Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Ciencias Matemáticas
Escuela Profesional de Estadística

Comparación de modelos de clasificación para determinar las variables que intervienen en la población penitenciaria que ha cometido el delito de robo agravado

TRABAJO DE SUFICIENCIA PROFESIONAL

Para optar el Título Profesional de Licenciado en Estadística

AUTOR

Carmen Johana HUARANGA VILCAS

ASESOR

Dr. Helfer Joel MOLINA QUIÑONES

Lima, Perú

2021

Referencia bibliográfica

Huaranga, C. (2021). *Comparación de modelos de clasificación para determinar las variables que intervienen en la población penitenciaria que ha cometido el delito de robo agravado*. [Trabajo de Suficiencia Profesional de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Matemáticas, Escuela Profesional de Estadística]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	Carmen Johana Huaranga Vilcas
Tipo de documento de identidad	DNI
Número de documento de identidad	73183672
URL de ORCID	https://orcid.org/0000-0002-2024-5182
Datos de asesor	
Nombres y apellidos	Helfer Joel Molina Quiñones
Tipo de documento de identidad	DNI
Número de documento de identidad	40014631
URL de ORCID	https://orcid.org/0000-0003-1307-7253
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Oscar Antonio Robles Villanueva
Tipo de documento	DNI
Número de documento de identidad	32762171
Miembro del jurado 1	
Nombres y apellidos	Ricardo Luis Pomalaya Verastegui
Tipo de documento	DNI
Número de documento de identidad	10460674
Datos de investigación	
Línea de investigación	A.3.2.6. Análisis de Datos y Modelamiento de Problemas de la Sociedad (Empresa, Instituciones, Poblaciones locales, regionales y nacionales).

Grupo de investigación	No aplica.
Agencia de financiamiento	Sin financiamiento.
Ubicación geográfica de la investigación	Universidad Nacional Mayor de San Marcos País: Perú Departamento: Lima Provincia: Lima Distrito: Lima Coordenadas geográficas Latitud: -12.058333 Longitud: -77.083333
Año o rango de años en que se realizó la investigación	Marzo 2021 – Junio 2021
URL de disciplinas OCDE	Estadísticas, Probabilidad https://purl.org/pe-repo/ocde/ford#1.01.03

ACTA DE SUSTENTACIÓN DE TRABAJO DE SUFICIENCIA PROFESIONAL EN LA MODALIDAD VIRTUAL PARA OBTENCIÓN DEL TÍTULO PROFESIONAL DE LICENCIADA EN ESTADÍSTICA

En Lima, siendo las 14:30 horas del domingo 03 de octubre del 2021, se reunieron los docentes designados como Miembros del Jurado del Trabajo de Suficiencia Profesional (PROGRAMA DE TITULACIÓN PROFESIONAL 2021-I): Dr. Oscar Antonio Robles Villanueva (PRESIDENTE), Mg. Ricardo Luis Pomalaya Verastegui (MIEMBRO) y el Dr. Helfer Joel Molina Quiñones (MIEMBRO ASESOR), para la sustentación del Trabajo de Suficiencia Profesional titulado: "**COMPARACIÓN DE MODELOS DE CLASIFICACIÓN PARA DETERMINAR LAS VARIABLES QUE INTERVIENEN EN LA POBLACIÓN PENITENCIARIA QUE HA COMETIDO EL DELITO DE ROBO AGRAVADO",** presentado por la señorita **Bachiller Carmen Johana Huaranga Vilcas**, para optar el Título Profesional de Licenciada en Estadística.

Luego de la exposición del trabajo de suficiencia, el Presidente invitó a la expositora a dar respuesta a las preguntas formuladas.

Realizada la evaluación correspondiente por los miembros del Jurado Evaluador, la expositora mereció la aprobación de **BUENO,** con un calificativo promedio de **DIECISEIS (16)**.

A continuación, los miembros del Jurado dan manifiesto que la participante **Bachiller Carmen Johana Huaranga Vilcas** en vista de haber aprobado la sustentación del Trabajo de Suficiencia Profesional, será propuesta para que se le otorgue el Título Profesional de Licenciada en Estadística.

Siendo las 15:00 horas se levantó la sesión firmando para constancia la presente Acta.

Dr. Oscar Antonio Robles Villanueva
PRESIDENTE

Mg. Ricardo Luis Pomalaya Verastegui
MIEMBRO

Dr. Helfer Joel Molina Quiñones
MIEMBRO ASESOR

La Vicedecana de la Facultad de Ciencias Matemáticas, Mg. Zoraida Judith Huamán Gutiérrez, certifica virtualmente la participación del Jurado Evaluador, el titulando, el acto de instalación y el inicio, desarrollo y término del acto académico de sustentación, dejando constancia en el acta respectiva.

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
Universidad del Perú. Decana de América
FACULTAD DE CIENCIAS MATEMÁTICAS
ESCUELA PROFESIONAL DE ESTADÍSTICA

INFORME DE EVALUACIÓN DE ORIGINALIDAD

El Director de la Escuela Profesional de Estadística, Dr. Roger Pedro Norabuena Figueroa, informa lo siguiente:

1. Operador del programa informático de similitudes: Dr. Roger Pedro Norabuena Figueroa
2. Documento evaluado: Tesis para optar el Título Profesional de Licenciada en Estadística, titulado: "Comparación de modelos de clasificación para determinar las variables que intervienen en la población penitenciaria que ha cometido el delito de robo agravado"
3. Autor de la tesis: Carmen Johana Huaranga Vilcas
4. Fecha de recepción de la tesis: 19/07/2023
5. Fecha de aplicación del programa informático de similitudes: 19/07/2023
   - Software utilizado: Turnitin
6. Configuración del programa detector de similitudes:
   - Excluye textos entrecomillados
   - Excluye bibliografía
   - Excluye cadenas menores a 40 palabras
7. Porcentaje de similitudes según programa detector de similitudes: cinco por ciento (5%)
8. Fuentes originales de las similitudes encontradas:
   Fuentes de internet: 5 %
   Publicaciones: 1 %
9. Calificación de originalidad:
   - Documento cumple criterios de originalidad, sin observaciones

Lima, 20 de julio del 2023

UNMSM
Firmado digitalmente por NORABUENA FIGUEROA Roger Pedro FAU 20148092282 soft
Motivo: Soy el autor del documento
Fecha: 20.07.2023 15:56:39 -05:00

Dr. Roger Pedro Norabuena Figueroa
Director

RESUMEN

En los últimos años, nuestro país viene enfrentándose a la delincuencia, convirtiéndose así en uno de los problemas que más preocupa a los ciudadanos como a las autoridades, siendo el robo agravado el delito con mayor número de internos en los centros penitenciarios del Perú. Por ende, el objetivo de este trabajo fue determinar las variables que intervinieron que un interno haya cometido delito de robo agravado, aplicando regresión logística y árboles de clasificación. La base de datos a utilizar fue extraída del portal del INEI; y corresponde al Primer Censo Nacional Penitenciario 2016, se delimito como ámbito de estudio, a todos los internos peruanos que cometieron algún delito en Lima Metropolitana contando con 23 mil 735 personas que se encuentran privadas de su libertad, considerando 23 mil 491 internos después del tratamiento de datos. De los cuales el 34.7% cometieron el delito de robo agravado y 94.1% son varones, Se determinó como mejor modelo a la regresión logística, con un 75% de sensibilidad, identificando a las siguientes variables: antes de cumplir los 18 años, sus mejores amigos cometían delitos; el barrio donde vivía había pandillas y si algún familiar se encontraba preso se asocian significativamente con que el interno haya cometido el delito de robo agravado.

Palabras clave: Robo agravado, regresión logística, árboles de clasificación.

ABSTRACT

In recent years, our country has been facing crime, thus becoming one of the problems that most worries citizens and authorities, with aggravated robbery being the crime with the highest number of inmates in prisons in Peru. Therefore, the objective of this work was to determine the variables that intervened that an inmate had committed the crime of aggravated robbery, applying logistic regression and classification trees. The database to be used was extracted from the INEI portal; and corresponds to the First National Penitentiary Census 2016, it was defined as the scope of study, all Peruvian inmates who committed a crime in Metropolitan Lima, with 23 thousand 735 people who are deprived of their liberty, considering 23 thousand 491 inmates after treatment of data. Of which 34.7% committed the crime of aggravated robbery and 94.1% are men, the best model was determined to be logistic regression, with 75% sensitivity, identifying the following variables: before reaching 18 years of age, their best friends committed crimes; There were gangs in the neighborhood where he lived, and if a relative was in prison, they are significantly associated with the inmate having committed the crime of aggravated robbery.

Keywords: Aggravated robbery, logistic regression, classification trees.

ÍNDICE DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

I. INTRODUCCIÓN

La empresa consultora desarrolla estudios de tipo cuantitativo, para encontrar soluciones que ayuden en la toma de decisiones de sus clientes, de los rubros de banca, retail, educación, inmobiliaria, seguros y gobierno. Actualmente me encuentro trabajando en el puesto de analista estadístico durante dos años en el área de Analitys.

Nuestro país viene enfrentándose a la delincuencia a lo largo de la historia, convirtiéndose en uno de los problemas que más preocupa tanto a los ciudadanos como a las autoridades, la inseguridad ciudadana constituye un obstáculo para el desarrollo humano, no solo resta calidad de vida y salud mental, sino produce altos gastos del Estado en seguridad y conservación de los centros penitenciarios, siendo de interés para las autoridades municipales tener conocimiento sobre los perfiles asociados de la población que comete delitos, para poder implementar programas sociales enfocados en la prevención de delitos.

El robo agravado es el delito con mayor número de internos (29.5%) en el Perú. Se encuentra dentro de los delitos contra el patrimonio, es el acto de apoderarse de un bien total o parcialmente ajeno, empleando violencia que ponga en riesgo la vida de la otra persona.

El objetivo de este trabajo fue determinar las variables que intervinieron para que los internos cometieran el delito de robo agravado aplicando modelos de clasificación. A partir del Primer Censo Nacional Penitenciario 2016, base de datos recolectada por el Instituto Nacional de Estadística e informática (INEI); se comparan los resultados obtenidos de la regresión logística y árboles de clasificación, mediante la evaluación de la sensibilidad y el AUC calculado para cada caso. Encontrándose en la línea de investigación de análisis de datos y modelamiento de problemas de la sociedad.

II. DESCRIPCION DE LA ACTIVIDAD

2.1 Datos de la empresa

La empresa de origen chileno con más de veinte años de experiencia que desarrolla consultoría para encontrar soluciones de localización, marketing de base de datos, estudios de mercado en distintos rubros, tales como: banca, consumos de venta, retail, educación, inmobiliario, seguros y gobierno; ofreciendo herramientas para mejorar en la toma de decisiones gubernamentales, ayudándoles a identificar oportunidades.

La empresa busca ser líder en generar información confiable, mediante el análisis de datos busca ofrecer herramientas tecnológicas que ayuden a los clientes a minimizar los riesgos y aumente su progreso.

