 2020) esta metodología puede definirse como la actividad realizada por algoritmos que aprenden partiendo de un conjunto de entrenamiento sin etiquetar, haciendo uso de las características de las entradas para categorizarlas de acuerdo con algunos criterios geométricos o estadísticos; el aprendizaje no supervisado comúnmente es utilizado para resolver problemas de agrupación, reducción de dimensionalidad, detección de valores atípicos y detección de novedades.

4. Referente Normativo y legal

El desarrollo de la presente investigación tendrá como marco de referencia la ley 599 de 2000, la cual dispone las normas rectoras de la Ley Penal Colombiana, siendo esta una fuente del derecho existente en el ordenamiento jurídico Colombiano, asimismo, se entrelaza entre sí y se dota de validez junto con la Constitución Política de 1991, tal como lo plantea (Kelsen, 2009): *"una norma pertenece, pues, a un orden determinado únicamente cuando existe la posibilidad de hacer depender su validez de la norma fundamental que se encuentra en la base de este orden."*

Colombia al configurarse como Estado social y democrático de Derecho, ha estado sometida a diferentes cambios legislativos penales de índole procesal y sustancial, registrados entre el procedimiento anterior y posterior a la Constitución de 1991, en ese sentido, se utilizaron como fundamento normativo las reformas que se relacionarán a continuación:

- Ley 75 de 1968. De la filiación, la investigación de la paternidad y los efectos del Estado Civil.
- Ley 30 de 1986. Por la cual se adopta el Estatuto Nacional de Estupefacientes.
- Artículo 246 de la Constitución Política de 1991. De las Jurisdicciones Especiales de los Pueblos Indígenas.
- Ley 44 de 1993. Por la cual se modifica y adiciona la Ley 23 de 1982 (derechos de autor) y se modifica la Ley 29 de 1944 (se dictan disposiciones sobre prensa).
- Ley 99 de 1993. Por la cual se crea el Ministerio del Medio Ambiente, se reordena el Sector Público encargado de la gestión y conservación del medio ambiente y los recursos naturales renovables, se organiza el Sistema Nacional Ambiental, SINA y se dictan otras disposiciones.
- Decreto 1900 de 2002. Por el cual se adoptan medidas en materia penal y procesal penal contra las organizaciones delincuenciales.
- Ley 906 de 2004. Por la cual se expide el Código de Procedimiento Penal.
- Ley 1263 de 2008. Por medio de la cual se modifica parcialmente los artículos 26 y 28 de la Ley 99 de 1993.
- Ley 1851 de 2017. Por medio de la cual se establecen medidas en contra de la Pesca Ilegal.

- Ley 1407 de 2019. Por la cual se expide el Código Penal Militar.
- Ley 2111 de 2021. De los delitos contra los recursos naturales y el medio ambiente de la ley 599 de 2000, se modifica la ley 906 de 2044 y se dictan otras disposiciones.
- Ley 599 de 2000. Por la cual se expide el Código Penal.
- Ley 600 de 2000. Por la cual se expide el Código de Procedimiento Penal

Por otro lado, para lo relacionado con manejo de la información se trabajó bajo las directrices impartidas en el CONPES[5] 3920 que define la política de explotación de datos (Big Data) en el Estado colombiano. Esta política pública habilita el aprovechamiento de datos para generar desarrollo social y económico.

La política presentada en este documento busca superar la carencia de una visión a largo plazo y la tendencia al cumplimiento mínimo de mandatos legales, con los mecanismos y herramientas que procuren las transformaciones institucionales para la generación de valor, entendido como la provisión de bienes públicos para brindar respuestas efectivas y útiles frente a las necesidades sociales. En esta medida, la explotación de datos se supedita al cumplimiento de los fines del Estado.

Cuando se hace mención a la explotación de datos o el aprovechamiento de datos para la generación de valor social y económico, debe entenderse que excluye aquellos cuyo tratamiento se encuentra proscrito. Concretamente, se refiere a aquellos datos que se encuentran bajo el reconocimiento de la protección de los datos personales sometidos a los principios definidos en la Ley 1581 de 2012.

[5] Consejo Nacional de Política Económica y Social.

5. Hipótesis

A través de un modelo basado en analítica de datos es posible explicar la reincidencia delictiva de una persona condenada que haya estado bajo la vigilancia del Instituto Nacional Penitenciario y Carcelario INPEC

6. Metodología

6.1 Enfoque metodológico

Esta investigación se abordó desde un enfoque metodológico cuantitativo en el cual se obtuvieron los datos de una sola fuente de información correspondiente a un archivo de extensión .xlsx obtenido mediante correo electrónico; la investigación se soportó en la metodología CRISP-DM la cual es una de las más comunes en proyectos de data mining[6].

6.2 Tipo de estudio

La investigación se enmarcó en un tipo de estudio correlacional y explicativo donde se pretendía encontrar la relación existente entre un grupo determinado de variables y la variable objetivo, buscando además entender de mejor manera las causas o factores que afectan la reincidencia delictiva. Además de conocer la relación entre variables se pretendía cuantificar la importancia de cada variable frente al fenómeno estudiado. El alcance de la investigación por tanto llegó hasta la identificación de relaciones y clasificación de la reincidencia, invitando para que sea punto de partida en trabajos futuros que aborden fases de implementación y relacionadas.

6.3 Diseño de la investigación

El desarrollo de esta investigación se abordó desde la metodología para proyectos de data mining llamada CRISP-DM. Es importante resaltar que el alcance del estudio llegó hasta la fase cinco de Evaluación, siendo la siguiente fase de implementación una oportunidad para desarrollar trabajos futuros relacionados.

[6] Traducido del inglés: Minería de datos

6.3.1 Fase O1: Análisis exploratorio de la base de datos

6.3.1.1 Entendimiento del negocio

Primera fase de la metodología CRISP-DM. En esta fase se pretendía entender el contexto general y especifico relacionado con el fenómeno que íbamos a estudiar; para entender de mejor manera la reincidencia delictiva se debía comprender el papel de la o las instituciones involucradas y los objetivos a nivel institucional que podrían desprenderse de esta investigación. Por ello se investigó su función principal como institución y sus objetivos institucionales.

Determinación de los objetivos institucionales

Fue necesario definir los objetivos institucionales de la investigación con el fin de aterrizar la investigación a un plano menos académico y más institucional. Estos se plantearon en función de la naturaleza y función principal del INPEC.

Evaluación de la situación

Con el fin de obtener un panorama real en términos de recursos disponibles, tiempos, alcance, posibles dificultades y ventajas, se enlistan una serie de elementos que se consideraron relevantes para la investigación a realizar.

Determinar objetivos de la minería de datos

Se plantearon los objetivos de la minería de datos en relación con lo que se pretendía desarrollar desde el objetivo general y los objetivos específicos. Estos objetivos van ligados al éxito de la investigación por lo que su planteamiento y consecución es de gran importancia.

Producir un plan de proyecto

El plan de proyecto busca dar un orden secuencial al trabajo desde una mirada amplia y general. Para esta investigación se tuvo el plan de proyecto que se muestra en la Figura 3:

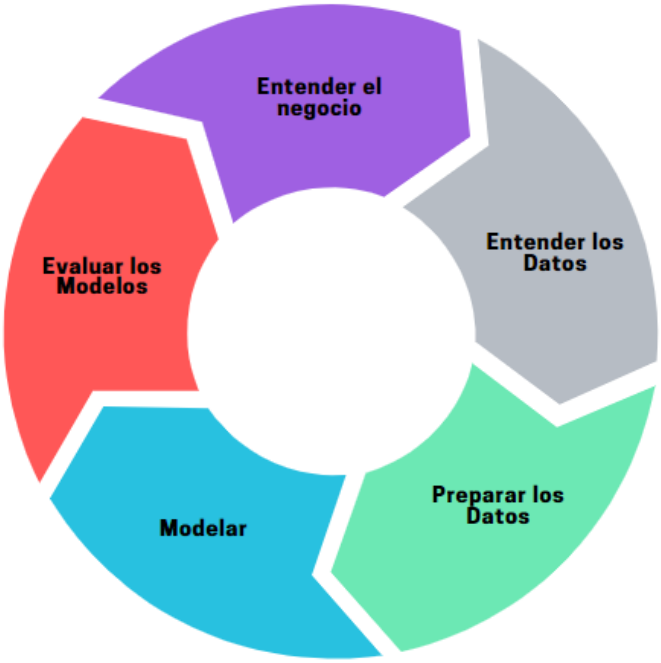

Figura 3. Plan de Proyecto. Fuente: construcción propia

6.3.1.2 Entendimiento de los datos

Posterior a la fase de entendimiento del negocio se encuentra la segunda fase, el entendimiento de los datos. Con ella se pretendía conocer de una manera más extensa y precisa los datos de los que se disponían. Esto se logró a través de análisis estadísticos básicos como medidas de tendencia central (media, mediana, moda), varianzas, frecuencias y también mediante el uso de gráficos bivariados y multivariados. A su vez proporcionó información importante sobre las variables que componen los datos y su naturaleza.

Para el análisis bivariado se construyeron histogramas de frecuencia y graficas de barras en porcentajes, además para los análisis multivariados se usaron gráficos de cajas para analizar combinaciones entre variables categóricas y continuas, gráficos de dispersión, densidad y barras.

Recolección inicial de datos

La investigación parte de una sola fuente de datos contenida en un documento Excel, por lo que no requiere hacerse ningún proceso de integración con extracción, transformación y carga de datos.

Descripción de los datos

Se analizó la totalidad de la base de datos con la que se cuenta haciendo un recuento total de registros y revisando totas las variables que la componen. Además, se definió la naturaleza de cada variable y las categorías que contiene cada variable.

Mediante el conteo de valores únicos, frecuencias y valores más comunes se analizaron las variables categóricas logrando entender mejor su naturaleza y comportamiento. Unido a lo anterior, se utilizaron técnicas de gráficos bivariados y multivariados como gráficos de frecuencias y boxplot para observar comportamientos y posibles relaciones entre variables.

Para las variables continuas se hicieron análisis de medidas de tendencia central como la media, visualización de distribución por cuartiles, desviación estándar de los datos. También nos apoyamos en análisis gráficos multivariados como el de distribución de los datos.

6.3.2 Fase O2: Implementación de modelo clasificatorio que permita explicar la reincidencia delictiva en Colombia de las personas que hayan estado bajo la vigilancia del Instituto Nacional Penitenciario y Carcelario INPEC

6.3.2.1 Preparación de los datos

En esta fase de la investigación se busca que los datos cumplan con las necesidades técnicas para su posterior modelado y evaluación. Por ello y según los análisis en fases previas se deben preparar, formatear y construir algunas variables.

6.3.2.2 Modelado

En la fase de modelado se preparan y configuran los modelos clasificatorios que se van a utilizar. Se toman como insumo las variables formateadas según los requerimientos de la investigación.

Diseño de la prueba

Se realizaron las construcciones necesarias para la implementación de los diferentes modelos. Tales como el listado de variables a utilizar, set de entrenamiento (Train) y el set de prueba (Test) entre otras.

Creación de la Variable X y Y

Se crearon la variable X que está compuesta por todas aquellas variables que se usarían en el modelado según las necesidades y criterios de la investigación y la variable Y que corresponde a la variable objetivo la Reincidencia.

Creación de Test y Train

Con el fin de entrenar los modelos y posteriormente testear los resultados, se dividió la base de datos original en dos conjuntos llamados test y train en una proporción de 20% y 80% respectivamente.

Balanceo de datos

Se verificó el estado de balanceo de los datos respecto la variable objetivo. Luego de ello y según la cantidad de registros disponibles se utilizó la técnica de balanceo más apropiada que para la investigación fue undersampling. Esta técnica usa la clase mayoritaria para el balanceo de los datos (Liu et al., 2009).

Modelos Implementados

Se implementaron los modelos de clasificación Regresión logística, Random Forest, Decision Tree, Naive vayes, Stochastic Gradient Descent, Gradient Boosted Machines GBM, K-Nearest Neighbors y Support Vector Machine SVM.

Se utilizaron los modelos anteriormente descritos dado el diseño de la investigación y la variable objetivo o respuesta que se tenia para ella. La variable reincidencia delictiva en la base de datos tiene dos únicos valores: Si y No, que posteriormente se transformarían en 1=Si, 0=No por lo que se buscó clasificar a las personas analizadas en alguno de estos dos grupos, necesitando para ello, modelos de tipo clasificatorio.

6.3.2.2 Evaluación

En la investigación se utilizaron diferentes métricas para evaluar los modelos, una de ellas la matriz de confusión. Partiendo de la matriz de confusión se analizaron métricas tales como exactitud, sensibilidad, precisión y especificidad. También se usó el área bajo la curva como indicador para medir los resultados de cada modelo.

6.3.3 Fase O3: Descripción de estrategias que puedan ayudar a disminuir la reincidencia delictiva en Colombia

Luego de obtenidos los resultados de los modelos clasificatorios, se analizan las variables seleccionadas y se proponen estrategias para mitigar el riesgo de reincidencia.

7. Resultados

7.1 Fase O1: Análisis exploratorio de la base de datos

7.1.1 Entendimiento del negocio

El Instituto Penitenciario y Carcelario INPEC es la institución del gobierno encargada de garantizar la ejecución de penas, vigilancia, custodia, atención social y tratamiento de las personas privadas de la libertad. Los objetivos institucionales del instituto penitenciario y carcelario INPEC según (Decreto 1242, 1993) son:

1. Ejecutar y desarrollar la política carcelaria y penitenciaria dentro de los lineamientos que establezca el Gobierno Nacional.

2. Hacer cumplir las medidas de aseguramiento, las penas privativas de la libertad y las medidas de seguridad, que establezcan las autoridades judiciales.

3. Diseñar y ejecutar programas de resocialización, rehabilitación y reinserción a la sociedad, para los reclusos de los establecimientos carcelarios y penitenciarios.

4. Diseñar y establecer los mecanismos necesarios de control de los programas de resocialización, rehabilitación y reinserción de los internos a la sociedad.

Determinación de los objetivos institucionales

Fue necesario definir los objetivos institucionales de la investigación Se definieron los objetivos institucionales generales y específicos con el fin de aterrizar la investigación a un plano menos académico y más institucional. Con esto se logró además entender la utilidad práctica del desarrollo del trabajo investigativo.

1. Objetivos institucionales de la investigación:

Disminuir la reincidencia delictiva a través de las variables accionables que más la impacten.

2. Objetivo del área institucional interesada:

Conocer las causas de la reincidencia delictiva con el fin de plantear estrategias a futuro que permita disminuirla.

Evaluación de la situación

Luego de hacer un análisis detallado del contexto general en el que se desarrollaría la investigación se obtuvieron los elementos relevantes que se debían tener en cuenta y que se mencionan a continuación:

- El dataset para realizar la investigación provenía de una sola fuente de datos del Instituto Penitenciario y Carcelario INPEC.

- Se utilizaron herramientas de código abierto como Python, Jupyter y Google Colab para las fases de Conocimiento de los Datos, Preparación de los Datos, Modelado y Evaluación.

- La mayoría de documentación científica relacionada y librerías de código abierto se encontraron en idioma inglés.

- La investigación solo se llevó hasta la fase de Evaluación, la implementación se podría realizar en trabajos futuros.

- Se debía realizar la investigación en un periodo de 12 meses.

Determinar objetivos de la minería de datos

Los objetivos de la minería de datos se dieron en función de qué se pretendía lograr a nivel técnico en el transcurso del proyecto. Son estos mismos objetivos los que fueron planteados como objetivo general y específicos del trabajo de grado.

- Describir las características de las personas que reinciden delictivamente.

- Clasificar las personas a cargo del INPEC según su probabilidad de reincidir.

7.1.2 Entendimiento de los datos

En la Tabla 2 se pueden observar las variables que componen el conjunto de datos con el que se trabajó la investigación:

Nº	NOMBRE DE LA VARIABLE	DESCRIPCIÓN	CATEGORIAS
1	**INTERNO ENCRIPTADO**	Numero único de identificación de la persona interna	Gran número de categorías, para detallarlas ver anexo A.
2	**DELITO**	Tipo de delito cometido	Gran número de categorías, para detallarlas ver anexo A.
3	**CLASIFICACION LUHMAN DELITO**	Agrupación de delito cometido según Teoría Luhman	Gran número de categorías, para detallarlas ver anexo A.
4	**CLASIFICACION JURIDICA DELITO**	Agrupación de delito cometido según el Código Penal	Gran número de categorías, para detallarlas ver anexo A.
5	**TENTATIVA**	Indica si el delito tiene Tentativa o No	S - N
6	**AGRAVADO**	Indica si el delito es Agravado o No	S - N
7	**CALIFICADO**	Indica si el delito es Calificado o No	S - N
8	**FECHA_INGRESO**	Fecha de Ingreso al establecimiento de reclusión	Gran número de categorías, para detallarlas ver anexo A.
9	**FECHA_SALIDA**	Fecha de Salida del establecimiento de reclusión	Gran número de categorías, para detallarlas ver anexo A.
10	**CONDENA CUMPLIDA MESES**	Tiempo en meses de la condena cumplida	Variable continua
11	**CONDENA INICIAL CUMPLIDA**	Comparación entre el tiempo de condena proferido y el tiempo cumplido realmente	SI – NO – En prisión
12	**FECHA_CAPTURA**	Fecha en que fue capturada la persona	Gran número de categorías, para detallarlas ver anexo A.
13	**ESTADO**	ALTAS= En reclusión, BAJAS= Ya no está en reclusión	ALTAS – BAJAS
14	**TIPO_SALIDA**	Motivo de salida del recluso del sistema penitenciario	Sin Salida - Libertad por Autoridad - Articulo 70 - Baja por Muerte - Informe Terminación - Baja por Fuga - Oficio Remisorio - Extradición
15	**SITUACION_JURIDICA**	Indica si es una persona Condenada o Sindicada	Condenado
16	**ANO_NACIMIENTO**	Año en el que nació la persona	Variable continua
17	**GENERO**	Indica el género de la persona	Masculino - Femenino
18	**PAIS_INTERNO**	País de origen de la persona	Gran número de categorías, para detallarlas ver anexo A.
19	**DEPARTAMENTO**	Departamento de origen de la persona	Gran número de categorías, para detallarlas ver anexo A.
20	**CIUDAD**	Ciudad de origen de la persona	Gran número de categorías, para detallarlas ver anexo A.
21	**MESES_CONDENA**	Tiempo de la condena dada en Meses	Variable continua
22	**REINCIDENTE**	Indica si la persona es Reincidente o No	SI – NO
23	**ESTADO_INGRESO**	Indica el tipo de reclusión de la persona	Prisión Domiciliaria - Intramuros - Detención Domiciliaria - Vigilancia Electrónica - Espera Traslado
24	SISTEMA_PENAL_ACUSATORIO	Indica el tipo de sistema penal acusatorio	inquisitivo - NSPA
25	**ACTIVIDADES_TRABAJO**	Indica si la persona Trabaja o No, mientras cumple la condena	SI – NO
26	**ACTIVIDADES_ESTUDIO**	Indica si la persona Estudia o No, mientras cumple la condena	SI – NO
27	**ACTIVIDADES_ENSENANZA**	Indica si la persona Enseña o No, mientras cumple la condena	SI – NO

28	**NIVEL_EDUCATIVO**	Nivel de escolaridad de la persona	ANALFABETA - CICLO I - CICLO II - CICLO III - CICLO IV - TECNICO - TECNICO PROFESIONAL - TECNOLOGICO PROFESIONAL - POST GRADO - ESPECIALIZACION - MAGISTER
29	**HIJOS_MENORES**	Indica si tiene Hijos Menores de Edad o No	SI – NO
30	**CONDIC_EXPECIONAL**	Indica el grupo poblacional excepcional al que pertenece la persona	SIN COND EXCEPCIONAL - INDIGENA AFRO - CON DISCAPACIDAD - ADULTO MAYOR – EXTRANJERIA – BISEXUAL - OTRAS MENOS FRECUENTES - HOMOSEXUAL
31	ESTABLECIMIENTO	Establecimiento donde la persona está cumpliendo la condena	Gran número de categorías, para detallarlas ver anexo A.
32	**REGIONAL**	Indica la Regional a la cual pertenece el Establecimiento	NORTE - CENTRAL - OCCIDENTE – NOROESTE - VIEJO CALDAS - ORIENTE

Tabla 2. Convención de Variables

El conjunto de datos está compuesto por treinta y dos variables dentro de las cuales hay de tipo categórico y continuo. La variable INTERNO_ENCRIPTADO tiene gran cantidad de categorías al tratarse de un número de identificación interno, por lo que se debe tener en cuenta al modelar su comportamiento. Para la variable SITUACION_JURIDICA tenemos una sola categoría “Condenado” ya que la investigación se está realizando con datos de personas a las que ya se les profirió una condena.