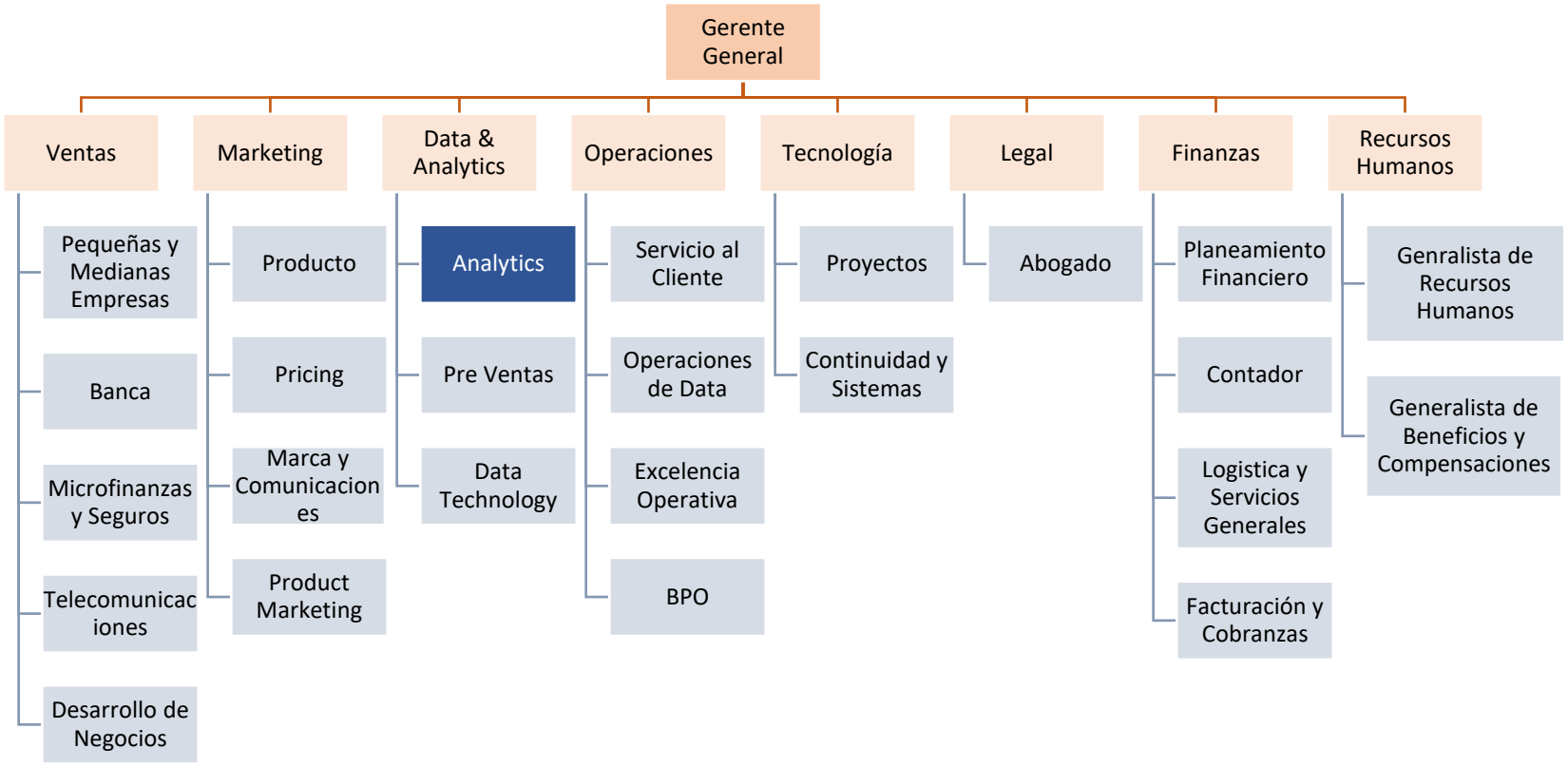
Su misión es tener un buen gobierno a nivel corporativo porque le permitirá evaluar las metas de desempeño de sus colaboradores y todos los procesos funcionen de manera eficiente, para enfrentar los desafíos de manera oportuna ante la competencia.

Este trabajo fue realizado durante el periodo de marzo a junio del 2021, para el cumplimiento de los objetivos de medio año del presente año, la propuesta fue desarrollada en el área de Analytics de la empresa que se encuentra ubicada en el distrito de San Isidro.

2.1.1 Organigrama de la empresa

Figura 1

Organigrama de la empresa



Nota. Área de Recursos Humanos de la Empresa

2.2 Descripción de la actividad

Actualmente me encuentro trabajando en el en el área de Analitycs; cuya finalidad es generar estudios y automatización de procesos, para resolver soluciones que faciliten la toma de decisiones minimizando el riesgo y maximizando oportunidades de crecimiento de los clientes.

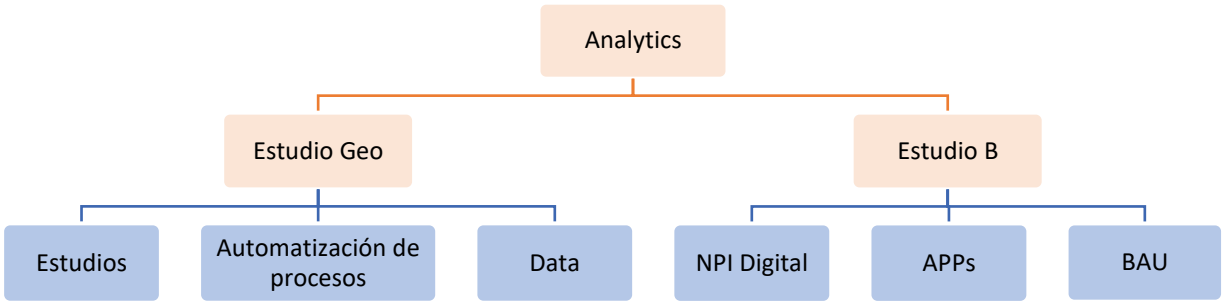
Desde hace dos años ocupó el puesto de analista estadístico y me encargó de analizar datos provenientes de entidades públicas o privadas, utilizando herramientas estadísticas o de minería de datos para encontrar soluciones que ayuden a la toma de decisiones; tales como perfilamiento, estimación de ventas, localización de tiendas, etc. Estos estudios tienen dos formas de ser entregados al cliente (empresa o persona de institución pública o privada), el primero mediante un informe y el segundo mediante una interfaz interactiva con el usuario. También realizo proyecciones de poblaciones y hogares a nivel manzanas mediante la técnica de desagregación de las proyecciones del departamento, para mantener actualizados los insumos de las plataformas que vende la empresa.

El sub- área al que pertenezco se llama Estudios tiene como finalidad encontrar soluciones innovadoras de localización, perfilamiento, marketing de base de datos.

2.2.1 Organigrama del área

Figura 2

Organigrama del área



Nota. Área de Recursos Humanos de la Empresa

2.2.2 Finalidad

La finalidad de este trabajo fue proponer a la empresa el desarrollo de un nuevo producto que pueda ofrecer a un sector no muy concurrido, que son clientes de entidades públicas, puesto que uno los objetivos de la empresa es ampliar su alcance de negocio. En la empresa las distintas áreas que trabajan con datos, se plantean objetivos que buscan implementar nuevos productos o generar insumos que ayudaran a generarlas. Esta propuesta sería ofrecida a clientes como funcionarios de municipalidades, los cuales deben definir su plan de acción distrital de seguridad ciudadana y este producto podría ser útil en la toma de sus decisiones.

2.2.3 Objetivos

2.2.3.1 Objetivo General

Determinar las variables que intervinieron para que la población penitenciaria peruana cometa el delito de robo agravado en Lima Metropolitana para el 2016.

2.2.3.2 Objetivos específicos

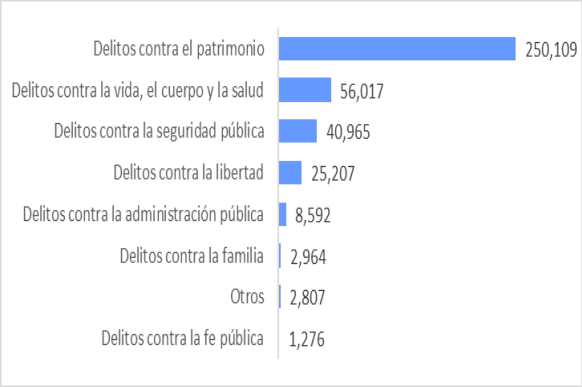
- Consolidar la información de la población penitenciaria peruana que cometió delitos en Lima Metropolitana para el 2016.
- Comparar los modelos de clasificación para encontrar el mejor modelo que discrimine a los internos peruanos que cometieron el delito de robo agravado.
- Identificar las variables más importantes que intervinieron para que la población penitenciaria peruana cometa el delito de robo agravado en Lima Metropolitana para el 2016.

2.2.4 Problemática

La delincuencia es un fenómeno social que está presente en todo el mundo y genera inseguridad en las personas, ya sea que vivan en las zonas urbanas o rurales, es un problema que está presente en todos los niveles socioeconómicos de nuestra sociedad. En nuestro país los delitos contra el patrimonio son los que más preocupa a las entidades públicas y privadas; porque es el delito con mayor número de denuncias año tras año, como se observa en la Figura 4.

Figura 3

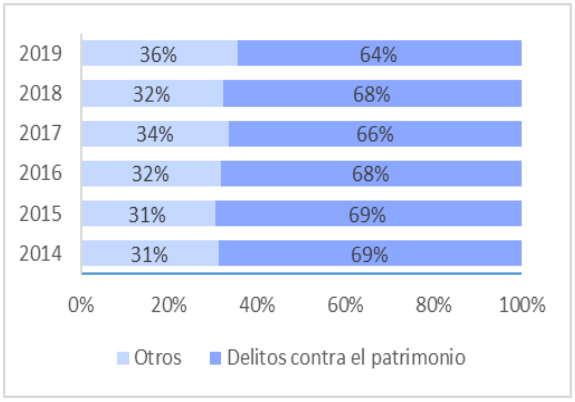
Cantidad de denuncias registras según la clasificación de delitos genéricos, durante el periodo del 2019.



Nota. Elaboración propia con datos de MININTER

Figura 4

Porcentaje de denuncias registrados de delitos contra el patrimonio durante el 2014 al 2019.



Nota. Elaboración propia con datos de MININTER

En la Figura 3 se puede observar que, en el 2019, los delitos contra el patrimonio representan un 64%, seguido de los delitos contra vida, el cuerpo y la salud. A parte de la inseguridad ciudadana que generan estos delitos, los delitos contra el patrimonio tienen mayor número de internos en los centros penitenciarios. Según la Defensoria del Pueblo (2020), existe un 140% sobrepoblación en los centros penitenciarios, siendo el robo agravado el delito con mayor número de internos. El 25.6% de la población privada de su libertad se debe a este delito específico. (Instituto Nacional Penitenciario, 2019)

2.3 Breve descripción de la metodología

2.3.1 Instrumento

El presente estudio obtuvo los datos a partir de una fuente secundaria, la cual corresponde al primer censo penitenciario ejecutado por el INEI, ejecutado durante el 18 al 24 de abril del 2016. Que consta de cinco secciones, pero para este trabajo solo se utilizamos las secciones que menciono a continuación:

- La primera sección recolecta información de datos personales del interno tales como, edad, genero, religión, estado civil, etc.
- La segunda parte son preguntas relacionadas a las condiciones sociales y familiares del interno.
- La tercera parte son preguntas relacionadas a la tipificación del delito, tales como: si perteneció a algún centro juvenil o si uso un arma durante el delito.

2.3.2 Herramientas

La herramienta que se utilizó para este trabajo fueron los softwares Hojas de cálculo desde Microsoft Excel y RStudio versión 4.1.0; los principales paquetes a utilizados son foreign para la lectura de los datos desde un formato (.sav), funModeling, VIM para el pre procesamiento de los datos y rpart se utilizó la estimación del modelo de árbol de clasificación aplicando el algoritmo CART, entre otros como pROC, dplyr, fastDummies, caret y readxl

2.3.3 Procedimientos

Este trabajo se realizó según la metodología CRIPS, como se detalla a continuación:

1) Compresión del negocio: Se valida la situación donde se determinó los objetivos de interés de la empresa y se realiza el plan de trabajo del proyecto con los objetivos analíticos alineados al negocio.
2) Comprensión de los datos: Se consolidó las bases de datos proveniente del primer censo penitenciario 2016, se realiza un análisis exploratorio de los datos, que nos permitirá delimitar el ámbito de estudio.
3) Preparación, procesamiento, limpieza y tratamiento de los datos: Se evalúa la calidad de los datos (detectando valores nulos), se realiza la limpieza de los datos, se recodifica o categoriza las variables, se selecciona las variables para la estimación del modelo.
4) Modelamiento: Se construyen los modelos que serán utilizados en el trabajo, como siguiente paso se evalúan según las pruebas de significancia o las métricas establecidas por la matriz de confusión.
5) Evaluación: Se compara los modelos mediante la sensibilidad y área bajo la curva obtenido los modelos.
6) Implementación: Se proporcionó la estructura cómo se presentaría los resultados en una interfaz de sus productos de la empresa y la realización de un informe final de la demo.

III. MARCO TEÓRICO

3.1. Minería de datos

Hand et al.(2001), menciona que la minería de datos se utiliza principalmente para analizar grandes cantidades de datos, de esta manera encuentra patrones, correlaciones y resume los datos obtenidos de la realidad de formas que sean entendible y útil para las personas que utilizaran esta información para la toma de decisiones. Por otro lado, Weiss y Davison (2010) expresa que es el proceso de tomar un conjunto de datos de entrada y convertirlos en conocimiento, para encontrar patrones de comportamiento útiles para poder tomar decisiones que solucionen problemas.

De acuerdo Kantardzic (2011) existen dos conceptos errados sobre la minería de datos: 1) Una gran cantidad de datos se separa entre sí esperando que ocurra el problema. 2) El problema involucra un solo método o algoritmo. Por el contrario, la minería de datos se centra en la integración de datos para comprender mejor el entorno en el que se produce el problema. También se considera un proceso iterativo porque puede utilizar diferentes técnicas hasta encontrar la que mejor se adapte a su modelo.

- Método de aprendizaje no supervisado: Son aquellas en las que no se dispone de una variable dependiente para realizar el estudio, el objetivo es explorar para encontrar alguna forma de organizarlos.
- Método de Aprendizaje supervisado: En este caso se tiene la variable dependiente o de estudio ya definida, cuyo objeto es realizar predicciones y clasificaciones. Aquí están los modelos utilizados en este trabajo árboles de decisión y la regresión logística.

3.1.1 Árbol de decisión:

El árbol de decisión es un modelo muy utilizado dentro de la minería de datos porque permite conocer las características de los usuarios respecto a la variable de estudio, se encuentra dentro de los métodos de aprendizaje supervidado. Es un modelo que busca subdividir el espacio en dos regiones lo más homogéneas posibles, dado que si una de las regiones presenta observaciones de diferentes clases; busca particionar en subregiones más chicas manteniendo el criterio anterior, hasta que en cada subregión solo se encuentren elementos de una clase; también se le conoce como árbol completo (Gironés et al., 2017)

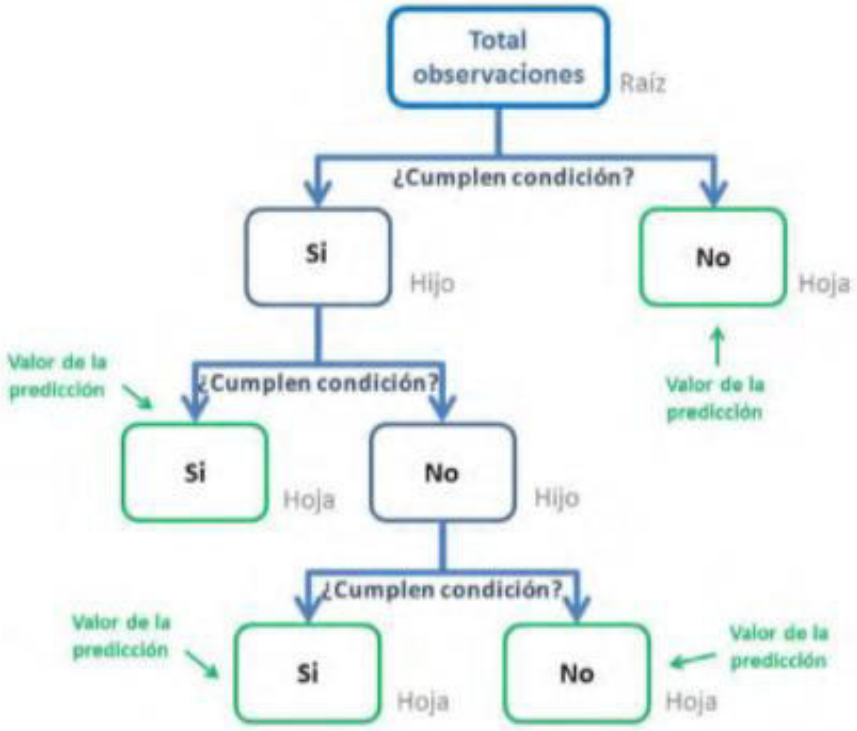
Son utilizados cuando la variable de estudio es cuantitativa o cualitativa, siendo llamados como arboles de regresión o de clasificación respectivamente. Ambos tipos de árboles de decisión se les conoce como CART (*Classification and Regression Tress*) según el trabajo original de (Breiman et al., 1984)

3.1.1.1 Estructura de un árbol de clasificación

Según Gironés et al. (2017) plante que un árbol de clasificación contiene nodos y hojas como se detalla a continuación: El *nodo raíz o padre*, contiene la condición por la cual se realizará la partición, es decir, definirá a que región pertenece el nuevo dato; hasta llegar a la siguiente condición en el *nodo hijo o intermedio*. Dicho nodo representa las condiciones que definirán a que clase pertenece cada elemento generando dos o más segmentos, los cuales volverán a ser divididos en las hojas terminales, donde el nodo ya no podrá ser dividido en más subregiones, la cual determina a que clase corresponde el nuevo dato. En la figura 5 podemos observar la estructura de un árbol de decisión con mayor claridad.

Figura 5

Árbol de clasificación



Nota. Representa la estructura de un árbol de clasificación. Tomado de Chavez, 2018, https://rpubs.com/elfenixsoy/arbol-veronica

3.1.1.2 Construcción del algoritmo de un árbol de clasificación

Para la construcción de un árbol de clasificación, el algoritmo general considera en primer lugar el *criterio de parada* el cual permite determinar cuándo parar de seleccionar nodos que serán subdivididos; después está el *criterio de selección*, este determinará qué nodo será elegido para ser dividido en dos o más subnodos; pero si el investigador busca construir un árbol completo este criterio no sería relevante; el siguiente es el *criterio de clasificación*, el cual definirá la clase que representará el nodo intermedio logrando minimizar los errores de

clasificación. El ultimo es el *criterio de partición*, el cual determina como se seguirá dividiendo un nodo en una o más subregiones. (Gironés et al., 2017)

Hay muchos algoritmos que permiten la construcción de árboles decisión, siendo los algoritmos más usados: El algoritmo C4.5 fue propuesto por Quinlan, basado en el algoritmo ID3. Este método maneja datos faltantes de manera eficiente, al mismo tiempo introdujo los métodos de poda basado en los errores y maneja fácilmente datos continuos y el algoritmo CART fue introducido por Breiman y otros en 1984, se pueden construir arboles de clasificación y regresión con este algoritmo, una de sus ventajas es que maneja atributos discretos y continuos, este algoritmo divide de forma binaria las regiones buscando que sean los más homogéneas posibles, considerando el índice de Gini. (Przemyslaw y Krzysztof, 2019, p.126)

Se podría decir que dado un conjunto de datos D= (X, Y), donde Y es la variable a explicar y $X = (X_1, \dots, X_k)$ es un vector de k variables explicativas, donde el CART busca predecir las variable dependiente a partir de los datos de las variables explicativas X.

Según Gironés et al. (2017), en el nodo t encontraremos n_t registros, por lo tando podemos definir a p_i como:

$$p_i(t) = \frac{n_i(t)}{n_t}$$

Donde p_i es la probabilidad que un elemento pertenezca a la clase i en el nodo t o como la proporción del número de elementos de la clase i presentes en el nodo t entre el total de elementos en este nodo n_t

3.1.1.3 Medidas de impureza

Las siguientes medidas de impureza $I_t(T)$ serán utilizadas en el árbol de clasificación para realizar la partición de las regiones lo más homogéneas posibles:

a) El índice de Gini:

$$I_t(T) = \sum_{k=1}^{k} p_i(t)(1 - p_i(t))$$

Este índice de gini mide la impureza del nodo; es decir, cuán desordenados o mezclados quedan los nodos una vez divididos. Considera una división binaria, para determinar la mejor división examina todos los posibles subconjuntos que puedan formarse. (Han et al., 2012)

b) La Entropía Cruzada:

$$I_t(T) = \sum_{i=1}^{k} p_i log p_i$$

"La entropía cruzada mide el grado de desorden de una distribución de elementos de k clases diferentes y es cero si todos los elementos son de una misma clase". (Gironés et al., 2017, p. 217),

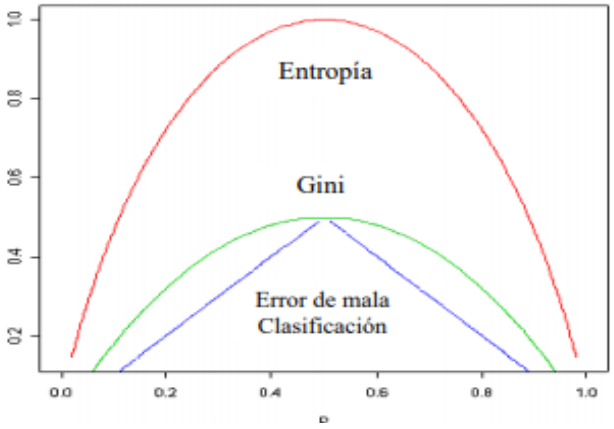
c) Error de mala clasificación

$$I_t(T) = \frac{1}{n_t} \sum_{t \in T} I(y_i \neq k) = 1 - p_i$$

Este error se basa en la proporción con la clase mayoritaria que presenta el nodo que se está evaluando.