Variables continuas: ANO_NACIMIENTO y MESES_CONDENA. En la Tabla 3 podemos observar algunas medidas estadísticas y el comportamiento por cuartiles de las variables.

	ANO_NACIMIENTO	**MESES_CONDENA**
count	614.908	614.916
mean	1980	93,503
std	11,724	103,724
min	1920	0
25%	1973	33
50%	1982	56
75%	1989	108
max	2002	1920

Tabla 3. Variables continuas

Se observa que la cantidad de registros que tiene la base de datos son 614.916, siendo un número importante y provechoso para el tipo de estudio que se quiere realizar. La variable MESES_CONDENA presenta un valor mínimo de cero y máximo de 1920; ambos valores son atípicos ya que al momento de realizar esta investigación la condena máxima en Colombia es de 60 años lo cual se traduce en 720 meses.

Variables Categóricas: El proyecto en su mayoría está compuesto por variables categóricas, en las tablas que se muestran a continuación se puede detallar el conteo de registros por variable, la cantidad de valores únicos, el valor que más se repite y su frecuencia.

	INTERNO ENCRIPTADO	DELITO	CLASIFICACION LUHMAN DELITO
count	614916	614916	614916
unique	398106	342	8
top	C8048DBF9B0138DF5107C210BBB93955646CBB2A	HURTO	Premeditación de la afectación social
freq	52	142611	275090

Tabla 4. Variables categóricas

Se observa que existen 398106 registros únicos de la variable INTERNO_ENCRIPTADO, lo que quiere decir que hay ese número de personas en la base de datos. La variable DELITO tiene 342 categorías únicas con el Hurto como la más frecuente, siendo esta variable de especial importancia se debe hacer uso de alguna de sus variables complementarias CLASIFICACION LUHMAN DELITO y CLASIFICACION JURIDICA DELITO para su posterior análisis.

	CLASIFICACION JURIDICA DELITO	TENTATIVA	AGRAVADO	CALIFICADO	FECHA_SALIDA
count	614916	614916	614916	614916	614916
unique	20	2	2	2	5252
top	DELITOS CONTRA EL PATRIMONIO ECONÓMICO	N	N	N	En prision
freq	163537	581353	450943	502721	170732

Tabla 5. Variables categóricas

La categoría de delito más frecuente en Colombia es DELITOS CONTRA EL PATRIMONIO ECONÓMICO representando el 27% de los datos. La mayoría de los registros presentan las condiciones TENTATIVA, AGRAVADO, CALIFICADO.

	CONDENA CUMPLIDA MESES	CONDENA INICIAL CUMPLIDA	ESTADO	TIPO SALIDA	SITUACION JURIDICA	GENERO
count	614916	614916	614916	614052	614916	614916
unique	4467	3	2	8	1	2
top	En prision	NO	BAJA	Libertad por Autoridad	Condenado	MASCULINO
freq	170732	388902	442971	440350	614916	551637

Tabla 6. Variables categóricas

El motivo de salida más frecuente de los centros de reclusión es Libertad por Autoridad y el género que más ha sido condenado es el MASCULINO representando el 90% de los datos.

	CONDIC_EXPECIONAL	ESTABLECIMIENTO	REGIONAL
count	614916	614916	614916
unique	9	191	6
top	SIN COND EXCEPCIONAL	COMPLEJO CARCELARIO Y PENITENCIARIO METROPOLIT...	CENTRAL
freq	561516	51479	213694

Tabla 7. Variables categóricas

Se observa en la a mayoría de personas no tienen ninguna condición excepcional, el establecimiento de reclusión más frecuente es el COMPLEJO CARCELARIO Y PENITENCIARIO METROPOLITANO DE BOGOTA que pertenece a la regional CENTRAL del INPEC.

	ACTIVIDADES_ TRABAJO	ACTIVIDADES_ ESTUDIO	ACTIVIDADES_ ENSENANZA	NIVEL_ EDUCATIVO	HIJOS_ MENORES
count	614916	614916	614916	614916	614916
unique	2	2	2	12	2
top	SI	SI	NO	CICLO II	SI
freq	309942	350115	600194	248050	494939

Tabla 8. Variables categóricas

La mayoría de las personas realiza ACTIVIDADES_TRABAJO y ACTIVIDADES_ESTUDIO. En cuanto al NIVEL_EDUCATIVO el Ciclo II es aquel que más se repite. En la variable HIJOS MENORES la categoría SI es la más frecuente lo que indica que al parecer la mayoría de personas tienen o tenían un hijo menor de 18 años al momento de la condena.

Entendimiento de la variable objetivo reincidencia

Se puede observar en la Tabla 9 que muestra variable objetivo reincidente tiene 2 valores únicos de los cuales el "NO" es el más frecuente apareciendo 441.734 veces en la base de datos:

Conteo	**614916**
único	2
top	NO
frecuencia	441734

Tabla 9. Variable objetivo.

Se puede observar la distribución de la variable reincidente en la Tabla 10. La categoría dominante como ya se había mencionado es “NO” representando el 71.84% del total de registros.

Reincidencia	Conteo	Porcentaje
NO	441734	71.84
SI	173182	28.16

Tabla 10. Variable reincidencia porcentajes

Verificación de la calidad de los datos

Con la finalidad de trabajar con datos de calidad, se hace una revisión en toda la base datos en búsqueda de valores faltantes o nulos. En la Tabla 11 vemos las inconsistencias encontradas.

NOMBRE DE LA VARIABLE	RECUENTO DE INCONSISTENCIAS	PORCENTAJE RESPECTO EL TOTAL DE LA VARIABLE
PAIS_INTERNO	17930	2,92 %
TIPO_SALIDA	864	0,14 %
ESTADO_INGRESO	52	0,01 %
SISTEMA_PENAL_ACUSATORIO	40	0,01 %
ANO_NACIMIENTO	8	0,00 %
TENTATIVA	0	0,00 %
AGRAVADO	0	0,00 %
CALIFICADO	0	0,00 %
CLASIFICACION JURIDICA DELITO	0	0,00 %
CLASIFICACION LUHMAN DELITO	0	0,00 %
FECHA_INGRESO	0	0,00 %
FECHA_SALIDA	0	0,00 %
CONDENA CUMPLIDA MESES	0	0,00 %
CONDENA INICIAL CUMPLIDA	0	0,00 %
DELITO	0	0,00 %
FECHA_CAPTURA	0	0,00 %
ESTADO	0	0,00 %
REGIONAL	0	0,00 %
SITUACION_JURIDICA	0	0,00 %
ESTABLECIMIENTO	0	0,00 %
GENERO	0	0,00 %
DEPARTAMENTO	0	0,00 %
CIUDAD	0	0,00 %

MESES_CONDENA	0	0,00 %
REINCIDENTE	0	0,00 %
ACTIVIDADES_TRABAJO	0	0,00 %
ACTIVIDADES_ESTUDIO	0	0,00 %
ACTIVIDADES_ENSENANZA	0	0,00 %
NIVEL_EDUCATIVO	0	0,00 %
HIJOS_MENORES	0	0,00 %

Tabla 11. Inconsistencias en variables

Se observa que la variable con mayor número de inconsistencias es PAIS_INTERNO, aunque representando solo el 3% aproximadamente de todos los registros contenidos en esta variable. A las variables que presentaron valores nulos se les deben aplicar técnicas de tratamientos de datos en la fase siguiente de la investigación.

Análisis grafico univariado

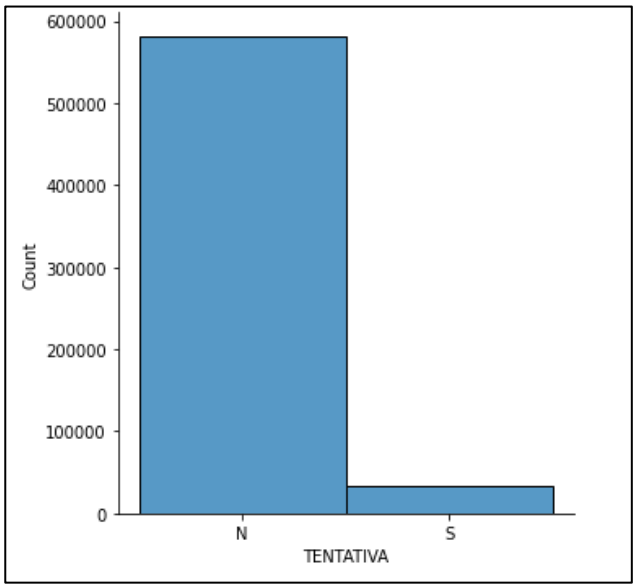

Figura 4. Variable tentativa

En la Figura 4 se evidencia que la gran mayoría de personas no tiene Tentativa asociada al delito cometido. El análisis bivariado permitirá obvservar la relación de esta variable con la reincidencia.

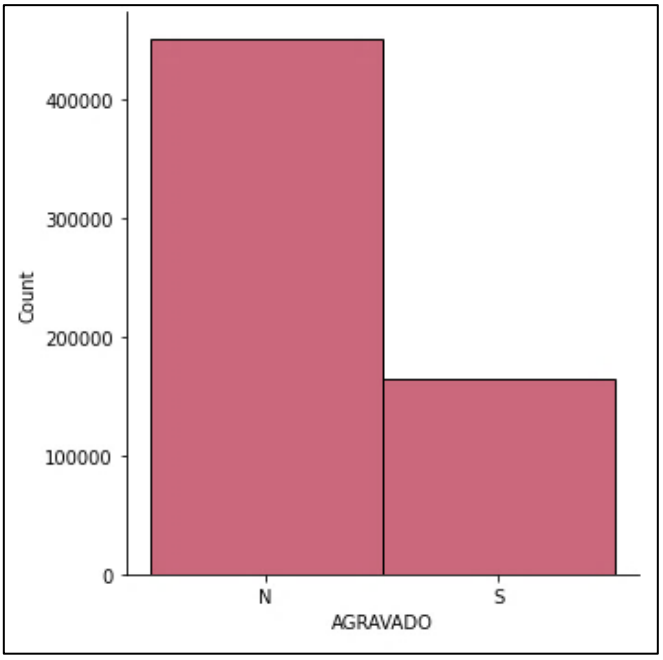

Figura 5. Variable agravado

En la Figura 5Figura 4 se evidencia que son más las personas que no tienen condición de Agravado asociado a su delito, esta variable denota un aumento considerable de las personas en condición de Si (S) respecto la variable Tentativa vista anteriormente. El análisis bivariado permitirá obvservar la relación de esta variable con la reincidencia.

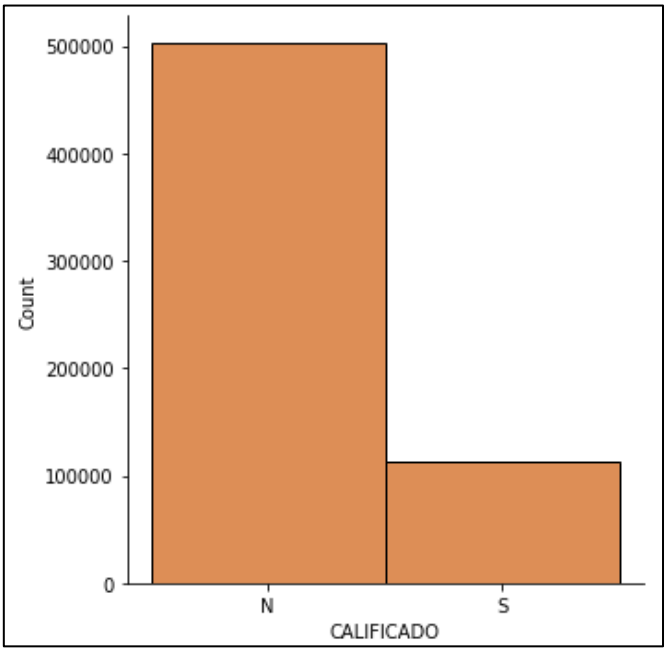

Figura 6. Variable calificado

En la Figura 6 se evidencia que la mayoría de personas no tiene condición de Calificado asociada al delito cometido. El análisis bivariado permitirá obvservar la relación de esta variable con la reincidencia.

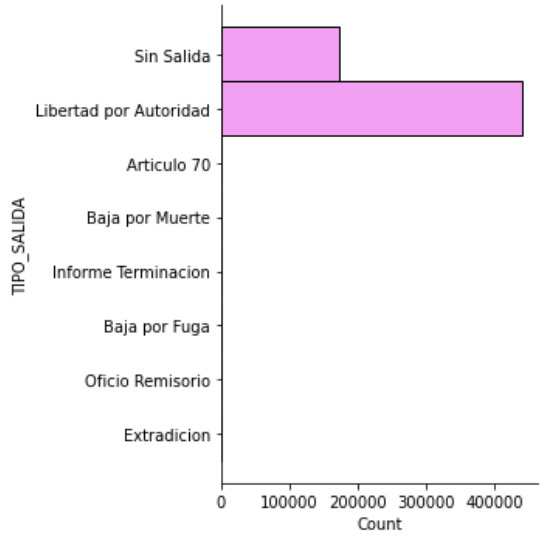

Figura 7. Variable tipo de salida

En la Figura 7 observamos que hay dos categorías predominantes en la variable como lo son “Sin Salida” y “Libertad por autoridad”. El análisis bivariado permitirá obvservar la relación de esta variable con la reincidencia.

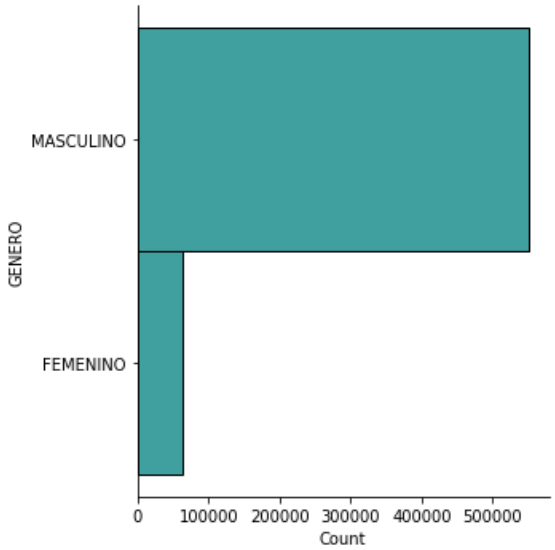

Figura 8. Variable genero

En la Figura 8Figura 7 observamos que la gran mayoría de personas de la base de datos pertenecen a la categoria “Masculino” .El análisis bivariado permitirá obvservar la relación de esta variable con la reincidencia.

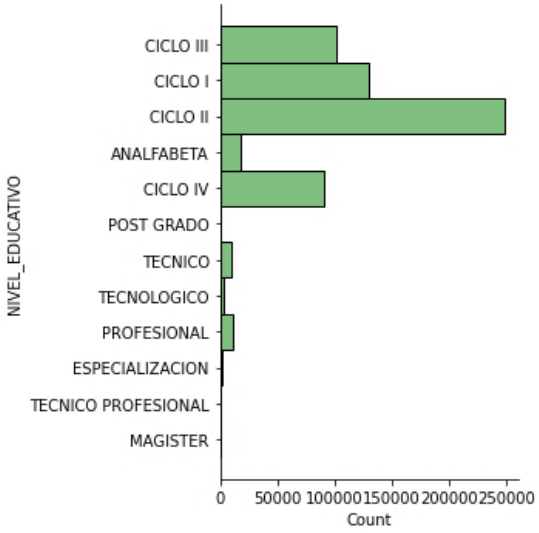

Figura 9. Variabe nivel educativo

En laFigura 7 Figura 9 se aprecian que las categorías del nivel educativo de las personas condenadas, el de mayor frecuencia es el “CICLO II”. El análisis bivariado permitirá obvservar la relación de esta variable con la reincidencia.

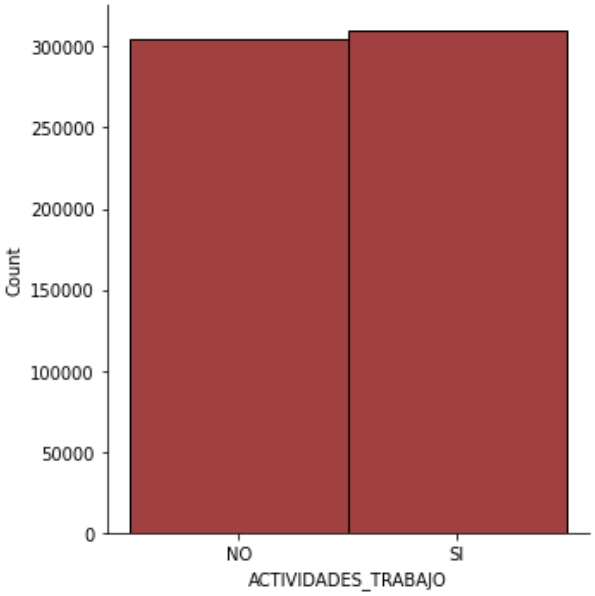

Figura 10. Variable actividades de trabajo

En la Figura 10 Figura 7se evidencia que las dos categorías que componen la variable Actividades de Trabajo tienen porcentajes de participación muy cercanos. El análisis bivariado permitirá obvservar la relación de esta variable con la reincidencia.

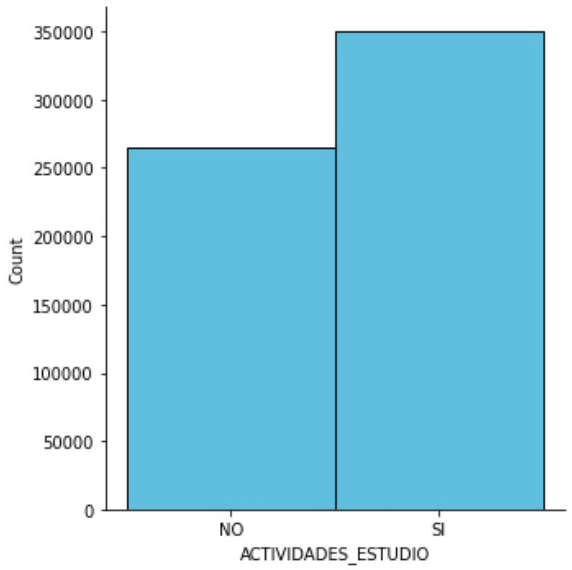

Figura 11. Variable actividades de estudio

En la Figura 11 Figura 7podemos apreciar que son mas las personas que Si realizan actividades de esdtudio adentro de las penitenciarias. El análisis bivariado permitirá obvservar la relación de esta variable con la reincidencia.

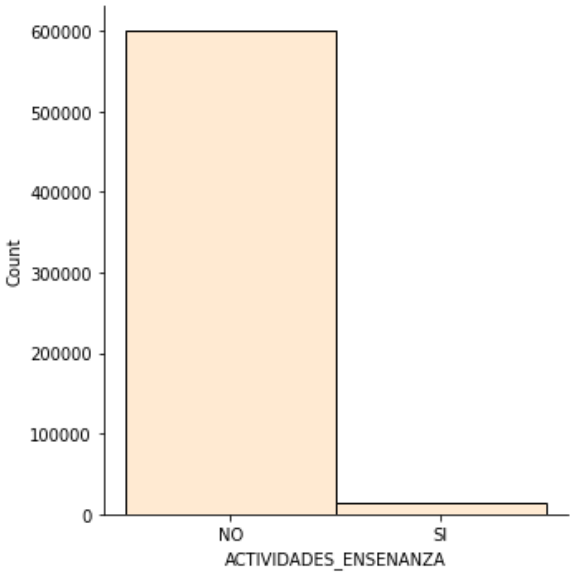

Figura 12. Actividades de enseñanza

En cuanto la variable Actividades de enseñanza se aprecia en la Figura 12 que la gran mayoría de personas no las realiza dentro de los centros de reclusion. El análisis bivariado permitirá obvservar la relación de esta variable con la reincidencia.

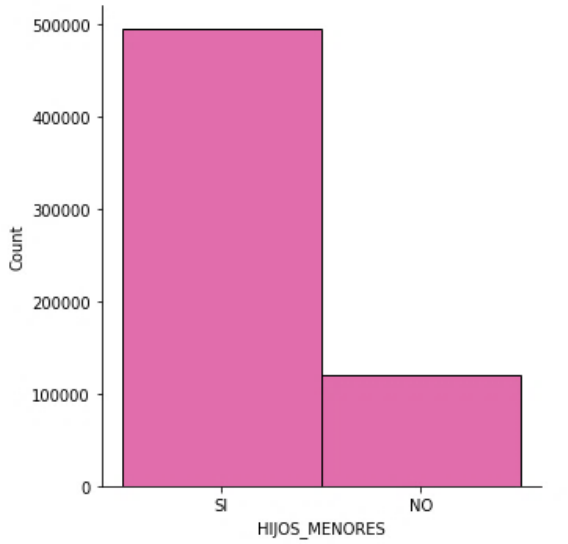

Figura 13. Variable hijos menores

Las personas que Si tienen hijos menores son la mayoría en la base de datos tal como se observa en la Figura 13. El análisis bivariado permitirá obvservar la relación de esta variable con la reincidencia.