Figura 6

Representación gráfica de las medidas de impureza



Nota. Representación gráfica de las medidas de impureza. Fuente: Vigo, 2010.

El árbol obtenido en el modelo inicial, también conocido como árbol completo; mayormente presenta un sobreajuste, entonces se buscará podarlo con la finalidad de encontrar el tamaño óptimo. Entonces, después de obtener el árbol completo T_0 se aplica un algoritmo de poda brinda una secuencia de subárboles, para después ser elegido el subárbol $T \subset T_0$ con menor tasa de error. (Gironés et al., 2017)

3.1.1.4 Discretización mediante arboles de decisión.

Otro uso que tienen los arboles de decisión es poder realizar discretizar las variables de tipo numérica para los métodos supervisados. Según Han et al. (2012)

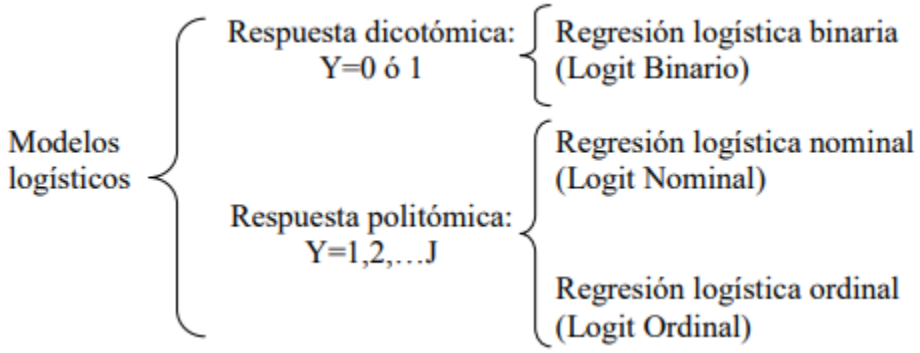
> Intuitivamente, la idea principal es seleccionar puntos de división para que una partición resultante contenga tantas tuplas de la misma clase como sea posible. La entropía es la medida más utilizada para este propósito. Para discretizar un atributo numérico, el método selecciona el valor que tiene la entropía mínima como punto de división y divide de forma recursiva los intervalos resultantes para llegar a una discretización jerárquica. Tal discretización forma una jerarquía de conceptos. (p.116)

3.1.2 Regresión Logística

Según Joseph Hilbe (2009), la regresión logística nos permite modelar cuando la variable dependiente es de tipo cualitativa a partir en un conjunto de variables explicativas que pueden ser cuantitativas o cualitativas.

Figura 7

Tipos de modelos logísticos



Nota. Tipos de modelos logísticos. Fuente: Aquino (2019)

3.1.2.1 Regresión Logística Binaria

Sea la variable respuesta Y una variable de tipo cualitativa que tomar dos posibles valores, 1 o 0 (Y=1: si ocurre el suceso; Y=0: no ocurre el suceso). Además, sea $X = (X_1, X_2, \dots, X_k)$ un conjunto de k variables explicativas; se busca determinar la probabilidad que una de las observaciones pertenezca a uno de los grupos, la cual es denotada por π_i, donde π_i es la probabilidad de éxito y $1 - \pi_i$ cuando no ocurre el suceso, según Hosmer et al. (2013). Entonces el modelo logístico está representado por:

$$\pi = P(Y = 1/X) = \left(\frac{\exp(\eta)}{1 + \exp(\eta)}\right)$$

Indica la probabilidad condicional de que un evento cuando Y=1 ocurra dado la ocurrencia de un conjunto X, de k variables explicativas. Siendo $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$ el

predictor lineal una combinación lineal de las variables explicativas, donde β_0 es una constante y los β_i son los coeficientes de las variables independientes x_i del modelo. Aplicando la transformación logit al modelo logístico se obtiene. (Joseph Hilbe, 2009)

$$logit(\pi) = ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

Dicha transformación permite linealizar la relación entre $P(Y = 1/X)$ y el predictor lineal η.

3.1.2.2 Estimación de parámetros

Para estimar los parámetros se emplea el método de máxima verosimilitud con lo cual se encuentra el valor de β que maximice la probabilidad de obtener el conjunto observado de datos. Se sabe que cada observación de la muestra y_i son variables aleatorias independientes de Bernoulli, entonces la distribución de probabilidad es definida por. (Hosmer et al., 2013).

$$f_i(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i};\ i = 1,2,\cdots,n$$

Entonces la función de verosimilitud es:

$$\ln\big(L(\beta)\big) = \sum_{i=1}^{n}\big[y_i X_i'\beta - ln\big(1 + exp(X_i'\beta)\big)\big]$$

Para encontrar el valor de β que maximiza $L(\beta)$ se deriva respecto a β_0 y β_j e igualar a cero las expresiones, entonces se tendrá $k+1$ ecuaciones de verosimilitud:

Derivando respecto a β_0:

$$\sum_{i=1}^{n}[y_i - \pi_i] = 0$$

Derivando respecto a β_j:

$$\sum_{i=1}^{n} xij[y_i - \pi_i] = 0\ , para\ j = 1,2,\ldots,k$$

Como las $k+1$ ecuaciones son no lineales en los parámetros β, se requieren de métodos iterativos para estimar los parámetros.

La estimación de varianzas de los parametros estimados se calcula a partir de la segunda derivada parcial del logaritmo de la función de máxima verosimilitud. (Hosmer et al., 2013).

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i (1-\pi_i)$$

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi_i (1-\pi_i), para\ j = 0,1,2, \dots, k$$

Se denota la matriz de información observada $I(\beta)$ que contiene los términos negativos de las ecuaciones anteriores; su expresión es matricial es:

$$I(\beta) = X'_{(k+1)\times n} \Gamma_{n\times n} X_{n\times(k+1)}$$

Las varianzas y covarianzas de los parámetros estimados se obtienen de la inversa de esta matriz, que se denota como $V(\beta) = I^{-1}(\beta)$.

3.1.2.3 Pruebas de Hipótesis

- **Prueba de significancia global del modelo**

Para evaluar la significancia global del modelo, se considera las siguientes hipótesis:

$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$

$H_1: \beta_i \neq 0$, para al menos un $i = 1, \dots, k$

El estadístico de prueba es la diferencia entre la desvianza del modelo nulo (modelo sin variables solo con la constante) y la desvianza del modelo ajustado:

$G = D(modelo\ nulo) - D(modelo\ ajustado)$

$$G = -2\ln\left(\frac{L(modelo\ nulo)}{L(modelo\ ajustado)}\right)$$

Donde la verosimilitud del modelo nulo es:

$$L(modelo\ nulo) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i},\ \ donde\ \pi_i = \frac{exp(\beta_0)}{1+exp(\beta_0)}$$

Donde la verosimilitud del modelo saturado es:

$$L(modelo\ ajustado) = \prod_{i=1}^{n} \hat{\pi}_i^{y_i}(1-\hat{\pi}_i)^{1-y_i},\ \ donde\ \hat{\pi}_i = \frac{exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k)}{1 + exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k)}$$

Entonces el estadístico G queda definido como:

$$G = -2\left\{n_1 \ln(n_1) + n_0 \ln(n_0) - n\ln(n) - \sum_{i=1}^{n} y_i \ln(\hat{\pi}_i) + (1-y_i)\right\} \sim \chi_k^2$$

Donde: $n_1 = \sum_{i=1}^{n} y_i \qquad y \qquad n_0 = \sum_{i=1}^{n}(1-y_i)$

El estadístico G presenta un distribución chi cuadrado con k grados de libertad. Se rechaza la H_0 si $G > c$; donde c es el cuantil $\chi^2_{(k,\propto)}$, el cual nos indica que al menos una de las variables predictoras contribuye significativamente al modelo. (Hosmer et al., 2013).

- **Prueba de significancia de los parámetros individuales del modelo**

Se evaluará la significancia de los coeficientes del modelo. Bajo la siguiente hipótesis:

$H_0: \beta_j = 0, \qquad para\ j = 0, \ldots, k$

$H_1: \beta_j \neq 0$

Se emplea la estadística de Wald, que tiene distribución chi cuadrado 1 grado de libertad.

$$W = \left(\frac{\hat{\beta}_j}{se(\hat{\beta}_j)}\right)^2 \sim \chi^2_{(1)}$$

Se rechaza la H_0 si $W > \chi^2_{(1,\alpha)}$, concluyendo que el parámetro es significativo.

- **Estimación por intervalos de los parámetros.**

Puesto que los parámetros β_j, siguen asintóticamente una distribución $N(\beta_j, se(\hat{\beta}_j))$ con lo que:

$$P\left(-Z_{\frac{\alpha}{2}} \leq \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \leq Z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

En base a la estadística de Wald, se obtiene el intervalo de confianza al $(1-\alpha)\%$ para el coeficiente β_j es:

$$IC(\beta_j) \in < \hat{\beta}_j \pm Z_{\frac{\alpha}{2}} se(\hat{\beta}_j) >$$

- **Interpretación de los parámetros.**

También conocido como razón de ventajas, donde podemos estimarlo aplicando el antilogaritmo al modelo de regresión logística, quedando definido de la siguiente manera:

$$\pi = \left(\frac{\exp(\eta)}{1+\exp(\eta)}\right) = exp(X_i'\beta) = OR$$

Si la razón de ventaja es igual a uno, indica que no existe asociación entre el factor X_i y la variable dependiente Y. Por otro lado, si es mayor a uno, entonces es más factor de riesgo y si el odds ratio es menor que uno, hay menos chance del resultado de Y dado el factor protector.

3.1.3 Evaluación de modelos

Para la evaluación de los modelos de clasificación, se dividirá los datos en una muestra de entrenamiento, donde se construyen los modelos y una muestra de validación donde se evaluará que tan bueno o preciso es el clasificador donde se utilizaran la matriz de confusión y el área bajo la curva AUC.

3.1.3.1. Matriz de confusión

La matriz de confusión es una tabla de doble entrada que nos permite contrastar los datos reales con los resultados obtenidos a partir de las predicciones. En la tabla 1 se presenta esta herramienta.

Tabla 1

Matriz de Confusión

Predicción	**Realidad**	
	Positivo	Negativo
Positivo	TP	FP
Negativo	FN	TN

Nota. Matriz de confusión para dos clases, adaptado de Han et al. (2012)

Según Han et al. (2012), detalla la tabla 1 de la siguiente manera: 1) Verdaderos positivos (TP), a las tuplas positivas que fueron clasificadas correctamente por el modelo. 2)Verdaderos negativos (TN), estas son las tuplas negativas que fueron clasificadas correctamente. 3) Falsos positivos (FP), representan las tuplas negativas que fueron clasificadas incorrectamente como positivas.4) Falsos negativos (FN), están son las tuplas positivas que fueron clasificadas erróneamente como negativas.