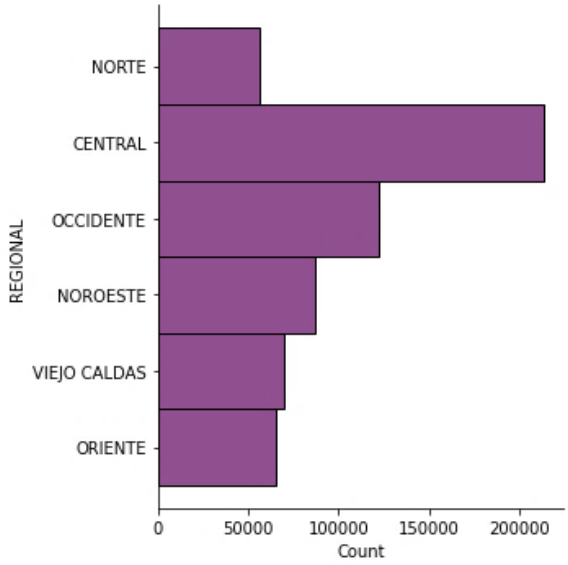

Figura 14. Variable regional

Como se evidencia en la Figura 14 la regional central es aquella donde más personas están cumpliendo su condena. El análisis bivariado permitirá obvservar la relación de esta variable con la reincidencia.

Análisis grafico bivariado

Mediante el contraste grafico de una variable contra la variable Reincidente, se analizan las posibles relaciones y se observan los comportamientos.

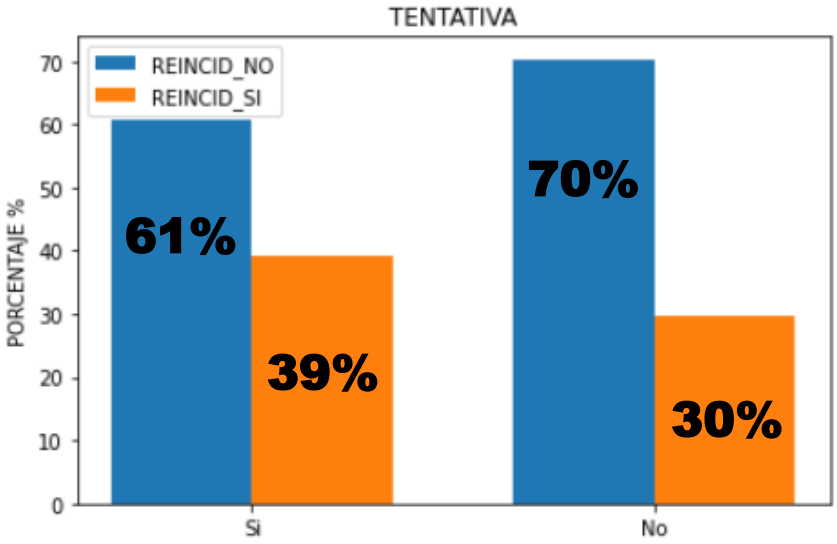

Figura 15. Tentativa y Reincidencia

En la Figura 15 se observa que las personas reincidentes con tentativa representan el 39% del grupo mientras las reincidentes sin tentativa el 30%; si bien no muestra una relación contundente es un comportamiento que se debe analizar con detalle en la fase de modelado.

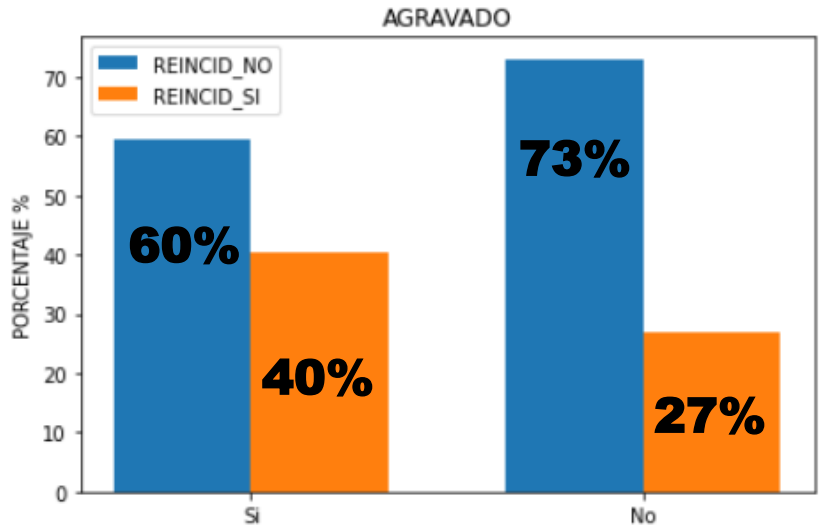

Figura 16. Agravado y Reincidencia

En la Figura 16 se observa mayor porcentaje de personas reincidentes en aquellas que SI tienen la connotación de agravado; esta posible relación se debe validar en la fase de modelado.

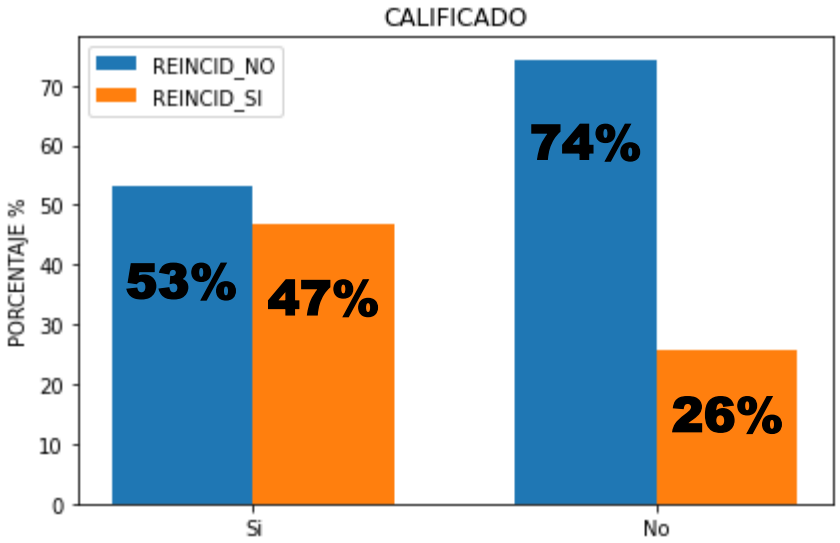

Figura 17. Calificado y Reincidencia

Observamos en la Figura 17 que la reincidencia en personas con delitos calificados se da en relación del 47% mientras que la reincidencia en las personas sin delito calificado corresponde al 26%, lo que muestra una posible incidencia fuerte de la variable calificado con la reincidencia.

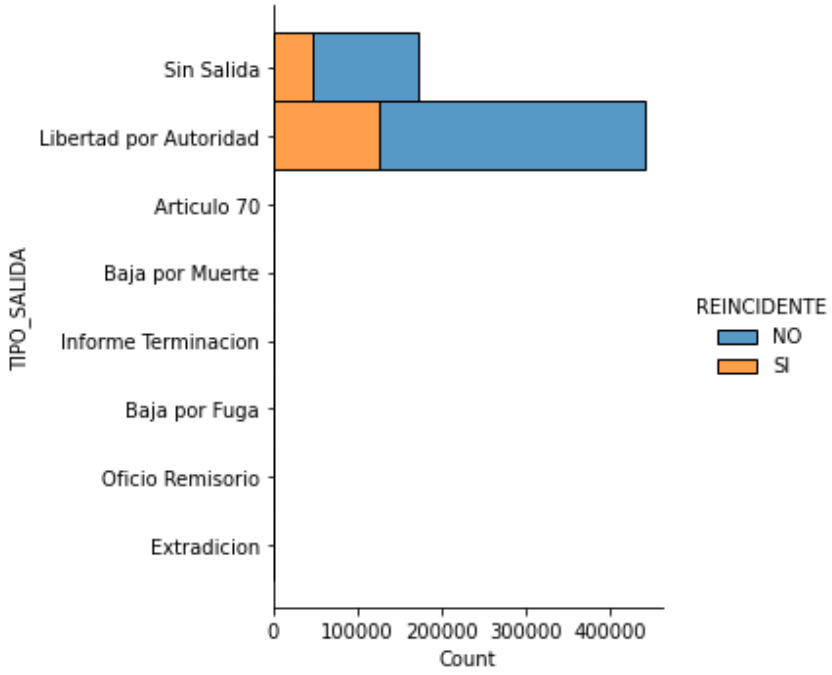

Figura 18. Tipo de Salida y Reincidencia

Observamos en la Figura 18 que el tipo de salida se concentra en solo 2 categorías por lo que sería útil hacer una agrupación en la transformación de variables.

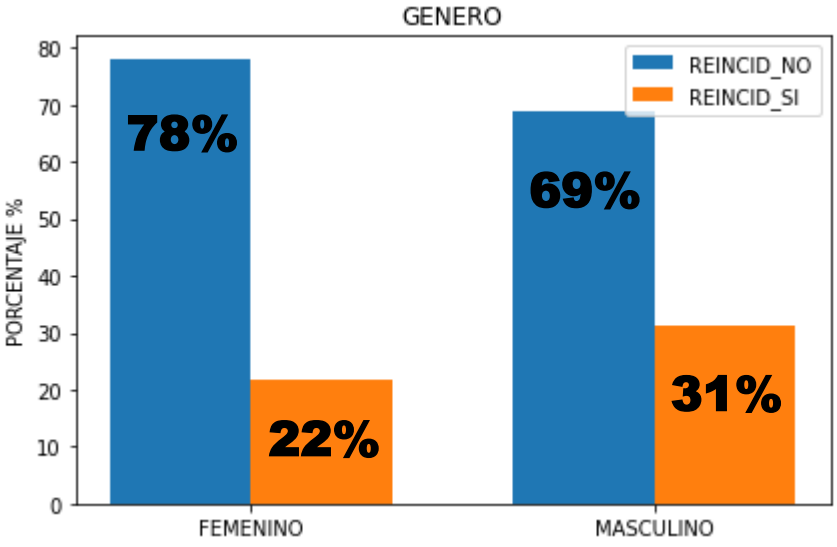

Figura 19. Sexo y Reincidencia

Se aprecia en la Figura 19 que en la variable genero se observa una leve incidencia del género masculino en la reincidencia. Las fases posteriores podrían ayudar a confirmarla.

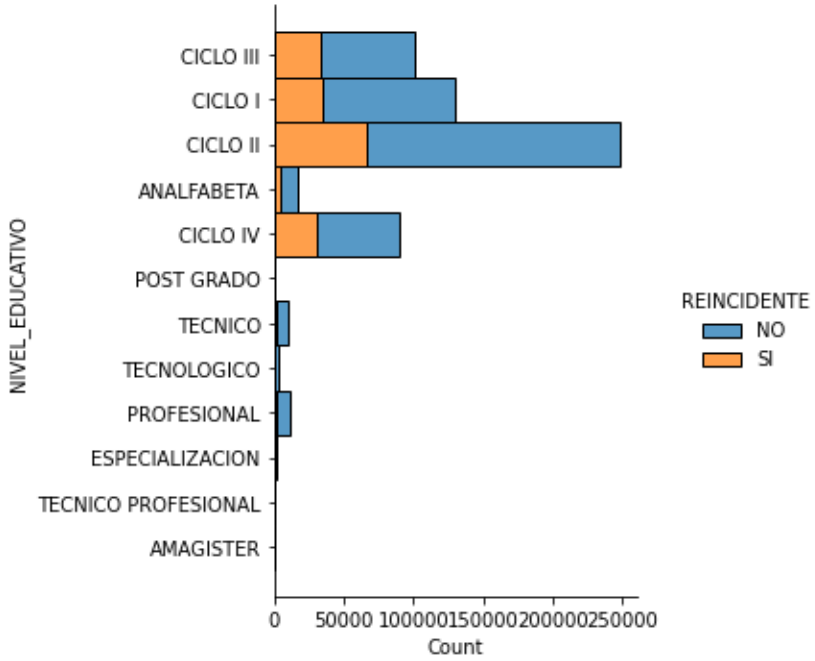

Figura 20. Nivel Educativo y Reincidencia

No se observa en la Figura 20 una relación clara entre las variables analizadas y la reincidencia. Transformaciones que se puedan hacer en la variable tales como agrupaciones podrían aportar más información y posibles relaciones.

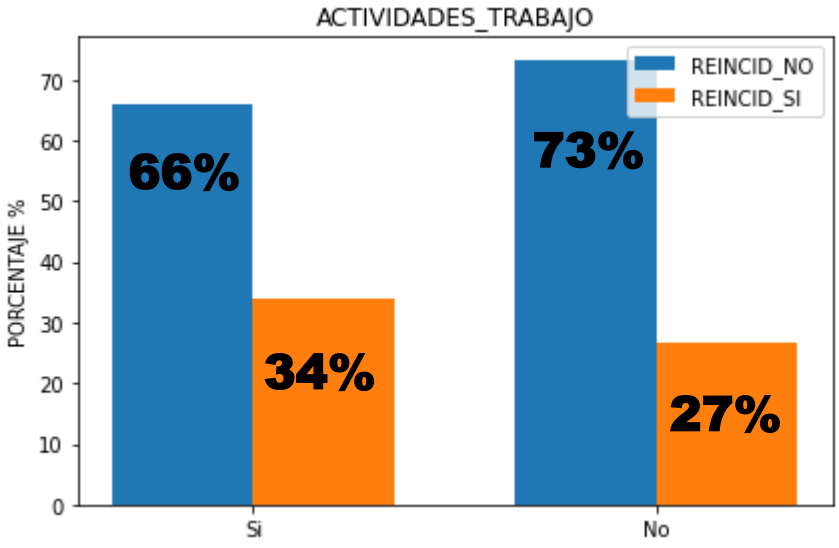

Figura 21. Actividades de Trabajo y Reincidencia

En la Figura 21 no observamos una relación de la variable actividades de trabajo, con la variable reincidencia, dado que el porcentaje varía tan poco en cada categoría, es necesario confirmar o refutar con otros análisis.

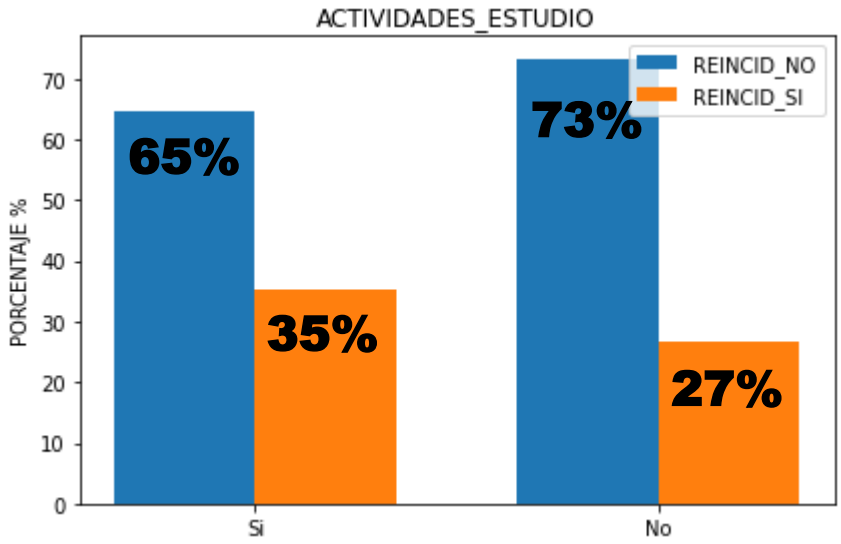

Figura 22. Actividades de Estudio y Reincidencia

En la Figura 22 la variable actividades de estudio no muestra una relación clara con la reincidencia, se debe profundizar posteriormente el análisis.

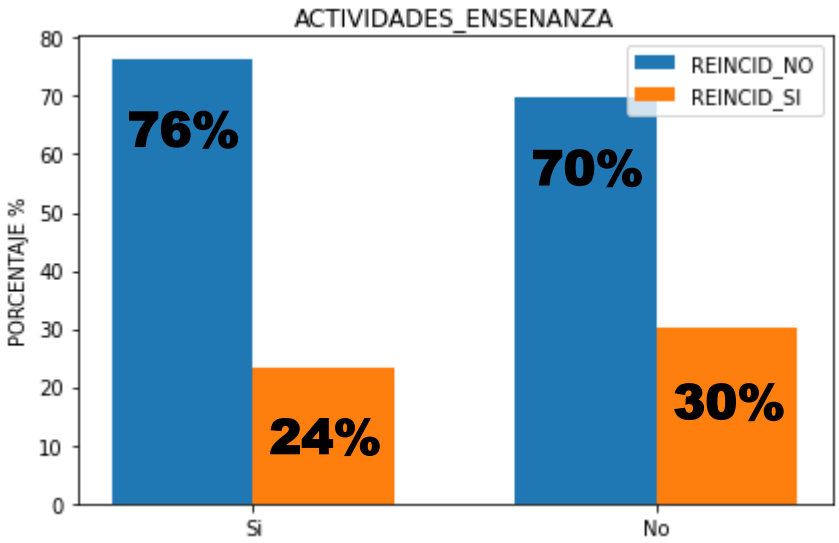

Figura 23. Actividades Enseñanza y Reincidencia

Según la Figura 23 pareciera que las personas que no realizan actividades de enseñanza tienden a reincidir en mayor medida. Esta relación se observará en el modelado para su confirmación o descarte.

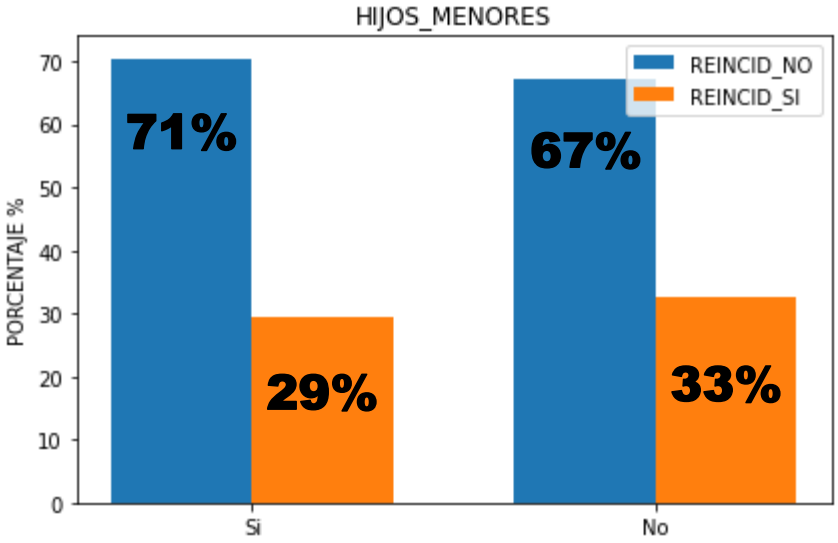

Figura 24. Hijos Menores y Reincidencia

No se observa en la Figura 24 una relación evidente de la variable hijos menores, con la variable reincidencia que permita suponer alguna hipótesis.

```
PAIS_INTERNO           REINCIDENTE
REPUBLICA DE COLOMBIA  NO            0.711260
                       SI            0.282682
VENEZUELA              NO            0.002886
ECUADOR                NO            0.000620
MEXICO                 NO            0.000444
                                       ...
CHINA                  SI            0.000002
HUNGRIA                NO            0.000002
PANAMA                 SI            0.000002
NUEVA ZELANDA          NO            0.000002
HONG KONG              NO            0.000002
Length: 111, dtype: float64
```

Figura 25. País y Reincidencia

Dado que País_Interno es una variable con gran cantidad de categorías, no se puede realizar una gráfica previa transformación de la variable. Además, dada la poca representatividad de la mayoría de categorías de la variable, es candidata a salir de la fase de modelado.

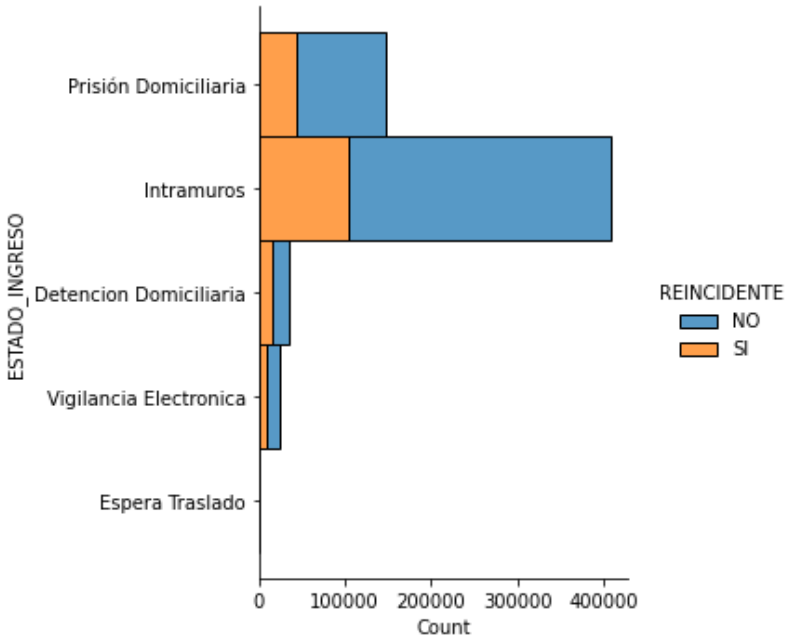

Figura 26. Estado Ingreso y Reincidencia

En la Figura 26 se aprecia que las personas cumpliendo su condena mediante la modalidad de detención domiciliaria y prisión domiciliaria parecieran reincidir en mayor medida que el resto de modalidades, esta posible relación debe ser confirmada en las etapas posteriores.

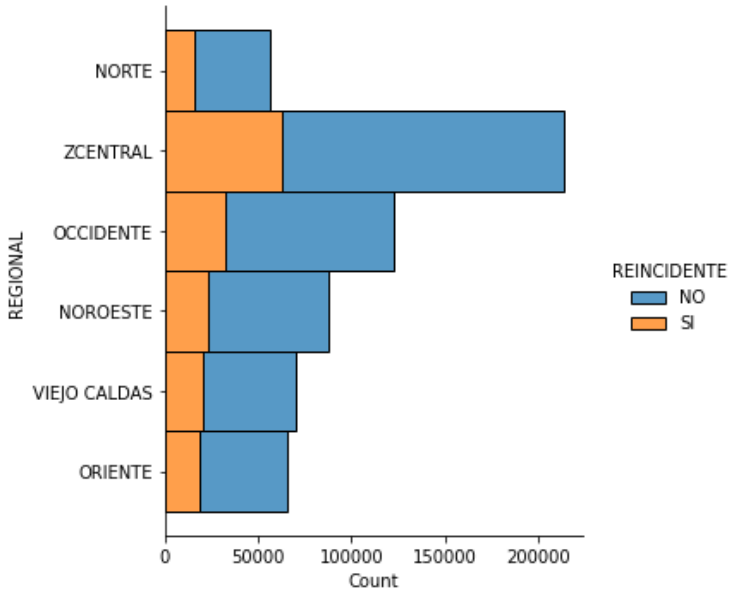

Figura 27. Regional y Reincidencia

Podemos observar un comportamiento relativamente uniforme en todas las regionales contrastadas con la reincidencia en la Figura 27. Se deben plantear agrupaciones o hipótesis que permita obtener más información de la variable.

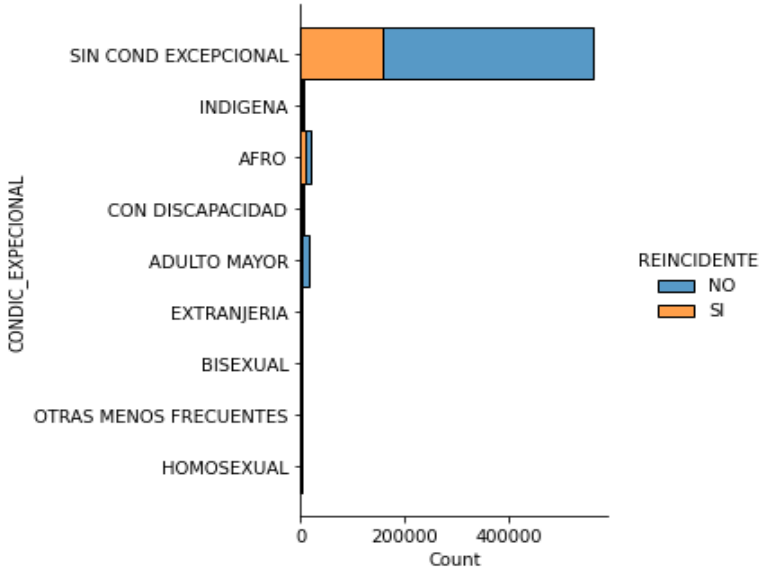

Figura 28. Condición Excepcional y Reincidencia

Dada la poca representatividad de la mayoría de categorías de la variable condición excepcional como se ve en la Figura 28; podría ser candidata a salir del modelado o realizarle discretización en la fase de preparación de los datos.

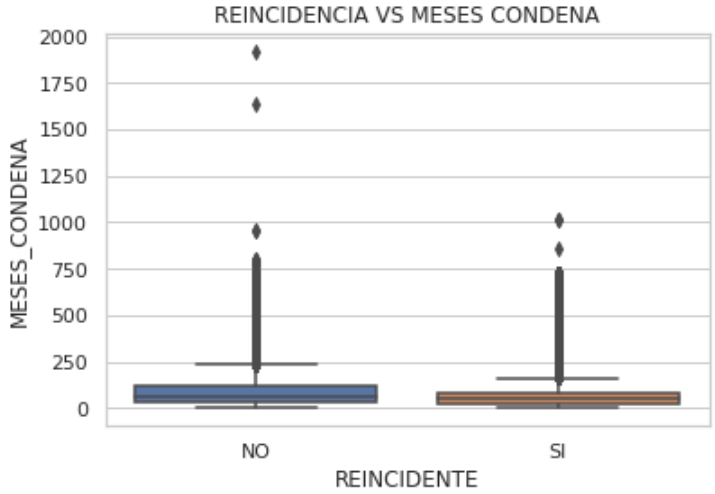

Figura 29. Meses Condena y Reincidencia

En la Figura 29 observamos valores atípicos a partir de los 150 meses de condena por lo que en la preparación de los datos se debe corregir.

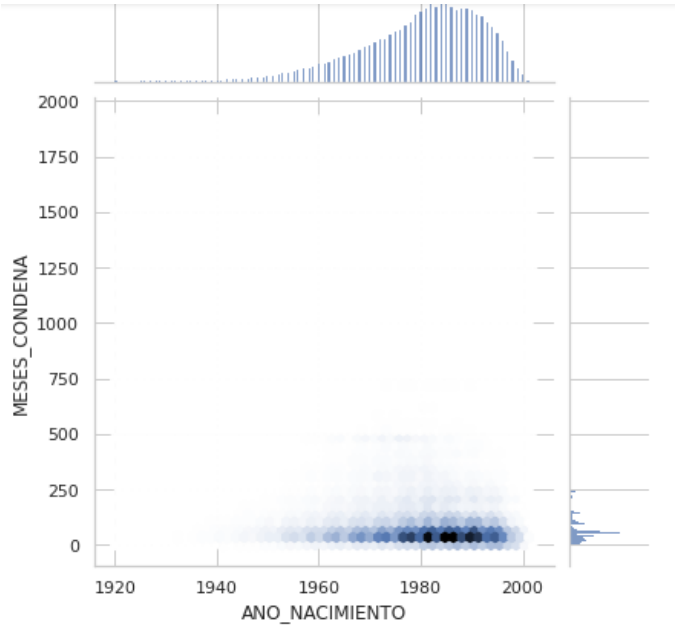

Figura 30. Meses Condena y Año Nacimiento

Se puede observar Figura 30 que hay mayor concentración de personas condenadas nacidas entre el año 1975 y 1995 aproximadamente, que además están cumpliendo condenas en mayor medida entre los 10 y 75 meses. Lo anterior podría relacionar la edad con el tiempo de condenas cumplidas, las cuales podrian pertenecer a grupos de delitos específicos. Estas conjeturas se revisarán en fases posteriores.

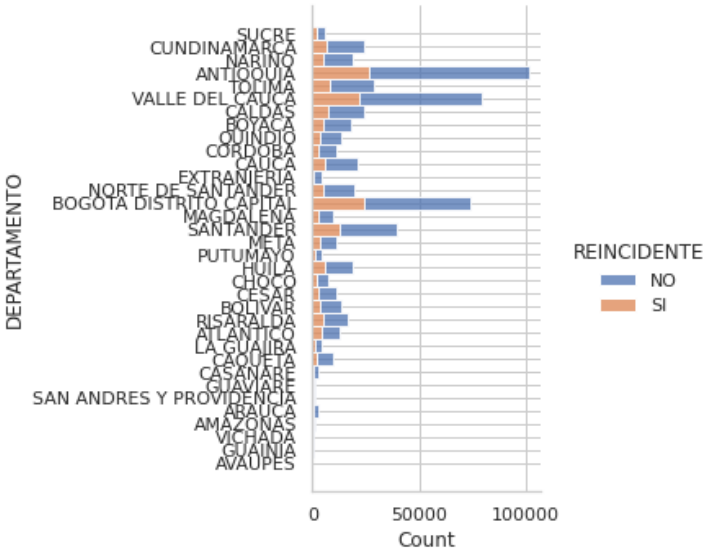

Figura 31. Departamento y Reincidencia

No se percibe un comportamiento claro en la Figura 31 del departamento vs la reincidencia. se podría hacer alguna discretización en la preparación de los datos que permita obtener más información.

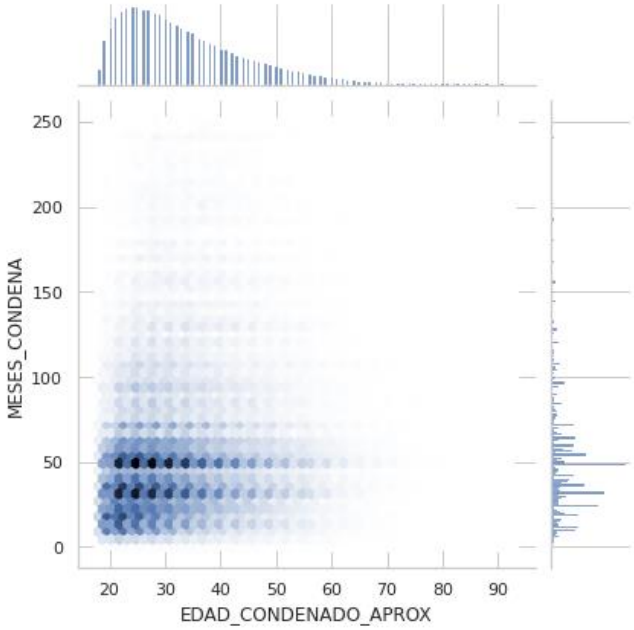

Figura 32. Meses Condena y Edad Condenado

Se observa en la Figura 32 que las personas con edades menores cumplen condenas más cortas. Ello podría explicarlo el tipo de delitos que se cometen a ciertas edades de la carrera criminal.

Análisis grafico multivariado

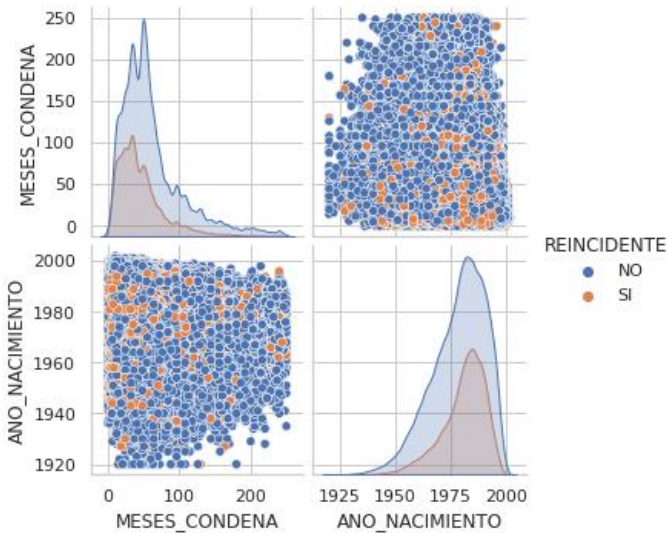

Figura 33. Año Nacimiento, Meses Condena y Reincidencia

Como era de esperarse en la Figura 33 vemos que la variable año nacimiento evidencia una distribución similar tanto en reincidentes como en no reincidente, por lo que se requería construir una variable llamada edad para sacar mejores conclusiones. Es interesante el comportamiento bivariado de la gráfica "meses_condena" vs reincidente, las personas reincidentes presentan meses de condena menores.

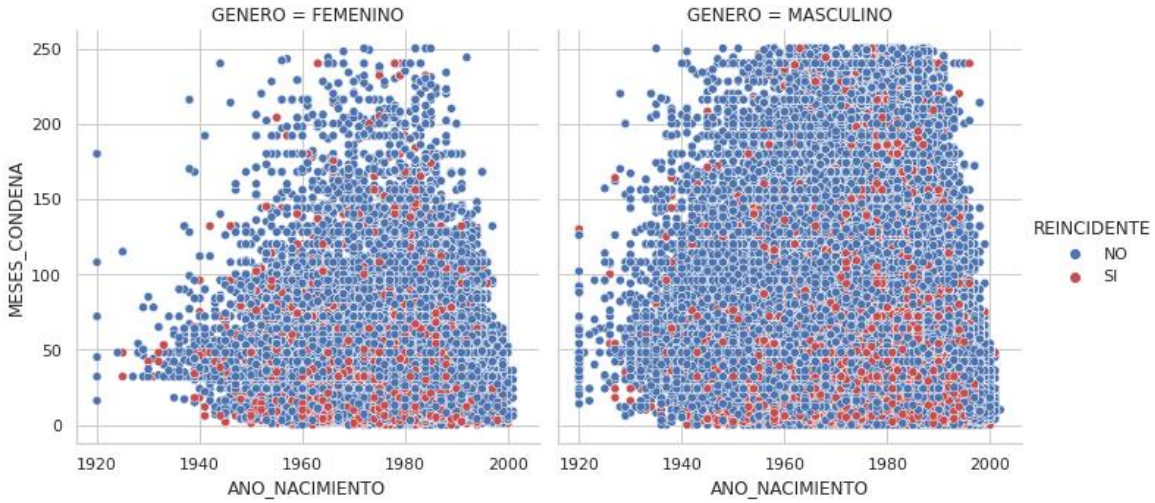

Figura 34. Densidad Año Nacimiento, Meses Condena y Reincidente

La Figura 34 muestra que el género femenino cumple condenas más cortas y con edades menores que el género masculino. una hipótesis es que el género femenino comete delitos de menor gravedad que a su vez conllevan condenas más cortas. La reincidencia en ambos grupos se comporta de manera similar y no evidencia relación clara con la combinación de variables "ano_nacimiento", "meses_condena", y "genero".

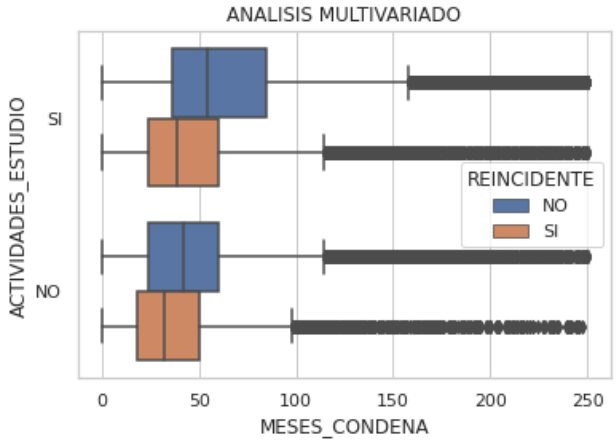

Figura 35. Actividades de Estudio, Meses Condena y Reincidencia

Contrastadas en la Figura 35 las variables meses_condena, actividades_estudio y reincidente, se observa que las personas reincidentes en el grupo que si realiza actividades de estudio está cumpliendo condenas más cortas que los no reincidentes de su mismo grupo.

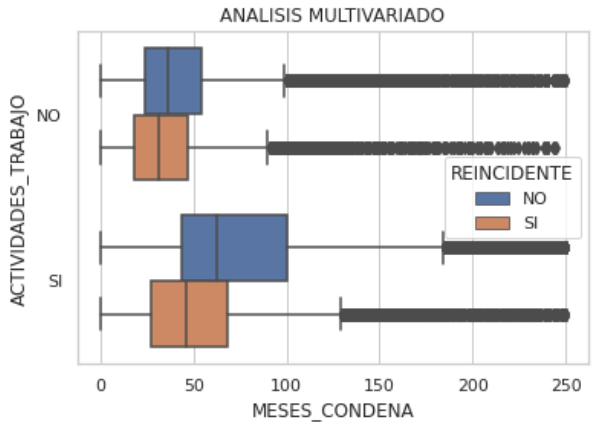

Figura 36. Actividades de Trabajo, Meses Condena y Reincidencia

Graficadas en la Figura 36 las variables meses_condena, actividades_trabajo y reincidente, se observa comportamiento similar de la reincidencia en ambos grupos.

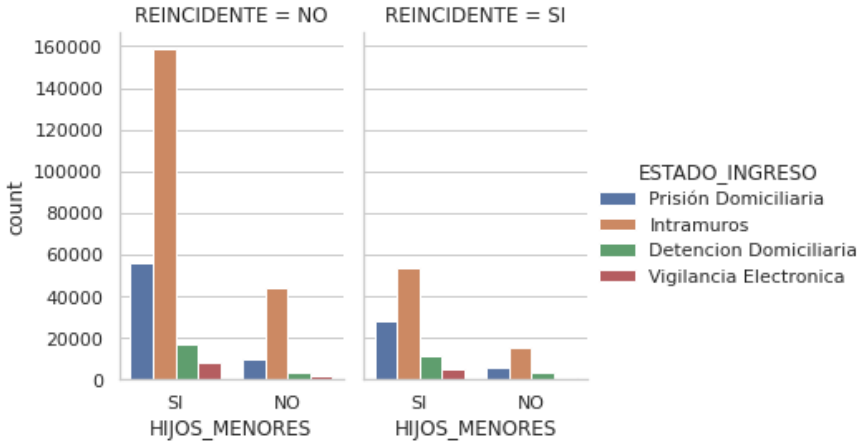

Figura 37. Hijos Menores, Estado Ingreso y Reincidencia

En la

Figura 37 el comportamiento de las variables es relativamente consistente en ambos grupos. No muestran una clara incidencia en la reincidencia.

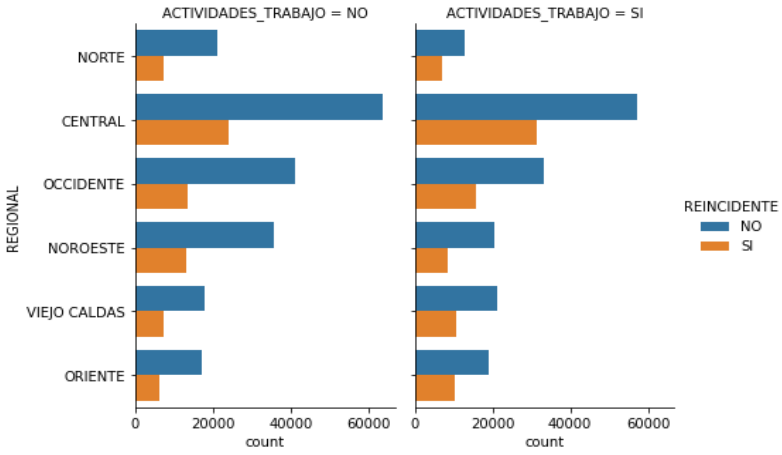

Figura 38. Actividades de Trabajo, Regional y Reincidente

Analizando la Figura 38 se observa que la reincidencia, es levemente mayor en los grupos de personas que realizan actividades de trabajo en la mayoría de las regionales, principalmente en la regional central. Parece que la regional y si realiza actividades de trabajo influye en la reincidencia. Esta observación tendrá que ampliarce en análisis posteriores.