A continuación, se presenta las medidas de evaluación que serán utilizadas en este trabajo

- Sensibilidad: Es la proporción de tuplas positivas que se identifican correctamente respecto a la clase positiva de los datos reales.$\frac{\text{TP}}{\text{TP+FN}}$
- Especificidad: Es la proporción de tuplas negativas que se identifican correctamente respecto a la clase negativa de los datos reales.$\frac{\text{TN}}{\text{FP+TN}}$

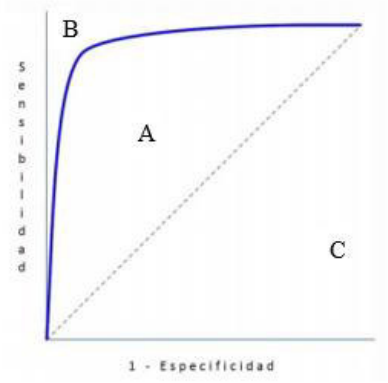
3.1.3.2. Curvas ROC

> "La Curva ROC (receiver operating characteristic curves) indica que cuanto más alejada este de la diagonal principal mejor es el método diagnóstico, ya que la curva ROC ideal sería la que con una especificidad de 1 tuviera una sensibilidad de 1, y cuanto más cercana este a dicha diagonal peor será el método de diagnóstico. Cabe recordar que la

diagonal principal es la que corresponde al peor test de diagnóstico y que tiene un área bajo de ella de 0.5.” (Lizares, 2017)

Figura 8

Curva de ROC



Nota. La diagonal de la curva ROC se interpreta como un modelo generado aleatoriamente, mientras que los valores inferiores se consideran peores. Fuente: Gironés et al.(2017)

Según Girones et al (2017), la curva de ROC permite determinar la precisión que tiene el modelo para discriminar respecto a la variable dependiente, considerando el área bajo la curva (AUC). Por otro lado, considera que un AUC discrimina bien si es mayor a 0.7.

3.1.4 Chi cuadrado de Pearson (test de independencia)

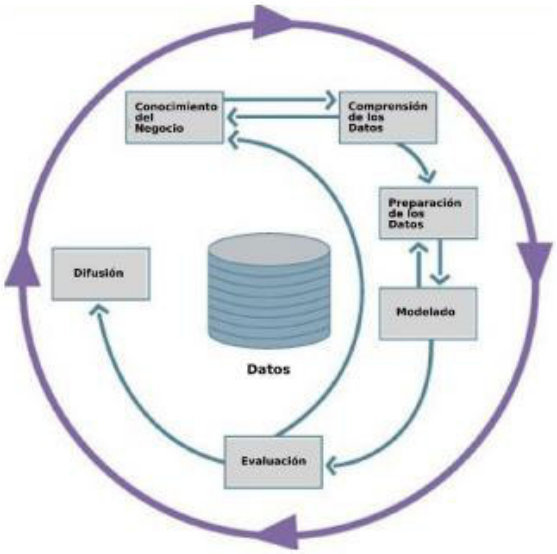
El test $\chi 2$ se emplea para estudiar si existe asociación entre dos variables categóricas, es decir, si las proporciones de una variable son diferentes dependiendo del valor que adquiera la otra variable.

3.1.5. Procesos de la metodología CRIPS-DM

Según Kimball (2002), el término CRISP-DM proviene del acrónimo Cross Industry Standard Process for Data Mining, es un proceso construido para desarrollar proyectos de minería de datos, consta de seis fases como se observa en la figura 9.

Figura 9

Etapas de la metodología CRIPS - DM



Nota. Etapas de la metodología CRIPS-DM. Fuente: Kimball (2002).

Según Kimball (2002), define las siguientes etapas:

1) Comprensión del Negocio: Se determina los objetivos del cliente, se entiende la problemática y se convierte el conocimiento en objetivos técnicos.
2) Comprensión de los Datos: Se procede a recolectar los datos a partir de las fuentes que provienen, luego se inicia la exploración de los datos sin perder de vista el objetivo del cliente y se identifica la calidad de los datos.

3) Preparación de la Data: Luego de haber realizado un análisis exploratorio de los datos, se procede a realizar transformaciones, tratamiento de valores perdidos mediante técnicas de imputación, generación de nuevas variables y selección de variables.
4) Modelamiento: Después de tener la base de datos limpia se aplican las diferentes técnicas de modelado que cumplan los objetivos planteados
5) Evaluación: Se procede a realizar la evaluación de los modelos para identificar el que mejor se ajuste a nuestro objetivo
6) Implementación: En esta etapa se implementa el modelo final que brinde solución al negocio.

3.2 Primer Censo Penitenciario 2016

El presente trabajo obtuvo los datos a partir de una fuente secundaria, la cual corresponde al primer censo penitenciario 2016 ejecutado por el Instituto Nacional de Estadística e Informática (INEI), en coordinación con el Instituto Nacional Penitenciario (INPE) y el Ministerio de Justicia y Derechos Humanos (MINJUS) entre el 18 y 24 de abril. El cuestionario utilizado comprende 5 secciones, utilizándose para este trabajo 3 secciones; datos personales del interno(a), condiciones sociales y familiares del interno(a) y la tipificación del delito.

La base de datos está conformada por 76 180 internos de 66 establecimientos penitenciarios con los que cuenta el Perú en el 2016, los establecimientos penitenciarios están ubicados en los 24 departamentos y la provincia constitucional del Callao, para este trabajo se consideró como criterio de inclusión a la población penitenciaria peruana que cometió algún delito en Lima Metropolitana, los cuales fueron 23 735 personas privadas de su libertad.

En la tabla 2, podemos observar que el 49% de la población penitenciaria peruana cometió delitos contra el patrimonio (delito genérico) en Lima Metropolitana. De los 11 mil 639

de internos que cometió el delito contra el patrimonio, el 71% de los internos cometió el delito específico de robo agravado

Tabla 2

Población penitenciaria peruana que cometió algún delito en Lima Metropolitana según delito genérico. 2016

Delito Genérico	n	%
Delitos Contra El Patrimonio	11,639	49%
Delitos Contra La Seguridad Publica	5,655	24%
Delitos Contra La Libertad	3,489	15%
Delitos Contra La Vida El Cuerpo Y La Salud	1,908	8%
Delitos Contra La Familia	310	1%
Delitos Contra La Tranquilidad Publica	257	1%
Delitos Contra La Administración Publica	237	1%
Delitos Contra El Orden Financiero Y Monetario	98	0%
Delitos Contra La Fe Publica	89	0%
Delitos Contra El Estado Y La Defensa Nacional	18	0%
Lavado De Activos	17	0%
Delitos Contra La Confianza Y La Buena Fe En Los Negocios	6	0%
Delitos Tributarios	5	0%
Delitos Contra La Humanidad	4	0%
Delitos Contra El Orden Económico	3	0%

Nota. Elaboración propia con datos del Primer Censo Penitenciario 2016. INEI

Siendo los delitos contra el patrimonio los más frecuentes en el país y como delito específico el robo agravado, este trabajo se enfocó en estudiar a dichos internos y de esta manera brindar información confiable al cliente para la toma de decisiones.

3.3. Robo agravado

El robo agravado es un delito contra el patrimonio, según Urqizo (2010), se define como el acto de apoderarse ilegalmente, abusar, quitar la propiedad de alguien, usar la violencia contra esa persona o amenazar su integridad física o poner en peligro su vida.

En el artículo 189º del Código de Procedimientos Penales se cuantifica en tres la pena del delito de robo agravado. Entre doce y veinte años de prisión si es cometido en: inmueble habitado, durante la noche o en lugar desolado, a mano armada, con el concurso de dos o más personas, cualquier medio de transporte, fingiendo ser autoridad o servidor público o privado, agravio de personas vulnerables, un vehículo automotor (sus partes o accesorios). 2) Entre veinte y treinta años si el robo es cometido: cuando se cause lesiones a la integridad física o mental de la víctima, con abuso de la incapacidad física o mental de la víctima o mediante el empleo de alguna sustancia, colocando a la víctima o a su familia en grave situación económica y sobre bienes de valor científico o que integren el patrimonio cultural de la Nación. 3) Cadena perpetua cuando el agente actúe en calidad de integrante de una organización criminal, o si, como consecuencia del hecho, se produce la muerte de la víctima o se le causa lesiones graves a su integridad física o mental.

Se define la variable dependiente como un interno que cometió robo agravado en Lima Metropolitana, para la clase positiva y para la clase negativa como un interno que cometió otro delito.

Variable	**Descripción**	**Tipo de variable**	**Categoría**
Cometió robo agravado	Interno que cometió el delito de robo agravado en Lima Metropolitana	Cualitativa Nominal	No, Si

En base a Molocho Luis (2017), Coronado Liz (2016) se consideran como características sociodemográficas de los internos a las siguientes variables:

Variable	**Descripción**	**Tipo de variable**	**Categoría**
Sexo	Sexo del interno	Cualitativa Nominal	Hombre, Mujer

Variable	Descripción	Tipo de variable	Categoría
Estado civil	Estado civil del interno	Cualitativa Nominal	Casado, Conviviente, Viudo, Divorciado, Soltero
Edad	Edad en años del interno	Cuantitativa Razón	
Grado de Instrucción	Último grado de instrucción que aprobó antes de ingresar al establecimiento penitenciario	Cualitativa Ordinal	Inicial, Primaria, Secundaria, Técnico, Universitario, Postgrado
Hijos	Tenencia de hijos del interno	Cualitativa Nominal	No, Si

Valderrama María (2013), Coronado Liz (2016) Molocho Luis (2017) consideraron como características familiares asociadas a los internos algunas de las siguientes variables:

Variable	Descripción	Tipo de variable	Categoría
Violencia	Interno golpeado por sus padres o tutor durante su niñez (5 a 12 años)	Cualitativa Nominal	No, Si
Percepción de consumo de alcohol de familiar	Interno vivió con padres o adultos que tomaban alcohol frecuentemente durante su niñez (5 a 12 años)	Cualitativa Nominal	No, Si
Percepción de consumo drogas de familiar	Interno vivió con padres o adultos que consumían drogas durante su niñez (5 a 12 años)	Cualitativa Nominal	No, Si
Huyo de su casa	El interno huyó de su casa cuando tenía menos de 15 años	Cualitativa Nominal	No, Si
Percepción de violencia familiar	Interno percibió agresión física del padre o pareja hacia la madre	Cualitativa Nominal	No, Si
Familiar con antecedentes penales	Algún familiar del interno estuvo en prisión.	Cualitativa Nominal	No, Si
Vivió con su madre	Interno vivió con su madre	Cualitativa Nominal	No, Si
Vivió con su padre	Interno vivió con su padre	Cualitativa Nominal	No, Si

Considerando los resultados de Escaff et al. (2013) y Molocho Luis (2017), se consideró las siguientes características de sociales de los internos.