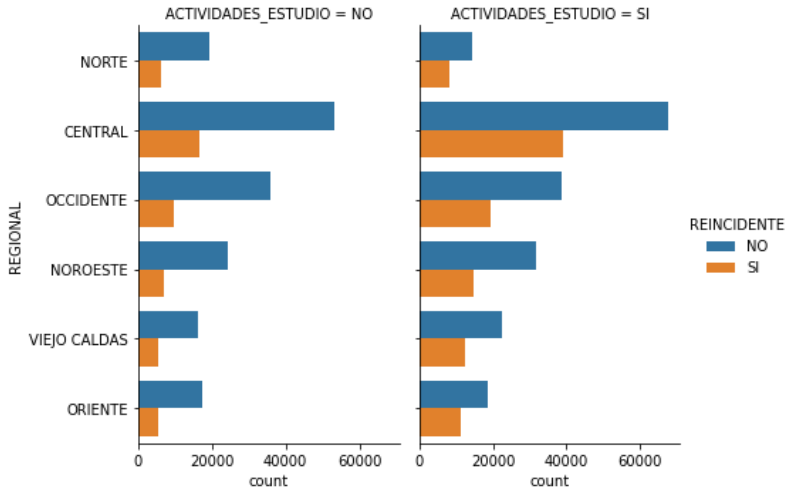

Figura 39. Actividades de Estudio, Regional y Reincidente

Vemos en la Figura 39 que la reincidencia es mayor en los grupos de personas que realizan actividades de estudio en la mayoría de las regionales. Parece que la regional y si realiza actividades de estudio influye en la reincidencia. Esta observación tendrá que ampliarce en análisis posteriores.

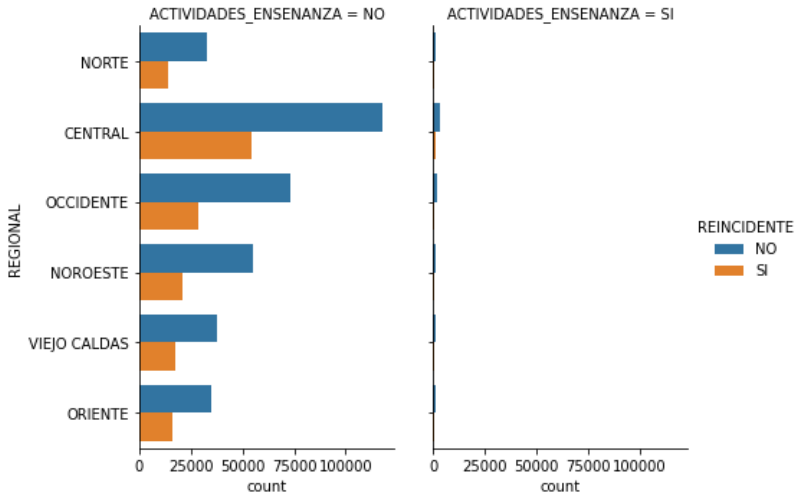

Figura 40. Regional, Actividades de Enseñanza y Reincidencia

En la Figura 40 la variable enseñanza tiene muy pocos valores en la categoría si, por lo que no es posible observar comportamiento grafico comparado.

7.2 Fase O2: Implementacion de modelo clasificatorio que permita explicar la reincidencia delictiva de las personas que hayan estado bajo la vigilancia del Instituto Nacional Penitenciario y Carcelario INPEC

7.2.1 Preparación de los datos

Limpieza variable SITUACION JURIDICA

En la Tabla 12 se puede observar que la variable está constituida por las categorías Altas y Bajas. Altas son aquellas personas que tienen una condena vigente, mientras que Bajas las personas que ya salieron del centro de reclusión.

VARIABLE ESTADO	CONTEO
ALTAS	171945
BAJAS	442971

Tabla 12. Variable estado

Se eliminaron de la base de datos las personas que tienen una condena vigente (ALTAS) con el fin de trabajar con datos de personas que hayan atravesado todo el ciclo penitenciario (BAJAS).

VARIABLE ESTADO	CONTEO
BAJAS	442971

Tabla 13. Variable estado depurada

Limpieza variable MESES CONDENA

Se realiza limpieza de esta variable basándose en el análisis gráfico y con el fin de eliminar outliers dado que a la fecha de obtención de la base de datos la condena máxima en Colombia era de 720 meses y tal como se aprecia en la Tabla 14 hay condenas de 1920 meses. Unido a ello según los gráficos de la variable Meses Condena, las condenas superiores a 150 meses figuran como datos atípicos por lo que se trabajará con condenas menores o iguales a 150 meses.

Media	**93,50**
Desviación	103,72
Valor Mínimo	0
Cuartil 25%	33
Cuartil 50%	56
Cuartil 75%	108
Valor Máximo	1920

Tabla 14. Variable meses condena

En la Tabla 15 vemos como quedó la variable luego de la limpieza realizada, obteniendo condenas máximas de 150 meses y una desviación estándar mucho menor.

Media	**49,93**
Desviación	31,38
Valor Mínimo	0
Cuartil 25%	26
Cuartil 50%	46
Cuartil 75%	64
Valor Máximo	150

Tabla 15. Meses condena depurada

- **Formateo de los datos**

Variables binarias

Dado que la fase de modelado requiere unas características técnicas específicas respecto las variables y sus datos, se hace necesario realizar transformaciones con técnicas cómo One Hot Encoding, con el fin de que las variables sean binarizadas. Esto permitirá que los modelos trabajen de mejor manera y los resultados sean más confiables. En la Tabla 16 se muestra el formateo realizado a las variables binarias de la base de datos.

NOMBRE DE LA VARIABLE	CATEGORIAS ORIGINALES	NUEVO NOMBRE DE LA VARIABLE	CATEGORIAS FORMATEADAS
REINCIDENTE	NO - SI	REINCI_1SI	0 - 1
TENTATIVA	N - S	TENTATIVA_1SI	0 - 1
AGRAVADO	N - S	AGRAVADO_1SI	0 – 1
CALIFICADO	N - S	CALIFICADO_1SI	0 – 1
GENERO	FEMENINO - MASCULINO	SEXO_1MASCUL	0 – 1
ACTIVIDADES_TRABAJO	NO - SI	ACT_TRABAJO_1SI	0 – 1
ACTIVIDADES_ESTUDIO	NO - SI	ACT_ESTUD_1SI	0 – 1
ACTIVIDADES_ENSENANZA	NO - SI	ACT_ENSENA_1SI	0 – 1
HIJOS_MENORES	NO - SI	HIJOS_MENOR_1SI	0 – 1

Tabla 16. Variables binarias

Variables de múltiples categorías

Como se pudo observar en la tabla anterior, la base de datos cuenta con unas variables de tipo binario que están compuestas únicamente por dos categorías, pero también existen variables con múltiples categorías que fueron transformadas y se muestran a continuación.

Variable TIPO DE SALIDA

La variable tipo de salida está compuesta por las diferentes causas por las que una persona puede salir del sistema penitenciario. Inicialmente la variable tenia las categorías: Baja por Fuga, Baja por Muerte, Extradición, Informe Terminación, Libertad por Autoridad, Oficio Remisorio y Sin Salida; para la conveniencia de la investigación se agruparon ciertas categorías, logrando con esto además de simplificar el contenido de la variable, observar el comportamiento de la categoría Baja por Fuga frente a un grupo menor de opciones.

TIPODESALIDA_1FUGA	
CATEGORIAS ORIGINALES	**CATEGORIAS FORMATEADAS**
Baja por Muerte	0
Extradición	
Baja por Fuga	1
Informe Terminación	2
Libertad por Autoridad	
Oficio Remisorio	
Sin Salida	

Tabla 17. Variable Tipo de Salida Formateada

En la Tabla 17 observamos que la variable Tipo de Salida fue reducida a 3 únicas variables para el interés de la investigación.

Variable NIVEL EDUCATIVO

Esta variable tiene doce categorías que se observan en la Tabla 18. Con el fin de formatearla se recurrió a la función get_dummies[7] de la librería de Python llamada Pandas. Con esta función se crea una columna por cada categoría de la variable asignándole el numero 1 a la categoría que corresponda para cada registro tal como se aprecia en el ejemplo descrito en la Tabla 19.

CATEGORIAS DE LA VARIABLE
ANALFABETA
CICLO I
CICLO II
CICLO III
CICLO IV
TECNICO
TECNICO PROFESIONAL
TECNOLOGICO
PROFESIONAL
ESPECIALIZACION
POST GRADO
MAGISTER

Tabla 18. Variable nivel educativo

[7] https://pandas.pydata.org/docs/reference/api/pandas.get_dummies.html

ID INTERNO ENCRIPTADO	NIVEL EDUCATIVO											
	ANALFABETA	CICLO I	CICLO II	CICLO III	CICLO IV	TECNICO	TECNICO	TECNOLOGICO	PROFESIONAL	ESPECIALIZACION	POST GRADO	MAGISTER
000001E4D7D5D64A8FD3AC2991B45947B823572C	0	0	0	1	0	0	0	0	0	0	0	0
000122B138FFDB01ACF5773EF7C67C566F4DE47E	0	0	1	0	0	0	0	0	0	0	0	0

Tabla 19. Variable Nivel Educativo Formateada

El formateo realizado con la función get_dumies se aplicó de la misma manera a las variables REGIONAL, CONDIC_EXPECIONAL, CLASIFICACION JURIDICA DELITO Y CLASIFICACION LUHMAN DELITO.

Variable PAIS INTERNO

Dado que la variable PAIS_INTERNO tiene demasiadas categorías y la gran mayoría de registros se ubican en la categoría REPUBLICA DE COLOMBIA, se formateo en una variable binaria agrupando todos los registros diferentes a REPUBLICA DE COLOMBIA en una misma categoría.

PAIS_INTERNO_1COL	
CATEGORIAS ORIGINALES	**CATEGORIAS FORMATEADAS**
Categorías distintas a Republica de Colombia	0
República de Colombia	1

Tabla 20. Variable País Interno

Variable ESTADO INGRESO

La base de datos tiene una variable llamada ESTADO_INGRESO, esta variable define la modalidad en la cual la persona está cumpliendo su condena, para lo cual cuenta con las categorías: Prisión Domiciliaria, Intramuros, Detención Domiciliaria, Vigilancia Electrónica y Espera Traslado. Según las necesidades y conveniencia para la investigación se le hizo a la variable el formateo que se muestra a continuación.

ESTADO_INGRESO_1INTRAMUROS	
CATEGORIAS ORIGINALES	**CATEGORIAS FORMATEADAS**
Prisión Domiciliaria	0
Detención Domiciliaria	
Vigilancia Electrónica	
Espera Traslado	
Intramuros	1

Tabla 21. Variable Estado Ingreso

En la Tabla 21 observamos que se crearon únicamente dos categorías que pueden ser catalogadas como aquellas personas que cumplen su condena en modalidad de extramuros, por fuera de un centro de reclusión, tal como lo son las categorías Prisión Domiciliaria, Detención Domiciliaria, Vigilancia Electrónica y Espera Traslado; y otra categoría denominada intramuros que son todas aquellas personas que cumplen la condena dentro de un centro penitenciario. Esto con el fin de observar el comportamiento de la reincidencia en ambos escenarios.

Variable MESES CONDENA

Debido a que MESES CONDENA es una variable continua se decidió agruparla en 3 rangos, con el fin de observar el comportamiento de la reincidencia en grupos de edades que permitan entender mejor su comportamiento.

MESES_CONDENA_CAT	
RANGO CONDENA	**CATEGORIA**
Menor igual a 36 meses	0
Mayor a 36 y Menor a 60 meses	1
Mayor a 60 meses	2

Tabla 22. Variable Rango Condena.

- **Construcción de Datos**

Hipótesis 1 Condena Inicial Cumplida

Se plantea la hipótesis de que aquellas personas que no cumplen su condena inicialmente proferida por un Juez, tienden a reincidir más; por cuanto perciben que el sistema de justicia es flexible. Por tanto, se agrupan las personas que cumplieron su condena inicial (categoría “SI”) y aquellas con condena vigente (categoría “En prisión”) en un solo grupo. La Tabla 23 se observa la nueva categorización.

H1_CondenaInicial_Cumpl_1NO	
CATEGORIAS ORIGINALES	**CATEGORIAS FORMATEADAS**
SI	0
En prisión	
NO	1

Tabla 23. Variable Condena Inicial Cumplida

Hipótesis 2: Tiempo de Condena

Se plantea que las condenas cortas influyen en la reincidencia de las personas. Aquellas que cumplen condenas muy cortas tienen a delinquir nuevamente. Por tanto, se redujo la variable a 2 únicas categorías como se muestran en la siguiente tabla.

RANGO CONDENA	CONVENCIÓN
Menor igual a 36 meses	0
Mayor a 36 meses	1

Tabla 24. Hipótesis 2 Tiempo Condena

Hipótesis 3: Condición Excepción

Aquellas personas que pertenecen a una minoría reinciden menos que aquellas que no tienen ninguna condición especial. Por eso se agruparon todas las condiciones excepcionales en una sola categoría quedando como resultado una variable binaria.

RANGO CONDENA	CONVENCIÓN
ADULTO MAYOR	0
AFRO	
BISEXUAL	
CON DISCAPACIDAD	
EXTRANJERIA	
HOMOSEXUAL	
INDIGENA	
OTRAS MENOS FRECUENTES	
SIN COND EXCEPCIONAL	1

Tabla 25. Hipótesis 3 Rango Condena

Hipótesis 4: Edad Persona Condenada

Se plantea la hipótesis de que las personas más jóvenes son aquellas que reinciden más, por lo que se procede a crear algunas variables de tiempo que permitan plantear dicha hipótesis.

RANGO EDAD	CONVENCIÓN
Entre 18 y 45 años	1
Mayores de 45 años	0

Tabla 26. Hipótesis 4 Edad Persona Condenada

Hipótesis 5: Tiempo de Condena Cumplido

El tiempo de condena realmente cumplido impacta la reincidencia. A mayor tiempo de condena cumplido, menor probabilidad de reincidir. Para construir esta hipótesis se creó la variable TiempoCumplidoAños, la cual está compuesta de restar la variable AÑO SALIDA – AÑO INGRESO.

TIEMPO CUMPLIDO	CONVENCIÓN
Menor o igual a 2 años	1
Mayor a 2 años	0

Tabla 27. Hipótesis 5 Tiempo Cumplido

Hipótesis 6: Nivel Educativo

El nivel educativo influye en la reincidencia, las personas con niveles educativos bajos reinciden delictivamente en mayor medida.

NIVEL EDUCATIVO	CONVENCIÓN MINIESTERIO DE EDUCACION[8]	NUEVA CONVENCIÓN	NUEVA CLASIFICACIÓN
ANALFABETA	SIN ESTUDIOS	BAJO	0
CICLO I	BASICA PRIMARIA (GRADO 1, 2, 3)	BAJO	0
CICLO II	BASICA PRIMARIA (GRADO 4, 5)	BAJO	0
CICLO III	BASICA SECUNDARIA (GRADO 6, 7)	ALTO	1
CICLO IV	BASICA SECUNDARIA (GRADO 8, 9)	ALTO	1
CICLO V	MEDIA (GRADO 10)	ALTO	1
CICLO VI	MEDIA (GRADO 11)	ALTO	1
TÉCNICO	EDUCACIÓN SUPERIOR	ALTO	1
TECNICO PROFESIONAL	EDUCACIÓN SUPERIOR	ALTO	1
TECNÓLOGICO	EDUCACIÓN SUPERIOR	ALTO	1
PROFESIONAL	EDUCACIÓN SUPERIOR	ALTO	1
POST GRADO	EDUCACIÓN SUPERIOR	ALTO	1
ESPECIALIZACION	EDUCACIÓN SUPERIOR	ALTO	1
MAGISTER	EDUCACIÓN SUPERIOR	ALTO	1

Tabla 28. Hipótesis 6 Nivel Educativo

En la Tabla 28 observamos que se creó la hipótesis con tan solo 2 categorías, lo cual permitirá en la etapa de modelado evaluar el impacto del Nivel Educativo en la reincidencia.

7.2.2 Modelado

Listado de Variables

Las variables que se utilizaran en el modelado de la investigación se describen en la Tabla 29. Variables para Modelado

[8] https://www.mineducacion.gov.co/1759/w3-article-233839.html?_noredirect=1

VARIABLES USADAS PARA X EN EL MODELADO		
NOMBRE DE LA VARIABLE	**NOMBRE DE LA VARIABLE**	**NOMBRE DE LA VARIABLE**
TENTATIVA_1SI	ESTADO_INGRESO_1INTRAMUROS	DELITOS CONTRA LA FAMILIA
AGRAVADO_1SI	NOROESTE	DELITOS CONTRA LA FEPÚBLICA
CALIFICADO_1SI	NORTE	DELITOS CONTRA LA LIBERTAD INDIVIDUAL Y OTRAS GARANTÍAS
TIPODESALIDA_1FUGA	OCCIDENTE	DELITOS CONTRA LA LIBERTAD INTEGRIDAD Y FORMACIÓN SEXUAL
SEXO_1MASCUL	ORIENTE	DELITOS CONTRA LA SALUD PÚBLICA
CICLOI	VIEJOCALDAS	DELITOS CONTRA LA SEGURIDAD PÚBLICA
CICLOII	AFRO	DELITOS CONTRA LA VIDA Y LA INTEGRIDAD PERSONAL
CICLOIII	BISEXUAL	DELITOS CONTRA LOS DERECHOS DE AUTOR
CICLOIV	CONDISCAPACIDAD	DELITOS CONTRA LOS RECURSOS NATURALES Y EL MEDIO AMBIENTE
ESPECIALIZACION	EXTRANJERIA	DELITOS CONTRA MECANISMOS DE PARTICIPACIÓN DEMOCRÁTICA
MAGISTER	HOMOSEXUAL	OTRAS JURISDICCIONES
POSTGRADO	INDIGENA	MESES_CONDENA_CAT
PROFESIONAL	OTRAS MENOS FRECUENTES	EDAD_CONDENADO_CATEG
TECNICO	SIN COND EXCEPCIONAL	H1_Condena Inicial Cumplida
TECNICOPROFESIONAL	DE LOS DELITOS CONTRA LOS ANIMALES	H2_Tiempo de Condena
TECNOLOGICO	DELITO CONTRA LA EFICAZ Y RECTA IMPARTICIÓN DE JUSTICIA	H3_Condición Excepcio
ACT_TRABAJO_1SI	DELITOS CONTRA EL ORDEN ECONÓMICO Y SOCIAL	H4_Edad Persona Condenada
ACT_ESTUD_1SI	DELITOS CONTRA EL PATRIMONIO ECONÓMICO	H5_Tiempo de Condena Cumplido
ACT_ENSENA_1SI	DELITOS CONTRA EL RÉGIMEN CONSTITUCIONAL Y LEGAL	H6_Nivel Educativo
HIJOS_MENOR_1SI	DELITOSCONTRALAADMINISTRACIÓNPÚBLICA	
PAIS_INTERNO_1COL	DELITOSCONTRALAEXISTENCIAYSEGURIDADDELESTADO	

Tabla 29. Variables para Modelado

Todas las variables mostradas en la tabla anterior se usaron para el modelado inicial y con ellas se evaluaron las diferentes métricas analizadas.

Creación de la Variable X y Y

Para modelar mediante las diéntenles técnicas a utilizar fue necesario crear la variable X que está compuesta de todas aquellas variables que se incluirán en el modelo las cuales se describen en la Tabla 29. Además, se crea la variable Y la cual es conformada por la variable objetivo del proyecto que es la Reincidencia y que luego de formatearse se llamó REINCI_1SI.

Creación de Test y Train

El conjunto de datos original se divide en un conjunto de Testeo y en un conjunto de Entrenamiento; en la investigación se realizará en una proporción de 70% Entrenamiento (train) y 30% Testeo (test). Ello con el fin de entrenar los modelos con una cantidad de datos suficiente y posterior a ello validar los resultados obtenidos frente al conjunto de datos de testeo.

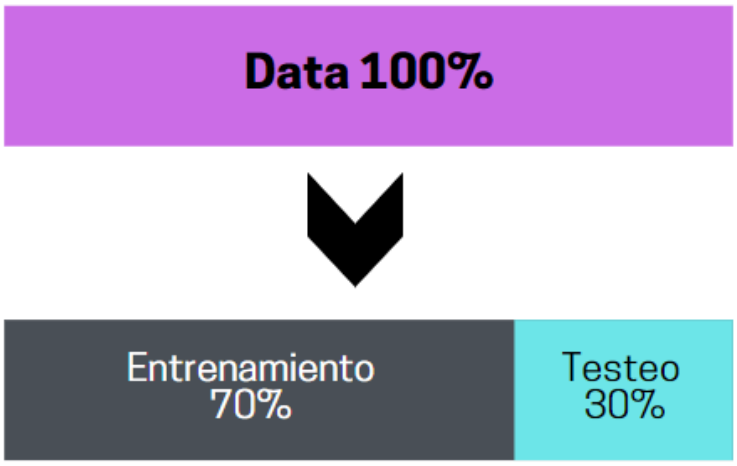

Figura 41. Partición Entrenamiento y Testeo

Balanceo de datos

En la Figura 42 observamos que la variable reincidente se encuentra Medianamente Desbalanceada siendo la categoría NO la predominante con un 69,9% de los datos y la categoría SI con el 30,1% restante.