Variable	**Descripción**	**Tipo de variable**	**Categoría**
Presencia de bandas o pandillas en su barrio	Presencia de bandas o pandillas en su barrio siendo menor de edad	Cualitativa Nominal	No, Si
Se relacionó con amigos que cometían delitos	Sus mejores amigos cometían delitos antes de cumplir los 18 años.	Cualitativa Nominal	No, Si
Se sintió discriminado	Internó se sintió discriminado alguna vez	Cualitativa Nominal	No, Si
Religión	Religión que practica el interno	Cualitativa Nominal	Católica, Evangélica, Otro, Ninguna
Se relacionó con compañeros que tuvieron problemas con la ley	En los últimos años de secundaria, se relacionó con compañeros que tuvieron problemas con la ley	Cualitativa Nominal	No, Si
Etnia	Identificación del interno según sus costumbres	Cualitativa Nominal	Quechua, Negro, Blanco, Mestizo, Otro
Pertenece a alguna comunidad campesina	Considera que pertenece a alguna comunidad campesina	Cualitativa Nominal	No, Si
Consumió Drogas	Consumió drogas antes de ingresar al centro penitenciario.	Cualitativa Nominal	No, Si
Consumió Alcohol	Consumió alcohol antes de ingresar al centro penitenciario.	Cualitativa Nominal	No, Si
Consumió Cigarro	Consumió cigarro antes de ingresar al centro penitenciario.	Cualitativa Nominal	No, Si
Alguna vez trabajo	Alguna vez trabajo, antes de ingresar al establecimiento penitenciario	Cualitativa Nominal	No, Si

En base a Escaff et al. (2013) se consideran las siguientes características asociadas al delito

Variable	Descripción	Tipo de variable	Categoría
Uso arma durante el delito	Interno utilizó arma cuando ocurrió el delito	Cualitativa Nominal	Si, No
Consumió drogas horas antes del delito	Interno había consumido drogas seis horas antes de cometer el delito	Cualitativa Nominal	Si, No
Consumió alcohol horas antes del delito	Interno había consumido alcohol seis horas antes de cometer el delito	Cualitativa Nominal	Si, No
Otras personas participaron del delito	Otras personas participaron del delito que está acusado	Cualitativa Nominal	Si, No
Estuvo internado en algún centro juvenil	Permaneció internado en algún centro juvenil anteriormente	Cualitativa Nominal	Si, No

3.4. Antecedentes

3.4.1 Antecedentes Nacionales

Según Ancco (2017) en su tesis titulada "Factores asociados al rendimiento académico en los cursos de Matemática Básica y Calculo I de los alumnos ingresante de la FCM-UNMSM utilizando regresión logística binaria", buscó identificar cuáles son los factores asociados al rendimiento de las mismas aplicando la regresión logística binaria, logrando encontrar que los factores que influyen son; el tipo de colegió de procedencia, preferencia por la carrera, condición laboral del alumno y con quien convive el alumno.

Caballero (2021) en "Caracterización del trabajo infantil en el Perú 2019, usando árboles de decisión" plantea caracterizar el trabajo infantil en el Perú en el año 2019, aplicado en 29962 menores de 5 a 17 años datos obtenidos de la ENAHO; obteniendo como factores asociados la región del hogar de residencia, la edad, el ultimo nivel de estudios y la profesión u oficio del jefe del hogar.

Límaco y Solano (2019) en su artículo titulado “Factores asociados a la violencia conyugal hacia la mujer en el Perú, utilizando regresión logistica” teniendo como objetivo identificar y analizar los factores asociados a la violencia contra la mujer en el Perú, utilizando la base de ENDES 2013 incluyendo a 22 mil 920 mujeres de 15 a 49 a años, obteniendo como factores asociados a las violencia psicológica, verbal y física el consumo de alcohol de sus esposos, que trabajen actualmente y si quedo embarazada.

3.4.2 Antecedentes Internacionales

Payam et al (2017) en “Prevalence and Determinants of Preterm Birth in Tehran, Iran: A Comparison between Logistic Regression and Decision Tree Methods”, comparan los dos modelos de regresión logística y árbol de clasificación para determinar las principales causas de muerte del neonatal, aplicado en 4419 mujeres embarazadas remitidas por los hospitales de maternidad de la provincia de Teherán en Irán entre el 6 y el 21 de julio del 2015. La regresión logística obtuvo mejores resultados para la clasificación; mostrando que las madres con preclampsia y aquellas que concibieron con reproducción asistida de la tecnología presentan mayor riesgo de parto prematuro.

IV. METODOLOGIA

4.1 Descripción de la Metodología

Este trabajo se realizó siguiendo la metodología CRIPS-DM definida según (Kimball, 2002).

4.1.1 Comprensión del Negocio

Dentro de los objetivos del área se busca implementar nuevos productos o procesar bases de datos de fuentes secundarias para la utilización de los estudios, en esta etapa se planteó la propuesta para generar una demo de estudio que tiene como potenciales clientes a funcionarios de entidades públicas, que buscan actualizar el plan de acción distrital de seguridad ciudadana, teniendo en cuenta estudios con un enfoque cuantitativo. En base a esto, se buscó determinar el perfil de las personas que cometen delitos, para lo cual se utilizó la base de datos extraída del portal del INEI, y siendo el robo agravado el delito más frecuente dentro de los delitos contra el patrimonio. Se buscó identificar las variables asociadas o que influyen que un interno incurra en el delito de robo agravado. Se delimita el ámbito de estudio solo para Lima Metropolitana por ser el área de influencia con mayor número de proyectos implementados en la empresa.

Población

Todas las personas privadas de su libertad en los centros penitenciarios del Perú durante el 2016, que cometieron algún delito dentro de Lima Metropolitana. (INEI: Primer Censo Penitenciario 2016)

Unidad de análisis

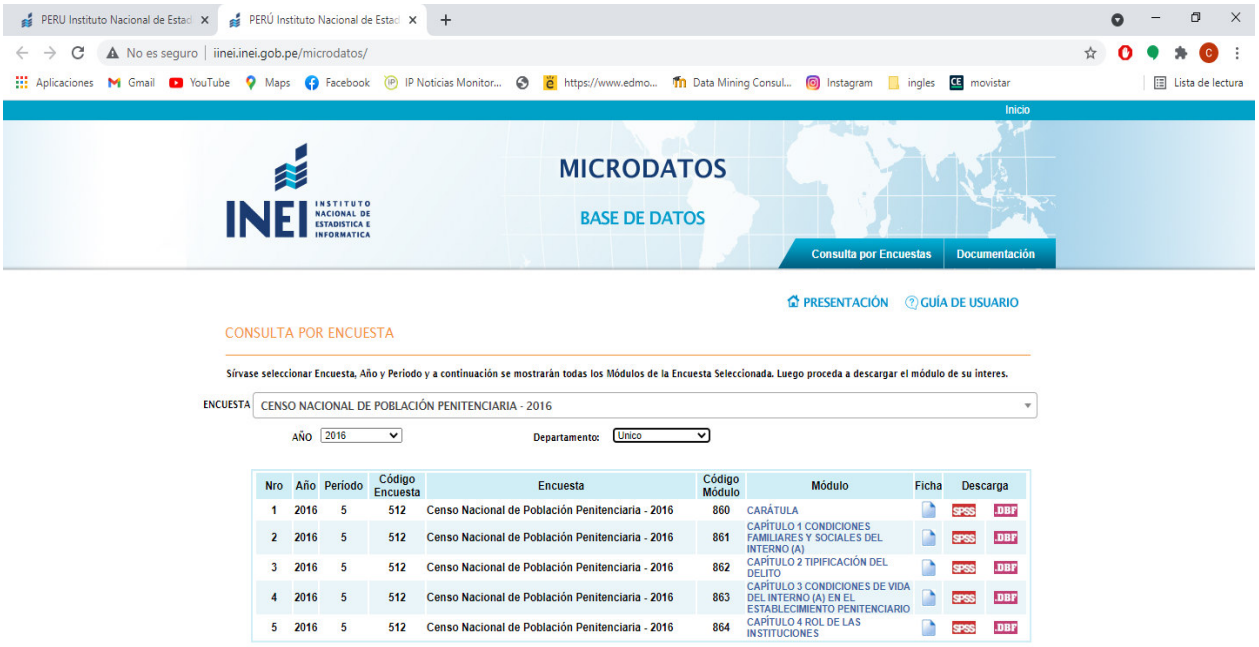
Cada interno que cometió algún delito dentro de Lima Metropolitana

4.1.2 Comprensión de los Datos

En esta etapa se procedió a recolectar los datos del Primer Censo Penitenciario 2016 desde el portal del INEI. En la figura 10 se puede observar el portal del INEI, para seleccionar y acceder a las bases de datos que se encuentra en diferentes módulos (archivos. sav).

Figura 10

Portal del INEI para acceder a las bases de datos

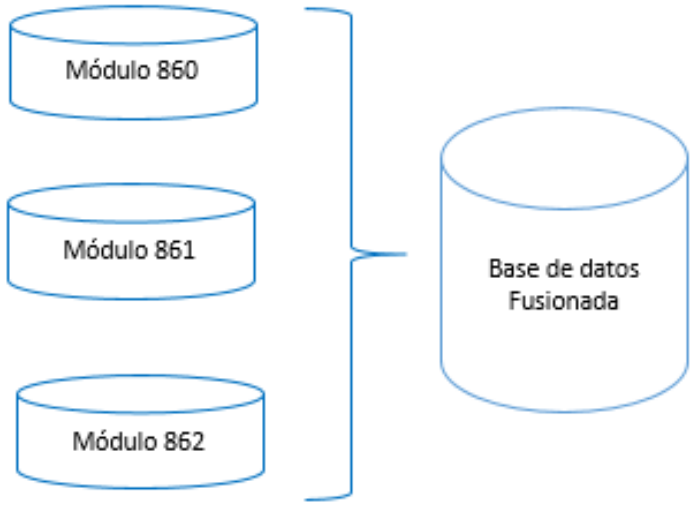


Nota. Obtenido del portal del INEI

Luego se procede a fusionar las tablas como se detalla en la figura 11 con los módulos y sus respectivas variables que se consideraron en este trabajo.

Figura 11

Proceso de fusión de la base de datos del Censo Penitenciario 2016



Nota. Elaboración propia

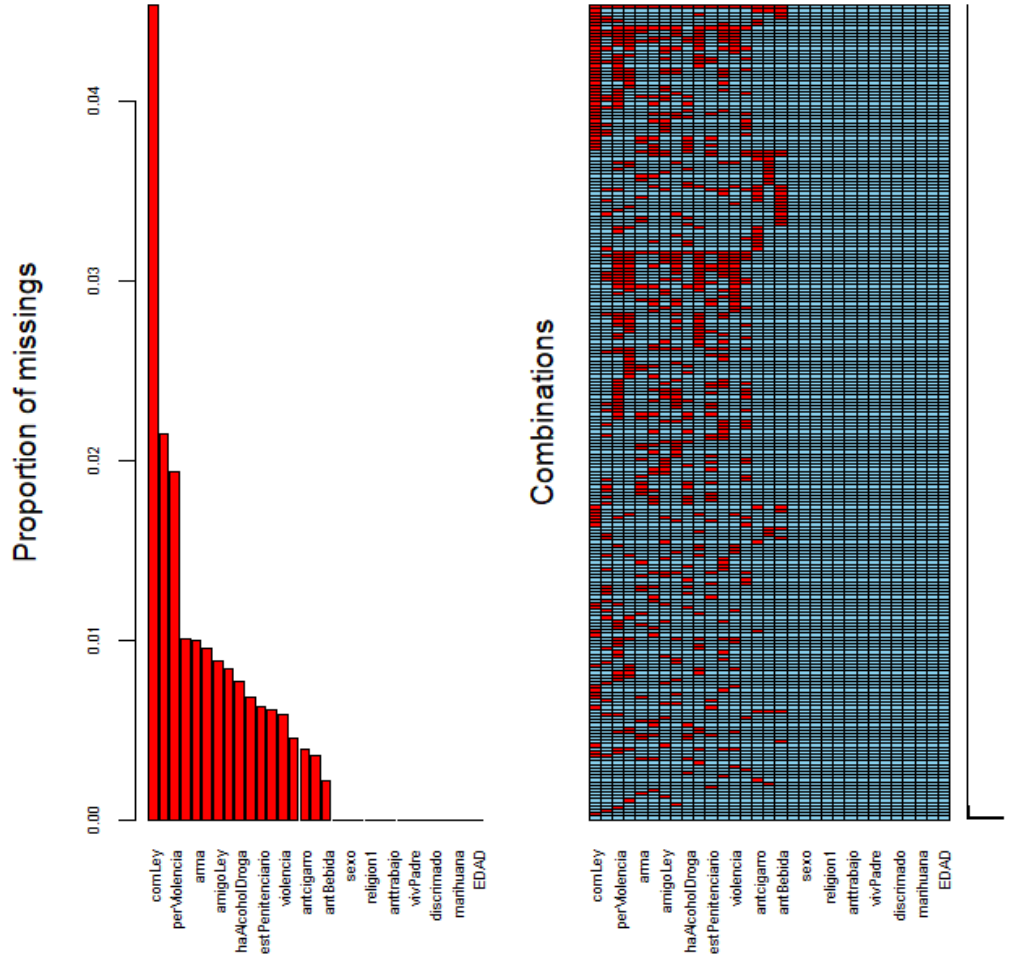
4.1.3 Preparación de los Datos

En esta etapa se realizó la limpieza de los datos de la base fusionada (tratamiento de datos faltantes) y se realizó la transformación de datos (recodificación de atributos y discretización)

Tratamiento de datos faltantes: Luego de realizar un control de calidad de los datos, se encontró 16 variables independientes de tipo cualitativas con datos faltantes; registrados inicialmente con valores indefinidos (No sabe / No contesta).

Figura 12

Identificación de datos faltantes en las variables independientes.



Nota. Elaboración propia

En la figura 12 se observa la presencia de datos faltantes en las variables independientes, donde el porcentaje de datos faltantes es menos del 5% en cada una. Por lo cual se evaluó realizar una imputación no paramétrica utilizando arboles de decisión, el algoritmo CART que se encuentra en el paquete paquete mice. En la tabla 3 podemos observar el porcentaje de valores perdidos en la base de datos.

Tabla 3

Valores faltantes en las variables explicativas

Variable	n	%
Compañeros que cometían delitos	1,076	4.53
Pertenece a alguna comunidad campesina	510	2.15
Percepción de violencia familiar	459	1.93
Percepción de consumo drogas de familiar	239	1.01
Uso arma	236	0.99
Hubo cómplices en el delito	226	0.95
Amigos que cometían delitos	209	0.88
Familiar con antecedentes penales	199	0.84
Percepción de consumo de alcohol de familiar	162	0.68
Permanecer internado en un centro juvenil	150	0.63
Huyo de su casa	145	0.61
Violencia	139	0.59
Presencia de bandas o pandillas en su barrio	108	0.46
Consumió Cigarro	93	0.39
Consumió Drogas	85	0.36
Consumió Alcohol	51	0.21

Nota. Elaboración propia

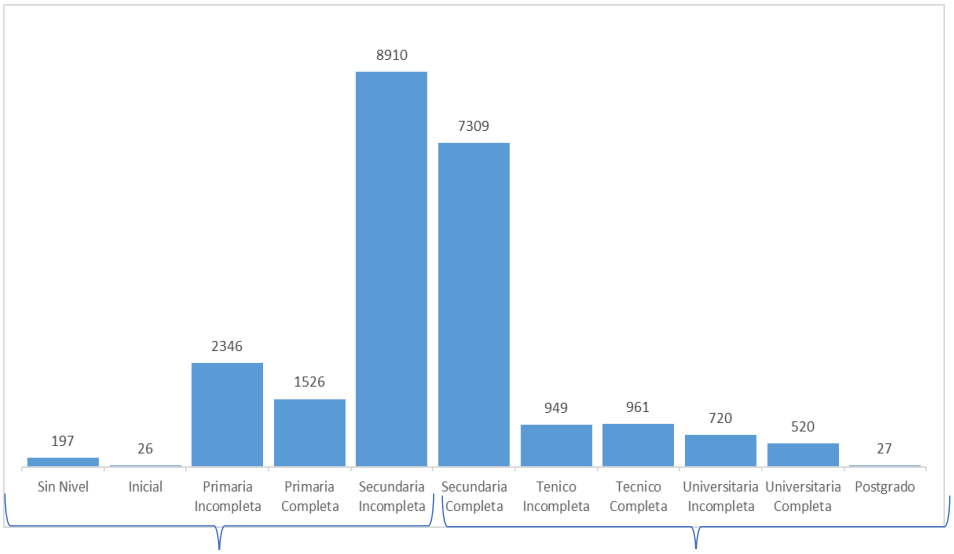
Evaluación de datos atípicos: Se encontraron datos atípicos en la variable edad, las cuales representan un 1 % del total de registros, se optó por eliminar los datos atípicos.

Transformación de datos: Se procedió a recodificar la variable nivel de estudios alcanzado antes de ingresar al establecimiento penitenciario.

El nivel educativo alcanzado por los internos peruanos que cometieron algún delito en Lima Metropolitana fue en su mayoría el nivel secundario incompleta (37.9%), seguido de secundaria completa (31.1%) y los otros niveles de estudios con menos del 10.5% como se observa en la figura 13. En el cuestionario se considera que un interno terminó de estudiar, si cuenta con secundaria completa. Por eso recodificamos como se presenta en la figura 13.

Figura 13

Nivel educativo alcanzado antes de ingresar a la cárcel



Después de recodificar las variables se continuó explorando los datos de manera univariada y bivariada. En la tabla 4, podemos observar que el 34.9 de la población penitenciaria peruana cometió el delito de robo agravado en Lima Metropolitana.

Tabla 4

Distribución de la variable delito que cometió fue robo agravado

		Frecuencia	Porcentaje
Interno cometió el delito de robo agravado	No	15,274	65.1
	Si	8,217	34.9
	Total	23,491	100

Nota. Elaboración propia

De los 23 mil 491 internos que se encontraron en los centros penitenciarios en el 2016, hay un mayor porcentaje de hombres (94.2%) que mujeres (5.8%), con una edad que varía entre los 18 a 66 años. También existe más internos solteros (47.7%), seguido de un 36.4% de internos casados y un 10.3% de internos convivientes, los internos que estudiaron hasta secundaria incompleta representan un 55.9%, del mismo modo el 74.5% de los internos tiene hijos. Ver tabla 5.

Tabla 5

Internos que cometieron algún delito en Lima Metropolitana según características sociodemográficas

Características Demográficas	Frecuencia	Porcentaje
Sexo		
Mujer	1374	5.8
Hombre	22117	94.2
Estado Civil		
Conviviente	2423	10.3
Casado	8561	36.4
Viudo	231	1.0
Divorciado	224	1.0
Separado	836	3.6
Soltero	11216	47.7
Nivel de estudios alcanzado antes de ingresar a la cárcel		
Por lo menos secundaria completa	10594	45.1
Hasta secundaria incompleta	13141	55.9
Tiene hijos		
No	5979	25.5
Si	17512	74.5

Nota. Elaboración propia

Respecto a las características familiares se puede observar en la tabla 6, que el 36.2 % de los internos que cometieron el delito de robo agravado, sus padres o tutores le golpeaban durante su niñez, así mismo el 63.8% de los internos que cometió otro delito sufrió violencia física en su niñez. El 42.8% de los internos que tenían padres o tutores que consumían drogas cometieron el delito de robo agravado, también el 40.7% de los internos que se fue de su casa antes de cumplir los quince años cometieron el delito de robo agravado, respecto a los internos que tenían familiares con antecedentes penales son el 40.4% que cometieron el delito de robo agravado. De los internos peruanos que cometieron el delito de robo agravado en Lima Metropolitana el 39.1% no vivió con su madre y el 37% no vivió con su padre.

Tabla 6

Internos peruanos que cometieron delitos en Lima Metropolitana según características familiares

Características Familiares	Cometió el delito de robo agravado		Chi cuadrado p_valor
	No n (%)	Si n (%)	
Interno fue golpeado por sus padres o tutor durante su niñez			
No	8286 (66.1)	4247 (33.9)	0.00
Si	6988 (63.8)	3970 (36.2)	
Percepción de consumo de alcohol de familiar			
No	10517 (65.6)	5527 (34.4)	1.28
Si	4757 (63.9)	2690 (36.1)	
Percepción de consumo drogas de familiar			
No	14666 (65.4)	7762 (34.6)	0.00
Si	608 (57.2)	455 (42.8)	
Huyo de su casa			
No	10153 (68.4)	4701 (31.6)	0.00
Si	5121 (59.3)	3516 (40.7)	
Percepción de violencia familiar			
No	10784 (65.6)	5647 (34.4)	0.28
Si	4490 (63.6)	2570 (36.4)	
Familiar con antecedentes penales			
No	10395 (67.9)	4908 (32.1)	0.00
Si	4879 (59.6)	3309 (40.4)	
Vivió con su madre			
Si	12703 (65.9)	6564 (34.1)	0.00
No	2571 (60.9)	1653 (39.1)	
Vivió con su padre			
Si	14551 (65.1)	7793 (34.9)	0.00
No	723 (63.0)	424 (37.0)	

Nota. Elaboración propia

Dentro de las características sociales podemos observar en la tabla 7; que el 41% de los internos peruanos que vivían en un barrio con presencia de bandas o pandillas antes de cumplir la mayoría de edad cometió el delito de robo agravado en Lima Metropolitana. También se puede observar que de los internos que cometieron el delito de robo agravado, el 45% tuvieron mejores amigos con problemas con la ley cuando eran menores de edad. El 48.7% de los internos que

tuvieron compañeros con problemas con la ley mientras estudiaban en la secundaria cometió el delito de robo agravado.

De los internos que cometieron el delito de robo agravado, el 35.5% no considera que pertenece a alguna comunidad campesina. Por otro lado, el 7.8% de los internos que consumía drogas antes de ingresar al centro penitenciario cometió el delito de robo agravado; así mismo como el 37.4% de los internos que consumía alcohol antes de estar en la cárcel cometió el delito de robo agravado y de los internos que cometió el delito de robo agravado el 47% de los internos peruanos no trabajo alguna vez antes de ingresar al centro penitenciario.

Además, observando la prueba Chi Cuadrado podemos ver que las características sociales se encuentran asociadas con el hecho que el interno cometió el delito de robo agravado; a excepción de las variables si se sintió discriminado y el tipo de etnia que considera pertenecer.

Por otro lado, observando la tabla 8 notamos que la prueba Chi cuadrado nos indica que todas las características correspondientes al delito están asociadas a al hecho que el interno cometió el delito del robo agravado.

De los internos peruanos que cometió el delito de robo agravado en Lima Metropolitana el 48.8% de los internos uso un arma mientras cometía el delito y el 49% había consumido drogas horas antes de cometer el delito. Así mismo, el 49.2% estuvo en algún centro penitenciario anteriormente y el 46.2% tenía compañeros que participaron del delito que fue acusado. El 37.4% de los internos que consumió alcohol seis horas antes de ocurrido el delito cometió el delito de robo agravado.