Figura 42. Desbalanceo Variable Reincidencia

Mediante la técnica random undersampler; la cual elimina de manera aleatoria una cantidad de muestras pertenecientes a la clase mayoritaria hasta igualar ambas clases(Galindo, 2018), se hizo el balanceo de la variable reincidencia. En la Figura 43 apreciamos la metodología aplicada.

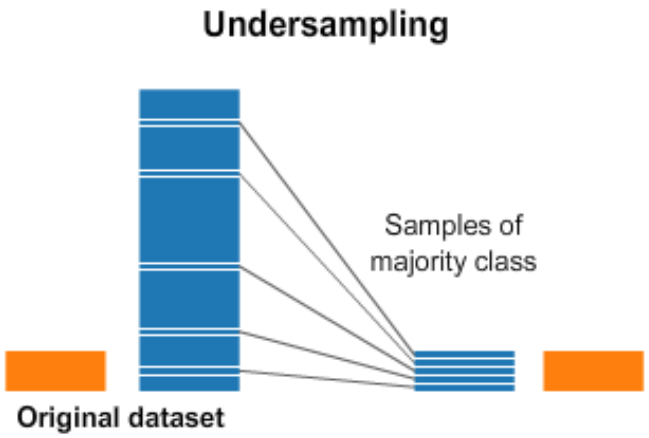

Figura 43. Undersampling. Fuente: www.kaggle.com/code/nikunjmalpan

Modelos implementados

Se implementaron los modelos de clasificación Regresión logística, Random Forest, Decision Tree, Naive vayes, Stochastic Gradient Descent, Gradient Boosted Machines GBM, K-Nearest Neighbors y Support Vector Machine SVM. Estos modelos se implementaron con la totalidad de variables previamente descritas en la Tabla 29.

7.2.3 Evaluación

En la investigación se utilizaron diferentes métricas para evaluar los modelos, una de ellas la matriz de confusión. En la Figura 44 se observa la matriz de confusión relacionada a la investigación.

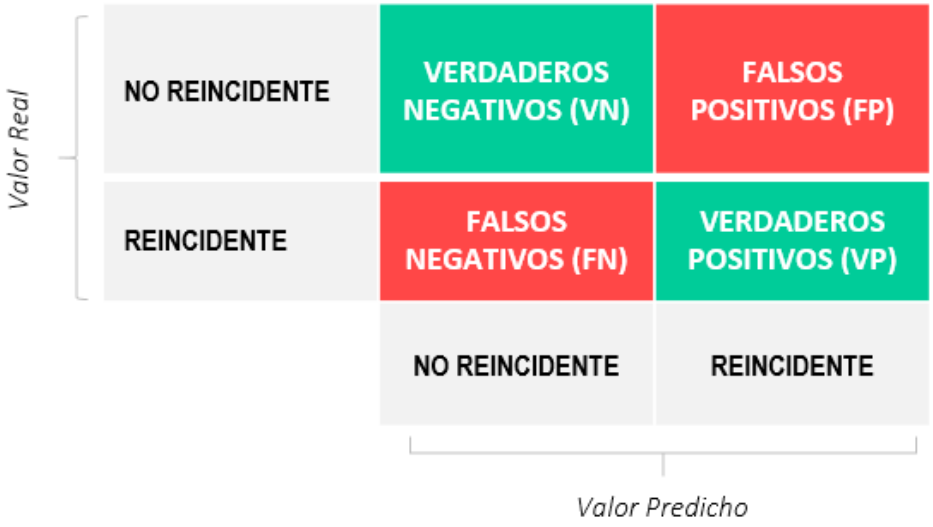

Figura 44. Matriz de Confusión Ejemplo

Como lo establecen Igual & Seguí, 2020 los significados del contenido de la matriz de confusión se observan a continuación:

- Verdaderos positivos: cuando el clasificador predice una muestra como positiva y realmente es positiva.
- Falsos positivos: cuando el clasificador predice una muestra como positiva, pero de hecho es negativa.
- Negativos verdaderos: cuando el clasificador predice una muestra como negativa y realmente es negativo.
- Falsos negativos: cuando el clasificador predice una muestra como negativa, pero de hecho es positivo.

De la matriz de confusión se desprenden las métricas exactitud, sensibilidad, precisión y especificidad las cuales se utilizaron en la evaluación de los modelos y se describen a continuación:

Métrica	Formula[9]
Exactitud (Accuracy) =	(VP+VN)/(VP+FP+FN+VN)
Precisión (Precision) =	(VP)/(VP+FP)
Sensibilidad (Recall) =	(VP)/(VP+FN)
Especificidad (Especificity) =	(VN)/(VN+FP)

Tabla 30. Fórmulas para Métricas

Si bien la Tabla 30 define las fórmulas para obtener el resultado, se explica a continuación la definición de estas métricas aplicadas a la variable objetivo, la reincidencia delictiva.

Métrica	Interpretación
*Exactitud (Accuracy)	Identificar los reincidentes y no reincidentes **en general**.
Precisión (precision)	Detectar correctamente la reincidencia **sin tener que clasificar erróneamente a personas que no van a reincidir**
*Sensibilidad (Recall)	Detectar correctamente la **reincidencia entre los verdaderos reincidentes**.
Especificidad (Especificity)	Identificar los casos de las **personas no reincidentes entre todas las no reincidentes**

Tabla 31. Interpretación Métricas

[9] Las abreviaciones descritas en esta columna son las referenciadas en la matriz de confusión de la.Figura 44.

Evaluación de modelos

Regresión logística

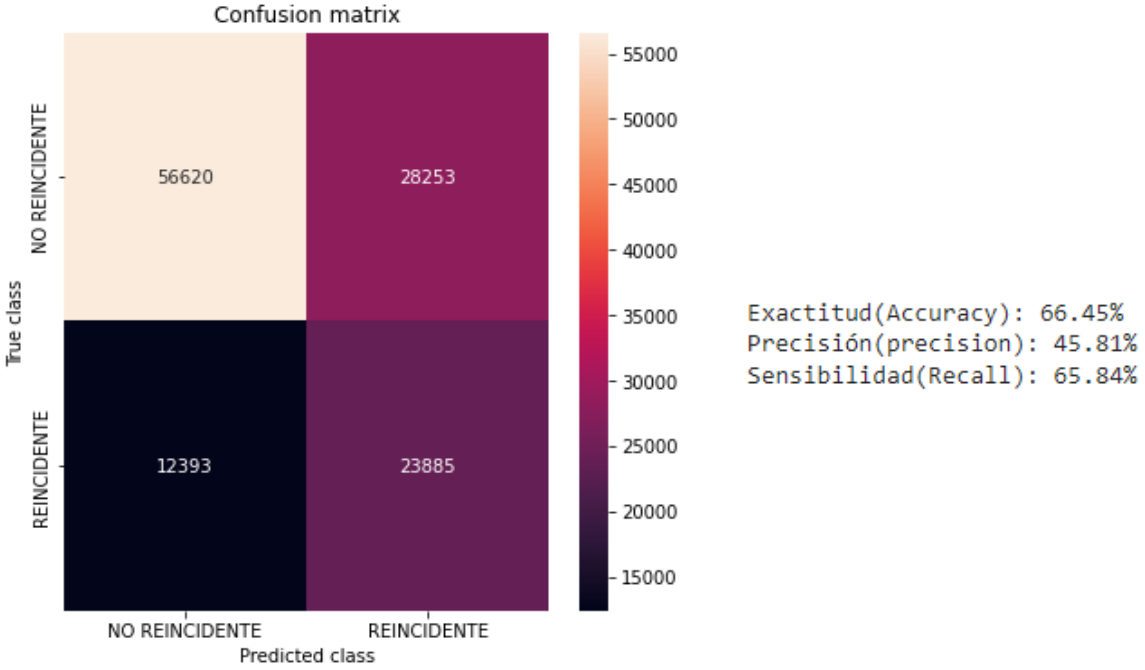

Figura 45. Matriz Confusión Regresión Logística

En la Figura 45 observamos que el modelo tiene una precisión muy baja fruto de su alta clasificación de falsos positivos. De igual manera la exactitud y la sensibilidad tienen valores bajos, menores a los esperados en el modelado.

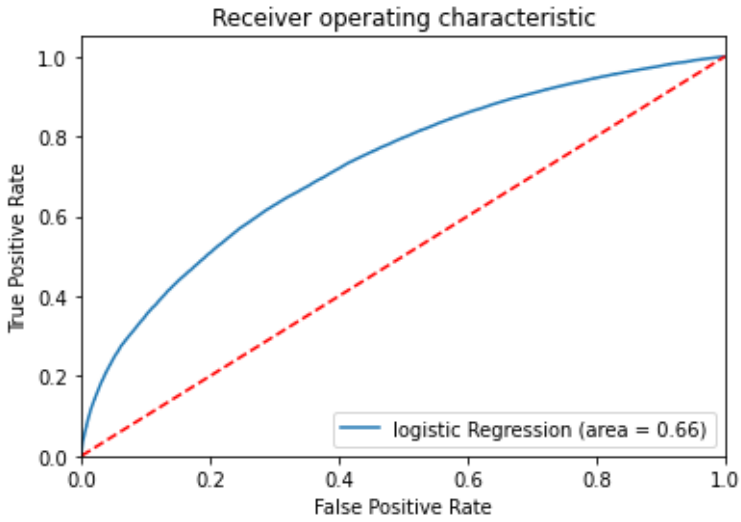

Figura 46. Área Debajo de la Curva Regresión Logística

El indicador de área bajo la curva de este modelo dio un resultado de 66% lo que indica que el modelo es capaz de clasificar los verdaderos reincidentes con una probabilidad de 66 por cada 100 personas.

Decision Tree

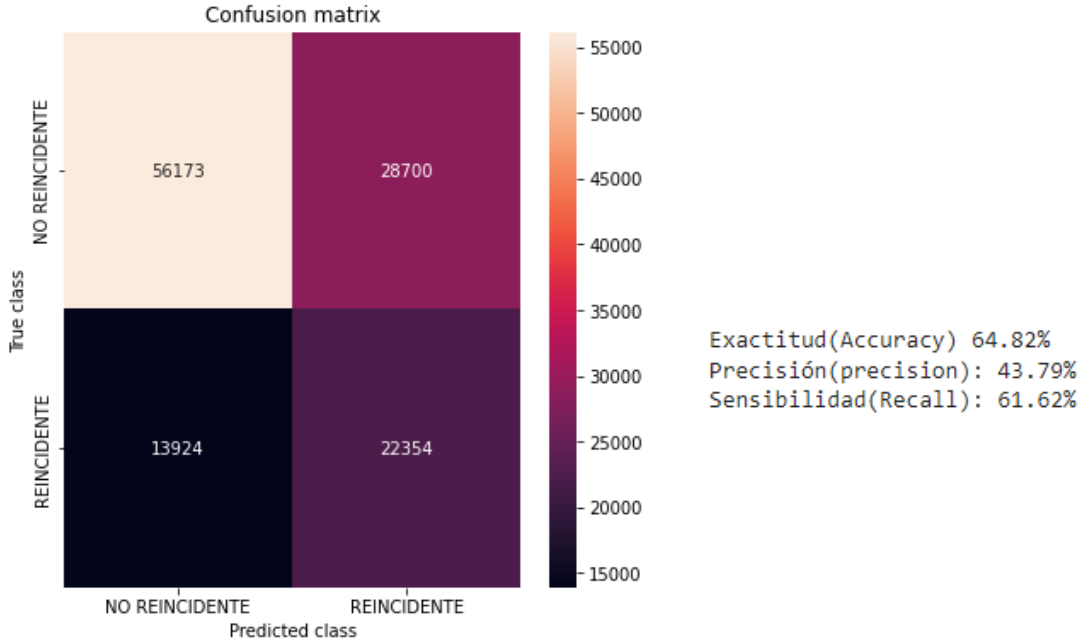

Figura 47. Matriz Confusión Decision Tree

De la matriz de confusión de la Figura 47 podemos decir que el modelo tiene una precisión muy baja, menor al 50% explicada por su alta clasificación de falsos positivos. De igual manera la exactitud y la sensibilidad tienen valores bajos, menores a los esperados en el modelado.

Figura 48. Área Bajo la Curva Decision Tree

El área bajo la curva de este modelo dio un resultado de 64% lo que lo sitúa en esta métrica por debajo del modelo de regresión logística y aún distante de ser el modelo ideal para clasificar la variable objetivo.

Random Forest

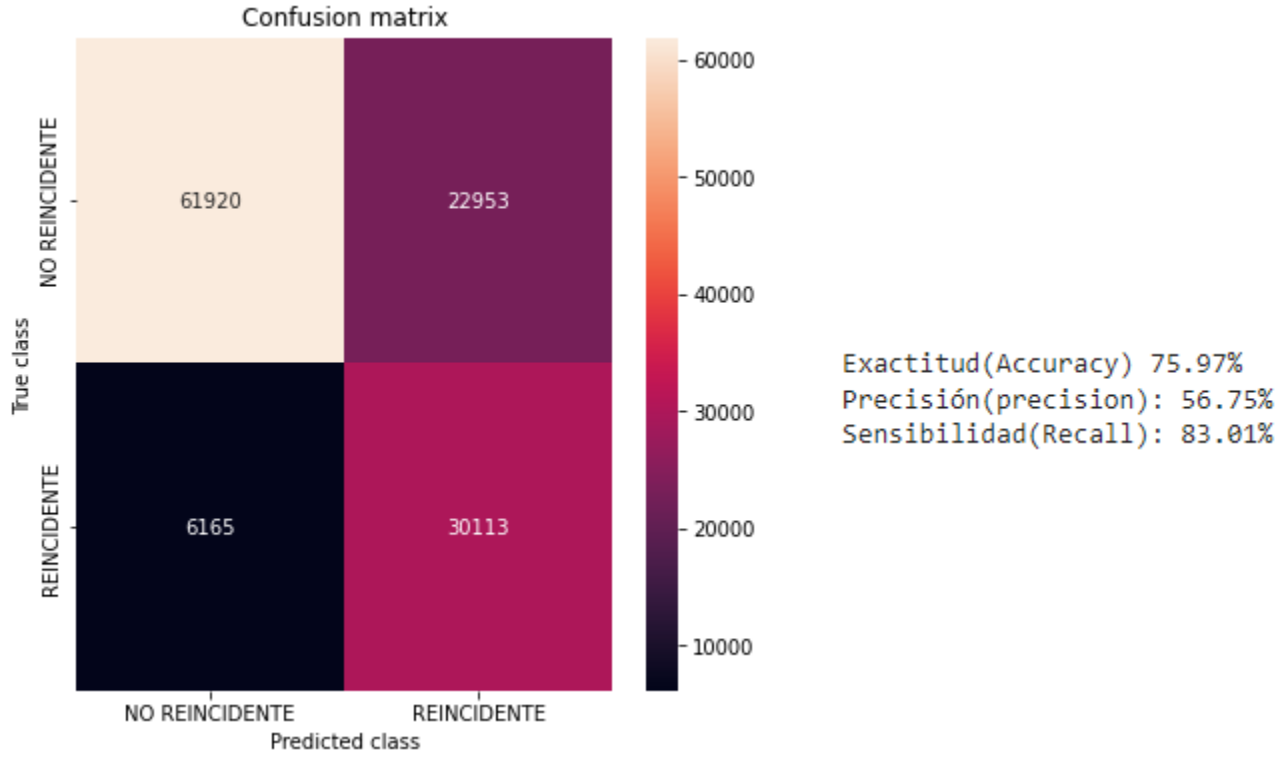

Figura 49. Matriz Confusión Random Forest

En la matriz de confusión del modelo Random Forest observamos los mejores resultados obtenidos hasta el momento. Se tiene la mayor proporción de verdaderos positivos y verdaderos negativos lo que se refleja en todas las métricas especialmente en la exactitud (capacidad de clasificar reincidentes y no reincidentes en general) y en la sensibilidad (capacidad de clasificar los reincidentes) la cual al criterio del investigador es la métrica más importante a evaluar.

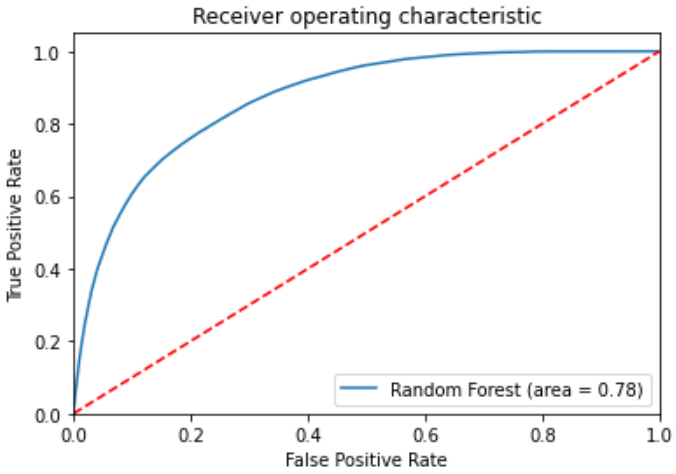

Figura 50. Área Bajo la Curva Random Forest

En la Figura 50 observamos la mejor área bajo la curva obtenida en el modelado, logrando un valor de 78% lo cual indica que el modelo clasificatorio por cada 100 personas analizadas está en la capacidad de clasificar correctamente 78 personas verdaderamente reincidentes.

Naive Bayes

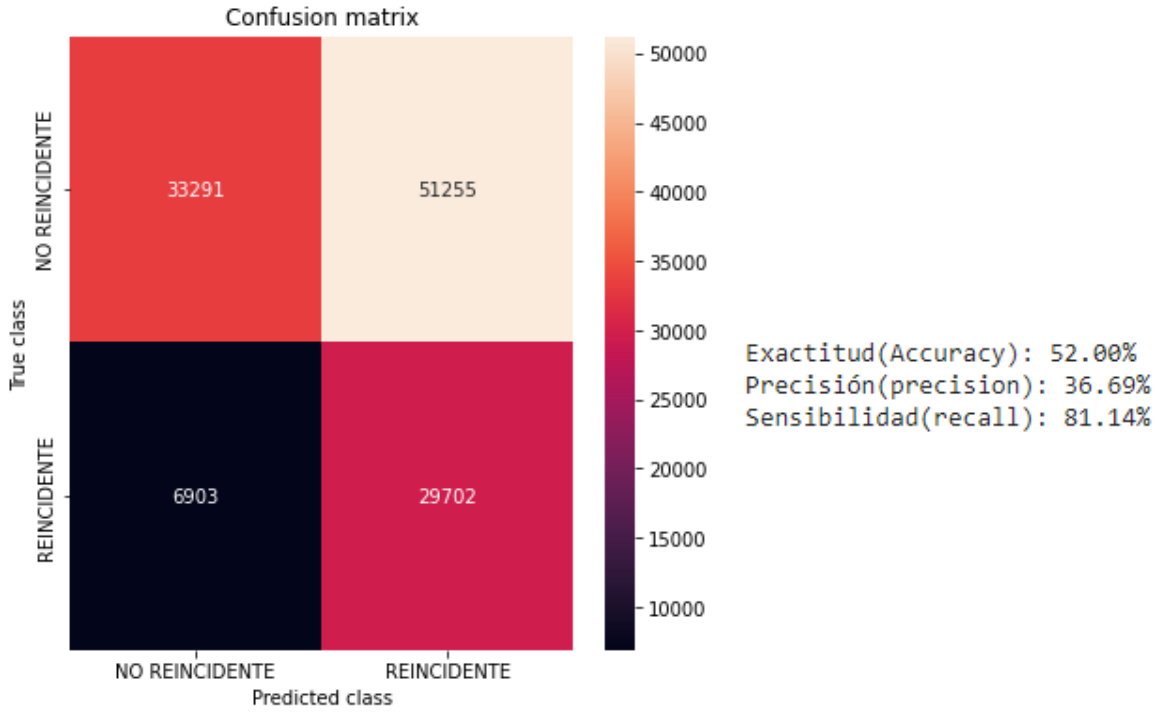

Figura 51. Matriz de Confusión Naive Bayes

En la Figura 51 se aprecia que la sensibilidad del modelo es alta por tanto tiene capacidad de clasificar bien a los reincidentes dentro de los verdaderamente reincidentes pero su capacidad de clasificar a nivel general (exactitud) es muy baja por lo que no tiene la capacidad de clasificar a los no reincidentes de manera correcta en alto porcentaje.

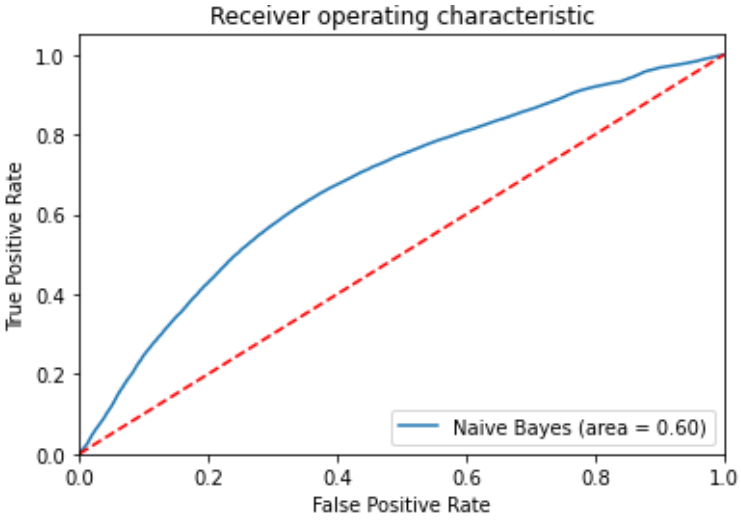

Figura 52. Área Bajo la Curva Naive Bayes

El área bajo la curva de este modelo observada en la Figura 52 dio un resultado de 60% lo que lo sitúa 18 puntos porcentuales por debajo del modelo de random forest.

Stochastic Gradient Descent

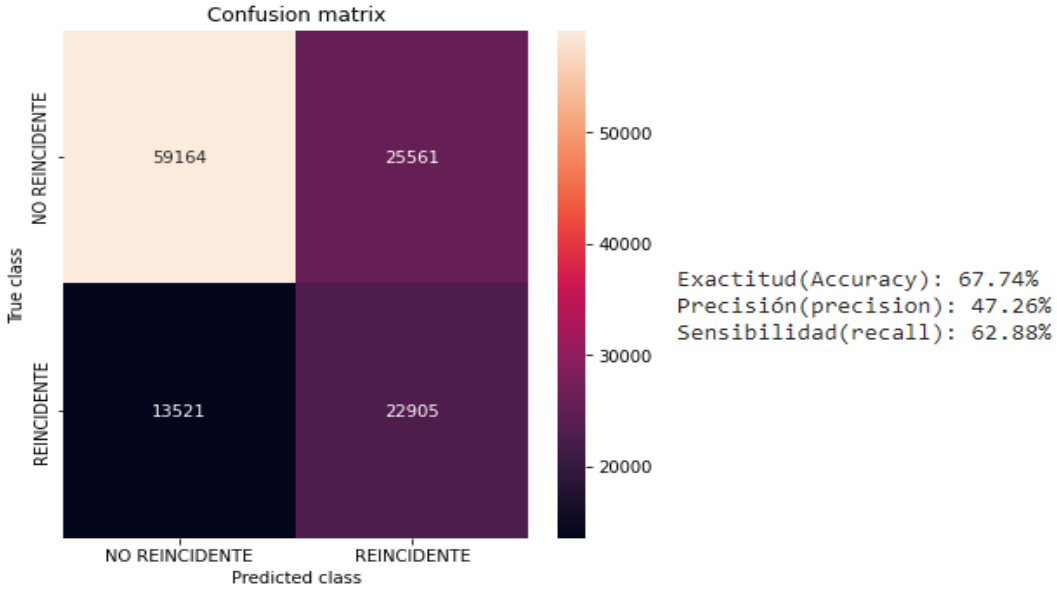

Figura 53. Matriz de Confusión Stochastic Gradient Descent

La matriz de confusión de este modelo arrojó un resultado más balanceado entre métricas, pero a pesar de ello sus resultados son bajos, en todos los casos menores a 70% por lo que no se usará como modelo definitivo para clasificar.

Figura 54. Área Bajo la Curva Stochastic Gradient Descent

En este modelo se obtuvo un puntaje de 62% en la métrica de área bajo la curva como se observa en la Figura 54. Este resultado se sigue ubicando por debajo del modelo más prometedor.

Gradient Boosted Machines GBM

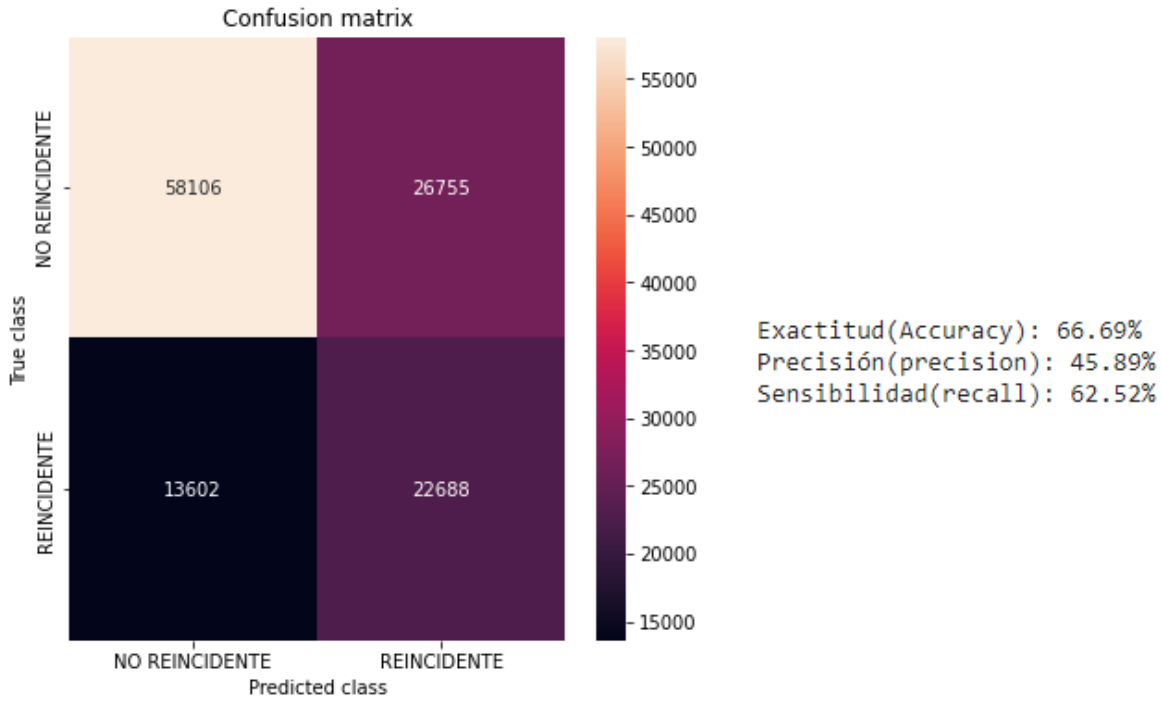

Figura 55. Gradient Boosted Machines Gbm

El modelo Gradient Boosted Machines GBM no tiene una alta capacidad para clasificar los reincidentes mostrando un resultado de clasificación correcto de 63 personas por cada 100 analizadas. La métrica de exactitud si bien es un poco mejor con un 67% aproximadamente está por debajo de la obtenida en el Random Forest y otros modelos.

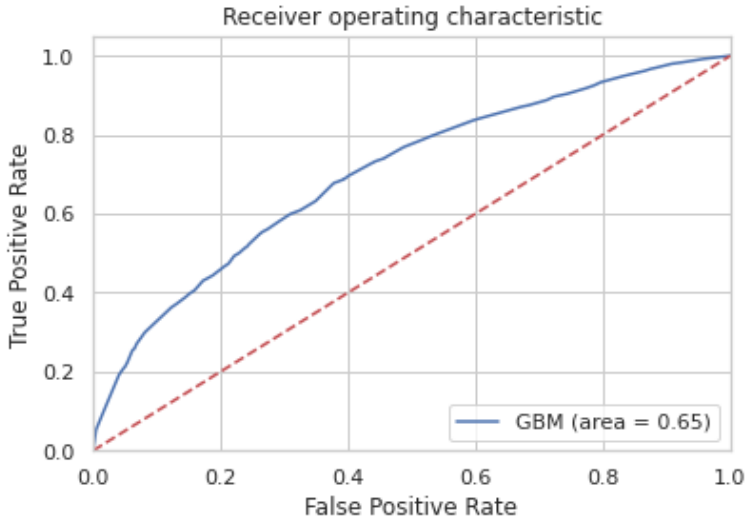

Figura 56. Área Bajo la Curva Gradient Boosted Machines Gbm

En este modelo se obtuvo un puntaje de 65% en la métrica de área bajo la curva. Este resultado es menor al obtenido en otros modelos y aunado a los valores de las métricas anteriormente evaluadas lo ubican por debajo de otros modelos.

K-Nearest Neighbors

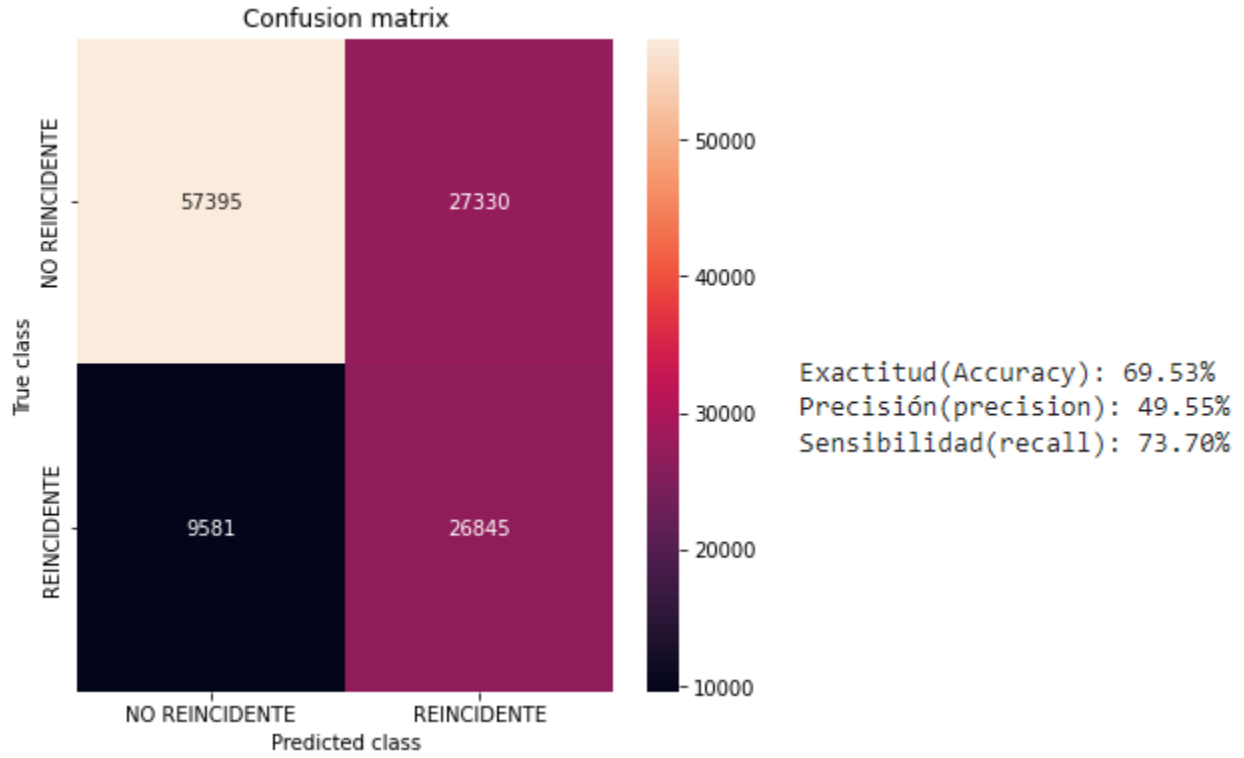

Figura 57. Matriz de Confusión K-Nearest Neighbor

La matriz de confusión del modelo k-nearest neighbor presenta una sensibilidad de 73,7% y una exactitud casi del 70% lo que lo sitúa dentro de valores a considerar. Sin embargo, existen otros modelos ya presentados con mejores resultados.

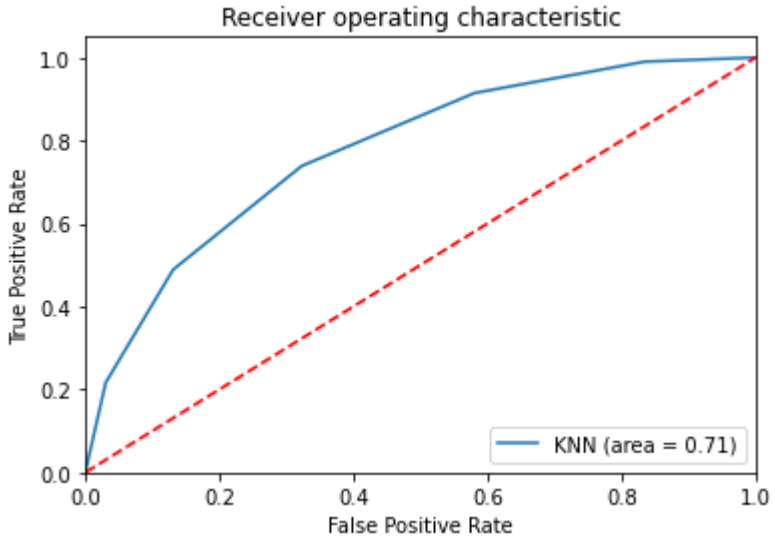

Figura 58. Área Bajo la Curva K-Nearest Neighbors

El área bajo la curva vista en la Figura 58 es mejor que la de otros modelos vistos, indicando que de cada 100 personas analizadas el modelo tiene la capacidad de clasificar correctamente 71 personas.

Evaluación de los resultados

Se evaluaron y compararon los resultados obtenidos en todos los modelos implementados. Se revisaron las métricas de Sensibilidad ("Recall") y Exactitud ("Accuracy") que en nuestro caso se consideran las más relevantes. Los resultados se observan en la Tabla 32.

	Métricas			
Modelo	**Exactitud (Accuracy)**	**Precisión (Precision)**	**Sensibilidad (Recall)**	**Área Bajo la Curva**
LogisticRegression	66,45%	45,81%	65,84%	66,00%
Decision_Tree	64,82%	43,79%	61,62%	64,00%
Random_Forest	75,97%	56,75%	83,01%	78,00%
Naive_Bayes	52,00%	36,69%	81,14%	60,00%
Stochastic Gradient Descent	67,74%	47,26%	62,88%	62,00%
GradientBoostingClassifier GBM	66,69%	45,89%	62,52%	65,00%
KNearest_Neighbors	69,53%	49,55%	73,70%	

Tabla 32. Evaluación modelos

En todas las métricas evaluadas, el modelo Random_Forest fue el de mejor desempeño, superando en exactitud, precision, sensibilidad y área bajo la curva a todos lo demás modelos. A continuación, se muestra el resumen de los resultados de mayor interés para la investigación:

Random Forest

Métrica	Valor	Interpretación
Accuracy (Exactitud)	76%	Tiene la capacidad de clasificar correctamente 760 personas como Reincidentes o No Reincidentes de manera general, por cada 1000 personas analizadas
Recall (Sensibilidad)	83%	El modelo clasifica correctamente 830 personas como Reincidentes por cada 1000 personas que en realidad iban a ser Reincidentes

Tabla 33. Interpretación random forest

Teniendo en cuenta que uno los objetivos específicos es *Implementar un modelo clasificatorio que permita explicar la reincidencia delictiva en Colombia de las personas que hayan estado bajo la vigilancia del Instituto Nacional Penitenciario y Carcelario INPEC.,* se hace importante evaluar los indicadores con los que se puede lograr este objetivo y por tanto la exactitud y la sensibilidad se convierten en métricas de importancia en la evaluación.

Con una exactitud superior al 75% se podrá garantizar que un porcentaje importante de las personas analizadas sean correctamente clasificadas como reincidentes o no reincidentes. Indicador que es susceptible a la mejora en un escenario donde se cuente con mayor cantidad de variables relacionadas y poder computacional.

La sensibilidad es a criterio nuestro el indicador más importante de todos, ya que nos da cuenta del grado de clasificación al que podemos llegar con las personas verdaderamente reincidentes que para nuestro modelo tiene un valor de 83%. Esto permitirá clasificar de manera correcta a la gran mayoría de las personas analizadas que verdaderamente van a ser reincidentes.

Con estas métricas se esta en la capacidad de explicar la reincidencia en los porcentajes descritos anteriormente. Teniendo la información de una persona asociada a las variables de esta investigación se podría clasificar cómo reincidente o no, con un exactitud del 76% de los casos.

Selección de variables (feature importantes)

El modelo Random Forest cuenta con una función llamada feature importances, la cual permite realizar la selección de un subconjunto de variables más representativas según su importancia dentro del modelo (Sotiroudis et al., 2020). De esta manera podemos darnos cuenta cuales son las variables que más impactan la reincidencia delictiva dentro de nuestro conjunto de datos. **¡Error! No se encuentra el origen de la referencia.**

En la Tabla 34 se describen las quince variables más importantes según el modelo random forest y las cuales en conjunto describen la reincidencia delictiva en un 63,21%. Para conocer la calificación total de las variables utilizadas, ver Anexo B.

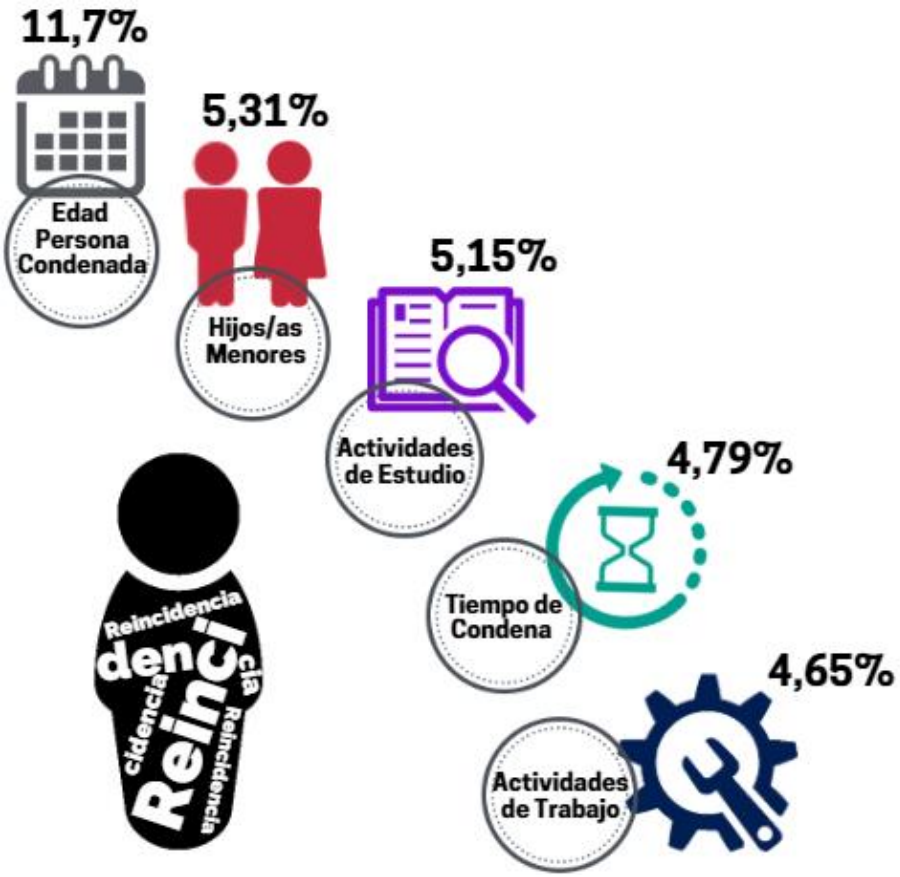

Figura 59. Principales Variables que Impactan la Reincidencia

Figura 59 observamos las cinco variables que más impactan la reincidencia delictiva según el modelado realizado con la técnica feature importances, variables que juntas explican la reincidencia en un 31,6%.

Variables de la persona:

EDAD_CONDENADO_CATEG, HIJOS_MENOR_1SI, ACT_ESTUD_1SI, ACT_TRABAJO_1SI, SEXO_1MASCUL: Se puede observar que hay variables relacionadas con la persona tales como la edad cuando fue condenado, si tiene hijos menores de edad, el sexo y si realiza actividades de estudio y trabajo dentro de los centros de reclusión. Las variables relacionadas con estudio y trabajo se deben analizar en profundidad ya que en primera instancia parecen ser un resultado atípico, pero podrían tener alguna explicación conociendo en profundidad su naturaleza y funcionamiento.

Variables de lo jurídico:

MESES_CONDENA_CAT, H6_TiempoCumplidoAños, AGRAVADO_1SI, DELITOS CONTRA EL PATRIMONIO ECONÓMICO, CALIFICADO_1SI, H1_CondenaInicial_Cumpl_1NO: Estas variables están relacionadas con lo jurídico en función con la condena. Las variables meses condena, tiempo cumplido en años y condena inicial cumplida, relacionan el tiempo categorizado que debía cumplir la persona condenada, el tiempo en años verdaderamente cumplido por la persona y si la persona cumplió o no el tiempo inicial de condena proferido. Agravado y calificado son variables que indican si los delitos cometidos por la persona tienen esa connotación jurídica. Por ultimo delitos contra el patrimonio económico describe que la persona cometió un delito relacionada con esta categoría jurídica.

Variables de lo institucional:

OCCIDENTE, VIEJO CALDAS, NOROESTE, ORIENTE: Estas variables están relacionadas con lo locativo e indican en qué regional del INPEC se está cumpliendo la condena, indicando que, en ciertas regionales las personas son más propensas a reincidir.

Los diferentes valores que puedan tomar las variables anteriormente mencionadas influyen en que una persona tenga mayor o menor posibilidad de cometer nuevamente una conducta delictiva, de ser reincidente.

7.3 Fase O3: Descripción de estrategias que puedan ayudar a disminuir la reincidencia delictiva en Colombia

- Revisión, análisis y evaluación por parte del INPEC y de las autoridades competentes de los programas de Estudio y Trabajo implementados al interior de las penitenciarias. Esto dado que fueron variables que resultaron ser influyentes en la reincidencia delictiva y al ser actividades participes del proceso de resocialización están en contravía de su naturaleza.

- Con las personas que tienen perfiles de mayor riesgo relacionado a variables demográficas y sociales tales como: EDAD_ CONDENADO_CATEG, HIJOS_MENOR_1SI, ACT_ESTUD_1SI, ACT_TRABAJO_1SI, SEXO_1MASCUL, realizar acompañamiento especial mediante programas educativos y laborales para la y el pospenado.

- Realizar por parte del INPEC un diagnostico que aborde las generalidades y particularidades de la reclusión en las regionales OCCIDENTE, VIEJO CALDAS, NOROESTE, ORIENTE ya que estas regionales demostraron tener influencia en la reincidencia delictiva.

- Realizar campañas gubernamentales de persuasión con los perfiles de mayor riesgo de reincidencia mediante la difusión de los programas educativos gratuitos, subsidios, líneas de orientación y todo tipo de ayudas estatales que puedan ser útiles para estas personas.

8. Impactos

8.1 Impactos Sociales

Conocer las variables que más impactan la reincidencia delictiva le puede permitir al estado disminuirla mediante la intervención con políticas de prevención y persuasión, logrando que cada vez sean menos las personas que continúen la carrera delictiva dentro del país. Como lo mencionan Garzón et al. (2018) existen gran cantidad de programas de rehabilitación en búsqueda de reducir la reincidencia, destacándose las oportunidades educativas, desarrollo cognitivo, apoyo psicológico, mejora de las habilidades para el trabajo y tutorías dentro y fuera de las prisiones, como las acciones con mejores resultados identificadas en Estados Unidos.

El uso de formas alternativas diferentes al cumplimiento intramuros de la pena, para aquellas personas con perfiles de menor riesgo de reincidencia podría rendir beneficios sociales para los núcleos familiares y círculos sociales cercanos de la persona condenada; tal como lo relata Larrota et al. (2018) los efectos y las consecuencias de la privación de la libertad afectan de manera importante a los núcleos cercanos del o la condenada, principalmente pareja e hijos, amigos y demás personas con alguna relación cercana.

El uso de las tecnologías de la información con vertientes tales como la analítica de datos, aporta un espectro amplio de recursos, en el que el eje social de nuestra sociedad se puede impactar positivamente encontrando soluciones a problemas desde lo trivial hasta lo complejo.

8.1 Impactos Económicos

En febrero del año 2022 la población reincidente con pago de condena intramuros sumaba 22.576 personas condenadas (INPEC, 2022) la cual representa un costo económico de $704 mil millones de pesos anuales (INPEC, 2022). Para poner en contexto las cifras anteriores y a manera de ejemplo, si el gobierno lograra reducir en un cincuenta por ciento la población actual reincidente, pasando de 22.576 personas a

11.288 personas, obtendría un ahorro de $352 mil millones de pesos anuales[10], lo que equivale a un poco más del presupuesto general de la nación asignado para el Ministerio de Ciencia, Tecnología e Innovación vigencia fiscal de 2022[11] el cual asciende aproximadamente a $331 mil millones de pesos anuales.

Estas cifras dan cuenta del alto sobrecosto económico que representa para el gobierno la población reincidente y el por qué es impactante desde el punto de vista fiscal la reducción de sus porcentajes.

[10] Cálculos realizados con un costo anual por persona condenada de $31.179.764 y tomando como base la población actual de reincidentes 22.576 personas, sin tener en cuenta la fluctuación futura que estos valores pudieran tomar. Fuente: (INPEC, 2022)

[11] Decreto 1793 de 2021 por el cual se liquida el Presupuesto General de la Nación para la vigencia fiscal 2022.

9. Conclusiones

Se pudo comprobar que haciendo uso de las variables y registros disponibles fue posible implementar un modelo que permitiera clasificar a las personas analizadas según su posibilidad de reincidencia delictiva. Con el modelo de analítica implementado se esta en la capacidad de explicar la reincidencia delictiva en porcentajes sobresalientes para esta primera investigación.

El modelo que mejores resultados arrojó fue el Random Forest, mostrando valores superiores en todas las métricas evaluadas. Con este modelo estamos en la capacidad de clasificar correctamente 76 personas condenadas por cada 100 evaluadas. Además, nos permite lograr un acierto de 81% cuando categorizamos personas como reincidentes.

Las variables EDAD_CONDENADO_CATEG, HIJOS_MENOR_1SI, ACT_ESTUD_1SI, MESES_CONDENA_CAT, ACT_TRABAJO_1SI son las cinco variables que mas impactan la Reincidencia delictiva en la investigación, explicándola por si solas en un 31,6%.

La variable EDAD_CONDENADO_CATEG es aquella que describe la edad que tenía la persona al ser condenada. Dicha variable fue construida en la investigación a partir de la FECHA_INGRESO y ANO_NACIMIENTO; según el feature importances esta es la variable que más aporta al modelo por lo que se trata de un descubrimiento importante. La edad de la persona condenada, influye en su posibilidad de reincidir, siendo las personas más jóvenes aquellas más propensas a hacerlo.

MESES_CONDENA_CAT que es una variable relacionada al tiempo de duración de la condena y según el feature importances una de las variables más relevantes, indica que la duración de la condena está relacionada con mayor o menor posibilidad de reincidir, siendo las condenas cortas aquellas más propensas por la reincidencia.

Las personas que estudian dentro de los centros de reclusión reinciden en mayor medida (ACT_ESTUD_1SI) lo cual va en contravía de las políticas penitenciarias de resocialización. Se debe hacer un estudio a fondo con esta variable para conocer su real naturaleza; el resultado podría deberse a factores que desconocemos en esta investigación. Este comportamiento también ocurre con las personas que realizan actividades de trabajo dentro de los centros de reclusión (ACT_TRABAJO_1SI).

Tener hijos menores (HIJOS_MENOR_1SI) es una condición que influye en la reincidencia delictiva. Estar a cargo de una responsabilidad como esta podría ser un elemento que condicione el comportamiento criminal de una persona. Para ahondar en los aspectos sociales de la persona se requerirían mayor cantidad de variables de este tipo, acompañado además de un trabajo multidisciplinar con otras áreas del conocimiento que permita enriquecer la intepretacion de cierto tipo de variables.

10. Recomendaciones

Para obtener mejores resultados en las métricas de los modelos, se requieren mayor cantidad de variables. Variables de las que no se disponía y relacionadas a factores sociales como el estado civil, estrato, nucleo familiar y variables relacionadas a los centros de reclusión, podrían ayudar a clasificar de mejor manera la reincidencia delictiva.

Implementar las modalidades de cumplimiento de condenas diferentes a la privación de la libertad intramuros para aquellas personas con baja posibilidad de reincidencia, podría liberar presión al sistema penitenciario, ayudando a reducir los índices de hacinamiento. Esto también podría reportar beneficios económicos para el gobierno entendiendo que las modalidades de prisión domiciliaria y la vigilancia electrónica son mecanismos con un menor coste que la detención intramuros.

Como se mencionó en la fase de evaluación, los resultados de esta investigación deben servir para que las entidades gubernamentales competentes realicen esfuerzos encaminados a la reducción de la reincidencia delictiva. El acompañamiento psicosocial mediante programas laborales, educativos, soporte psicológico y todas aquellas actividades que puedan dar ayuda a las personas pospenadas con mayor riesgo de reincidencia, deben ser una de las prioridades en la construcción de nuevas políticas penitenciarias y carcelarias.

Este trabajo investigativo tiene la única pretensión de aportar al debate académico en materia penitenciaria y se espera que sus resultados puedan ser de alguna utilidad en la búsqueda de la reducción de los índices de reincidencia, siempre desde una mirada social y en ningún momento pretende estigmatizar grupos o minorías ni que sus resultados se usen para políticas represivas de perfilamiento y persecución de personas.

11. Agradecimientos

Al profesor Juan Manuel Barco por su compromiso y dedicación en los módulos impartidos. Gracias a ese conocimiento adquirido pude llevar a cabo esta investigación.

A mis compañeros David y Héctor, quienes hicieron que el camino de aprendizaje fuera más enriquecedor y agradable.

A María Paulina Bernal por todo su apoyo jurídico en la investigación. Su disposición y voluntad siempre fueron una constante y una fortuna para este trabajo.

A Néstor Jaime Castaño por animarme e impulsarme a realizar esta maestría.

A Juan Alejandro Trujillo por los aportes realizados.

A. Anexo: base de datos INPEC.xlsx

https://docs.google.com/spreadsheets/d/1Nr-hs9GNQAfHqQQR2fQAXlljq5TDhK7H/edit?usp=sharing&ouid=108301764351664475371&rtpof=true&sd=true

B. Anexo: Google Colab con desarrollo técnico del proyecto

https://colab.research.google.com/drive/1Uvfse42nQu6bSie9uXu7mu3VYf137W-i?usp=sharing

Referencias bibliográficas

Acosta, C. (2021). *El hacinamiento en las cárceles colombianas*. https://www.asuntoslegales.com.co/actualidad/el-hacinamiento-en-las-carceles-colombianas-es-de-20-a-marzo-segun-cifras-del-inpec-3133024

Arancibia, J. A. G. (2009). *Metodología para la Definición de Requisitos en Proyectos de Data Mining (ER-DM)*. Universidad Politécnica de Madrid.

Ariza, L. J., Iturralde, M., & Tamayo Arboleda, F. L. (2020). De la cárcel al barrio. Caracterización cualitativa de la reincidencia criminal en Colombia. *Estudios de Derecho*, *78*(171). https://doi.org/10.17533/udea.esde.v78n171a03

Canhoto, A. I. (2020). Leveraging machine learning in the global fight against money laundering and terrorism financing : An affordances perspective. *Journal of Business Research*, *October*. https://doi.org/10.1016/j.jbusres.2020.10.012

Chen, H., Cai, D., Dai, W., Dai, Z., & Ding, Y. (2020). Charge-based prison term prediction with deep gating network. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 6362–6367. https://doi.org/10.18653/v1/d19-1667

Corporación Excelencia en la Justicia. (2021). *Tasa de criminalidad en Colombia*. https://cej.org.co/indicadores-de-justicia/criminalidad/tasa-de-criminalidad/

Cuervo, K., Villanueva, L., & Prado-Gascó, V. (2017). Predicción De La Reincidencia Juvenil Mediante El Yls/Cmi Y Baremos Para Su Valoración Youth Recidivism Prediction Using the Yls/Cmi and Norms for Assessment. *Revista Mexicana de Psicología*, *34*, 24–36.

de Graaf, R. (2019). Managing Your Data Science Projects. In *Managing Your Data Science Projects*. https://doi.org/10.1007/978-1-4842-4907-9

Decreto 1242, 1 (1993).

Galindo, A. (2018). *Algoritmos de Clasificacion para datasets*.

Garzón, J., María, L., & Suárez, M. (2018). ¿Qué hacer con la reincidencia delincuencial? *Fundación Ideas Para La Paz*, 1–39.

Ghasemi, M., Anvari, D., Atapour, M., Stephen wormith, J., Stockdale, K. C., & Spiteri, R. J. (2021). The Application of Machine Learning to a General Risk–Need Assessment Instrument in the Prediction of Criminal Recidivism. *Criminal Justice and Behavior*, *48*(4), 518–538. https://doi.org/10.1177/0093854820969753

Gholamzadeh Nabati, E., & Thoben, K. D. (2016). On applicability of big data analytics in the closed-loop product lifecycle: Integration of CRISP-DM standard. *IFIP Advances in Information and Communication Technology*, *492*, 457–467. https://doi.org/10.1007/978-3-319-54660-5_41

IBM. (2015). *CRISP-DM Help Overview*. https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=dm-crisp-help-overview

Igual, L., & Seguí, S. (2020). *Introduction to Data Science*. https://doi.org/10.4018/978-1-7998-3053-5.ch001

INPEC. (2022). *Informe Estadistico INPEC - Población pivada de la libertad Febrero 2022* (Issue 27). https://www.inpec.gov.co/web/guest/estadisticas/-/document_library/TWBuJQCWH6KV/view_file/1421306?_com_liferay_document_library_web_portlet_DLPortlet_INSTANCE_TWBuJQCWH6KV_redirect=https%3A%2F%2Fwww.inpec.gov.co%2Fweb%2Fguest%2Festadisticas%2F-%2Fdocument

INPEC, I. N. P. y C. (2020). *Informe Estadístico 2020 Poblacion Privada de la Libertad - INPEC - Oficina Aseora de Planeacion Grupo Estadistica* (Vol. 2). https://www.inpec.gov.co/documents/20143/965447/INFORME+ESTADISTICO+FEBRERO+.pdf/fcf8284f-99e6-81f8-d4de-763cfc7f052d

Karimi-Haghighi, M., & Castillo, C. (2021). Enhancing a recidivism prediction tool with machine learning: Effectiveness and algorithmic fairness. *Proceedings of the 18th International Conference on Artificial Intelligence and Law, ICAIL 2021*, 210–214. https://doi.org/10.1145/3462757.3466150

Kelsen, H. (2009). La Teoría Pura del Derecho. In *Revista de Ciencias Jurídicas* (Vol. 4, Issue 9789502308869).

Kourtit, K., Mazurencu, M., Pele, M., Nijkamp, P., & Traian, D. (2021). Safe cities in the new urban world : A comparative cluster dynamics analysis through machine learning. *Sustainable Cities and Society*, *66*(February 2020), 102665. https://doi.org/10.1016/j.scs.2020.102665

Larrota, R., Gaviria, A. M., Mora, C., & Arenas, A. (2018). Aspectos criminogénicos de la reincidencia y su problema. *Revista de La Universidad Industrial de Santander. Salud*, *50*(2), 158–165. https://doi.org/10.18273/revsal.v50n2-2018007

Li, S., Zhang, H., & Su, S. (2020). Prison Term Prediction on Criminal Case

Description with Deep Learning Prison Term Prediction on Criminal Case Description with Deep Learning. *Computers, Materials & Continua*, *62*(January 2019), 1217–1231. https://doi.org/10.32604/cmc.2020.06787

Li, Z., Zhang, T., Jing, X., & Wang, Y. (2021). Facial expression-based analysis on emotion correlations , hotspots , and potential occurrence of urban crimes. *Alexandria Engineering Journal*, *60*(1), 1411–1420.

Liu, X. Y., Wu, J., & Zhou, Z. H. (2009). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *39*(2), 539–550. https://doi.org/10.1109/TSMCB.2008.2007853

Lussier, P., Deslauriers-varin, N., Collin-santerre, J., & Bélanger, R. (2019). Using decision tree algorithms to screen individuals at risk of entry into sexual recidivism. *Journal of Criminal Justice*, *63*(May), 12–24. https://doi.org/10.1016/j.jcrimjus.2019.05.003

Medium. (2019). *Skills Needed for Data Science*. Introduction to Data Science. https://medium.com/@jrendz/pengenalan-data-science-b49a52eeef9c

Müller, A. C., & Guido, S. (2016). Introduction to with Python Learning Machine. In *Proceedings of the Speciality Conference on Infrastructure Condition Assessmenr: Art, Science, Practice*.

Pete, C., Julian, C., Randy, K., Thomas, K., Thomas, R., Colin, S., & Wirth, R. (2000). CRISP-DM 1.0. *CRISP-DM Consortium*, 76.

Prabhu, C. S. R., Chivukula, A. S., Mogadala, A., Ghosh, R., & Jenila Livingston, L. M. (2019). Big data analytics: Systems, algorithms, applications. In *Big Data Analytics: Systems, Algorithms, Applications*. https://doi.org/10.1007/978-981-15-0094-7

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, *1*(1), 51–59. https://doi.org/10.1089/big.2013.1508

Qazi, N., & Wong, B. L. W. (2019). An interactive human centered data science approach towards crime pattern analysis. *Information Processing and Management*, *56*(6). https://doi.org/10.1016/j.ipm.2019.102066

Reid, J. A., & Beauregard, E. (2020). Exploring a machine learning approach:

Predicting death in sexual assault. *Journal of Criminal Justice*, *71*(June), 101741. https://doi.org/10.1016/j.jcrimjus.2020.101741

Schröer, C., Kruse, F., Marx, J., Kruse, F., & Marx, J. (2021). A Systematic Literature Review on Applying Process Model on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*(2019), 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Silva, J., Zilberman, J., Romero, L., Bonerge, O., Herazo-, Y., Silva, J., Zilberman, J., Romero, L., Bonerge, O., & Herazo-, Y. (2020). ScienceDirect ScienceDirect Identification of Patterns of Fatal Injuries in Humans through Big Identification of Patterns of Fatal Injuries in Humans through Big Data Data and Networks. *Procedia Computer Science*, *170*, 893–898. https://doi.org/10.1016/j.procs.2020.03.114

Sotiroudis, S. P., Goudos, S. K., & Siakavara, K. (2020). Feature Importances: A Tool to Explain Radio Propagation and Reduce Model Complexity. *Telecom*, *1*(2), 114–125. https://doi.org/10.3390/telecom1020009

Swamynathan, M. (2019). Mastering Machine Learning with Python in Six Steps. In *Mastering Machine Learning with Python in Six Steps*. https://doi.org/10.1007/978-1-4842-4947-5

Tiwari, L., Ranjan, R., Sardana, N., Lal, S., Tiwari, L., & Ranjan, R. (2020). ScienceDirect ScienceDirect ScienceDirect Analysis and Classification of Crime Tweets and of Verma Crime a Classification a Tweets Sangeeta Lal Analysis. *Procedia Computer Science*, *167*(2019), 1911–1919. https://doi.org/10.1016/j.procs.2020.03.211

Tollenaar, N., & Van Der Heijden, P. G. M. (2019). Optimizing predictive performance of criminal recidivism models using registration data with binary and survival outcomes. In *PLoS ONE* (Vol. 14, Issue 3). https://doi.org/10.1371/journal.pone.0213245

Van der Aalst, W. (2016). Process mining: Data science in action. *Process Mining: Data Science in Action*, *April 2014*, 1–467. https://doi.org/10.1007/978-3-662-49851-4

Vijayalakshmi, C. (2020). ScienceDirect ScienceDirect Design and Analysis of Machine Learning Algorithms for the Design and Analysis of Machine Learning Algorithms for the reduction of crime rates in India reduction The 9 th World Engineering Education Forum (WEEF - 2019). *Procedia Computer Science*,

172, 122–127. https://doi.org/10.1016/j.procs.2020.05.018

Wang, H., & Ma, S. (2021). Socio-Economic Planning Sciences Preventing crimes against public health with artificial intelligence and machine learning capabilities. *Socio-Economic Planning Sciences*, *September 2020*, 101043. https://doi.org/10.1016/j.seps.2021.101043

Watts, D., Moulden, H., Mamak, M., Upfold, C., Chaimowitz, G., & Kapczinski, F. (2021). Predicting offenses among individuals with psychiatric disorders - A machine learning approach. *Journal of Psychiatric Research*, *138*(October 2020), 146–154. https://doi.org/10.1016/j.jpsychires.2021.03.026

Wirth, R., & Jochen, H. (2000). *CRISP-DM : Towards a Standard Process Model for Data Mining*. *24959*.

Xia, Z., Stewart, K., & Fan, J. (2021). *Computers , Environment and Urban Systems Incorporating space and time into random forest models for analyzing geospatial patterns of drug-related crime incidents in a major U . S . metropolitan area*. *87*(July 2020).