Tabla 7

Internos que cometieron delitos en Lima Metropolitana según características Sociales

Características Sociales	Cometió el delito de robo agravado		Chi cuadrado p_valor
	No n (%)	Si n (%)	
Presencia de bandas o pandillas en su barrio			
No	6388 (76.8)	1930 (23.2)	0.00
Si	8886 (58.6)	6287 (41.4)	
Sus mejores amigos cometían delitos antes de los 18 años			
No	9560 (73.0)	3533 (27.0)	0.00
Si	5714 (55.0)	4684 (45.0)	
Se sintió discriminado			
No	13120 (65.1)	7025 (34.9)	4.09
Si	2154 (64.4)	1192 (35.6)	
Religión			
Evangélica	3870 (63.3)	2248 (36.7)	0.00
Católica	9839 (66.8)	4892 (33.2)	
Otra	407 (68.3)	189 (31.7)	
Ninguna	1158 (56.6)	888 (43.4)	
Se relacionó con compañeros que tuvieran problemas con la ley			
No	13369 (67.6)	6412 (32.4)	0.00
Si	1905 (51.3)	1805 (48.7)	
Etnia			
Otra etnia	2789 (63.6)	1599 (36.4)	2.56
Se considera mestizo o blanco	12485 (65.4)	6618 (34.6)	
Pertenece a una comunidad campesina			
No	14336 (64.5)	7893 (35.5)	0.00
Si	938 (74.3)	324 (25.7)	
Consumía drogas antes de ingresar a la cárcel			
No	10407 (71.7)	4112 (28.3)	0.00
Si	4867 (92.2)	410 (7.8)	
Consumía alcohol antes de ingresar a la cárcel			
No	5764 (69.5)	2525 (30.5)	0.00
Si	9510 (62.6)	5692 (37.4)	
Consumía cigarro antes de ingresar a la cárcel			
No	10542 (67.6)	5043 (32.4)	0.00
Si	4732 (59.9)	3174 (40.1)	
Trabajo alguna vez			
No	639 (53.0)	567 (47.0)	0.00
Si	14635 (65.7)	7650 (34.3)	

Nota. Elaboración propia

Tabla 8

Internos que cometieron delitos en Lima Metropolitana según características asociadas al delito

Características asociadas al delito	Cometió el delito de robo agravado		Chi cuadrado p_valor
	No n (%)	Si n (%)	
Uso arma durante el delito			
No	12931 (68.4)	5985 (31.6)	0.00
Si	2343 (51.2)	2232 (48.8)	
Consumido drogas antes del delito (6 horas antes)			
No	14619 (65.8)	7588 (34.2)	0.00
Si	655 (51.0)	629 (49.0)	
Consumido alcohol antes del delito (6 horas antes)			
No	12443 (68.3)	5772 (31.7)	0.00
Si	2831 (53.7)	2445 (46.3)	
Otras personas participaron del delito			
No	9170 (75.6)	2966 (24.4)	0.00
Si	6104 (53.8)	5251 (46.2)	
Estuvo internado en algún centro juvenil			
No	13870 (66.9)	6855 (33.1)	0.00
Si	1404 (50.8)	1362 (49.2)	

Nota. Elaboración propia

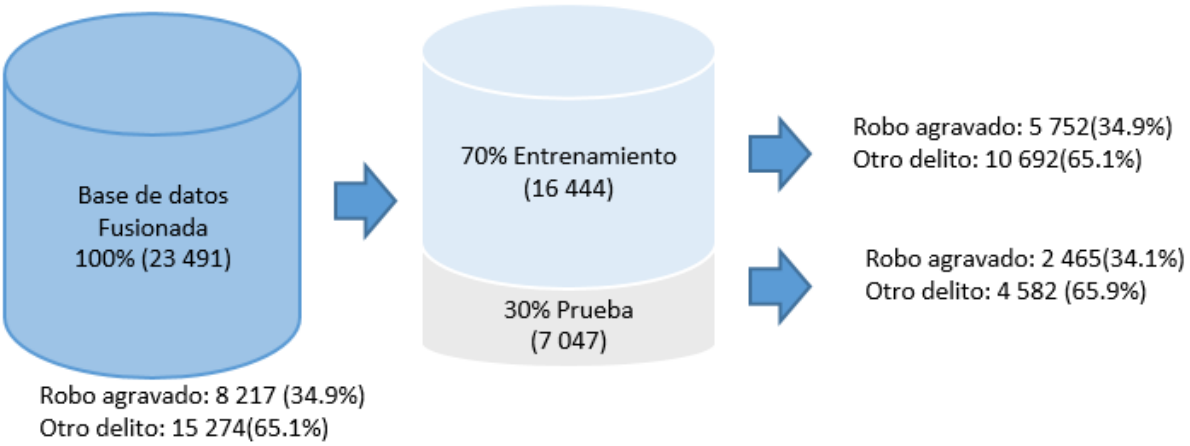
Como siguiente paso se procede a seleccionar las variables independientes que entraran al modelo; sin considerar las variables percepción de consumo de alcohol de familiar, percepción de violencia familiar, se sintió discriminado y etnia, puesto que mediante la prueba Chi Cuadrado nos indica que no existe asociación significativa a cometer el delito de robo agravado. Se realiza la selección de variables mediante el procedimiento de backward stepwise, basado en el criterio de Akaike, donde el modelo inicial considera todas las variables regresoras, así como todas las interacciones de orden inferior y se elige el modelo con menor AIC. Quedándonos con un modelo que contiene 14 variables independientes con un AIC de 26104.

4.1.4 Modelamiento

En esta etapa para evitar el sobreajuste en los modelos aplicados se dividió los datos en un 70% para la muestra de entrenamiento y en 30% para la muestra de prueba, como se puede observar en la figura 14.

Figura 14

Selección de la muestra de entrenamiento y prueba



Nota. Elaboración propia

En la muestra de entrenamiento se ajustan los modelos explicados de regresión logística y el algoritmo CART explicado en los ítems (3.1.1 y 3.1.2) y en la muestra de prueba se realizó la evaluación para ver la capacidad de discriminación de los modelos.

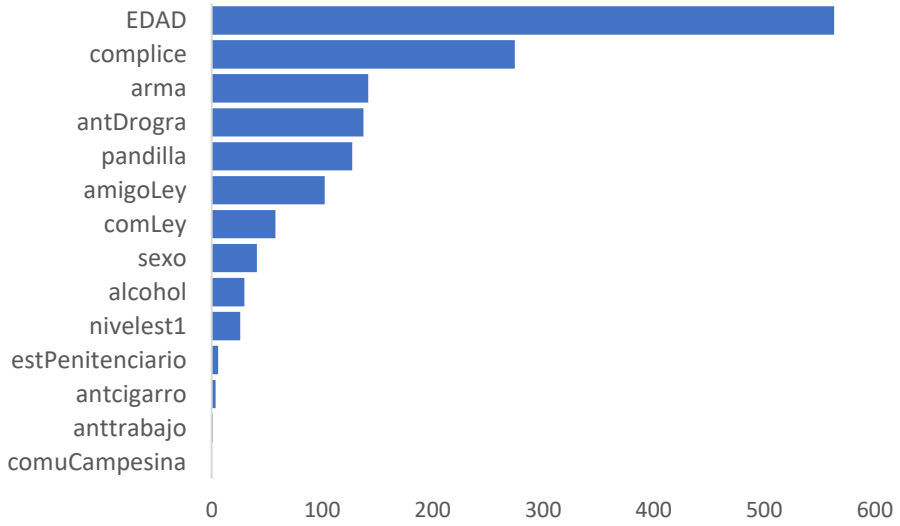
Modelo con el algoritmo CART

Se ajusta el algoritmo CART con las catorce variables construyendo un árbol completo en primera instancia, para después realizar la poda a un costo de complejidad de 0.001 obteniendo el mínimo error relativo, con un nodo hijo de mínimo 50 observaciones y para que sea un nodo padre mínimo 100 observaciones.

En la figura 15 podemos observar la importancia de variables del modelo, según el índice gini utilizado en cada partición del nodo.

Figura 15

Importancia de variables involucradas asociadas al robo agravado



Después de realizar el árbol de clasificación final, la edad es la variable más importante asociada a la acción de cometer el delito de robo agravado, en segundo lugar y tercer lugar tenemos características asociadas al delito; las cuales son: si se encontraban involucradas otras personas en el delito que fue acusado el interno y si uso arma en el momento que ocurrió el delito, respectivamente.

En cuarto lugar, tenemos a la variable si antes de ingresar al centro penitenciario consumía drogas, seguido de si cuando era menor de edad: había pandillas en al barrio donde vivía, sus mejores amigos cometían actos delictivos o si durante la secundaria tenia compañeros con problemas con la ley.

Tabla 9

Matriz de los internos que cometió el delito de robo agravado y la predicción basado en el árbol de clasificación

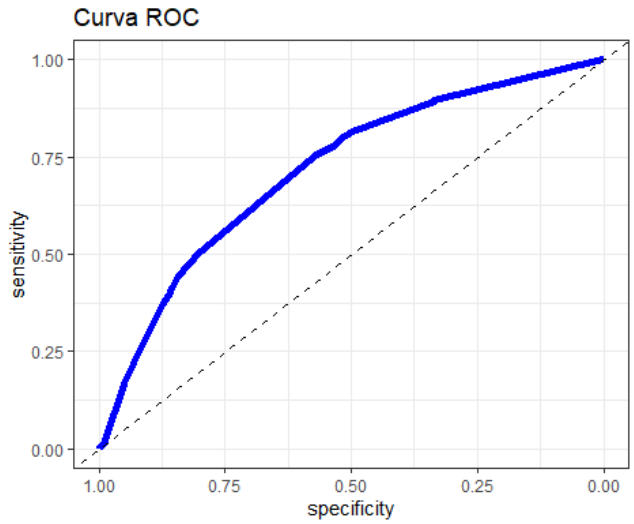
Predicción	Cometió robo agravado	
	No	Si
No	3859	1373
Si	723	1092

Nota. Elaboración propia

Se realiza la validación del modelo en la muestra de prueba, como se observa en la tabla 9 se observa que los internos que cometieron el delito de robo agravado se estimó con una precisión de 44% (sensibilidad) y aquellos internos que cometieron otro tipo de delito en Lima Metropolitana es de 84% (especificidad).

Figura 16

Curva característica de operación del modelo de árbol de clasificación. Algoritmo CART.



Nota. Elaboración propia

El área bajo la curva es de 0.72, lo cual es el estadístico por excelencia para medir la capacidad discriminante de la prueba.

Modelo de regresión logística

Como se definió en él ítem (3.1.2); se realizó el ajuste del modelo de regresión logístico, la cual se formula de la siguiente manera:

$$P(Cometio\ el\ delito\ de\ robo\ agravado = Si) = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

donde:

η=-0.756-0.055*Edad+0.869*Interno tuvo complices+0.178*Tenía amigos que cometían delitos+1.103*Interno es hombre+0.357*Interno si uso arma en el delito+0.293*Hasta secundaria incompleta+0.168*Había pandillas en su barrio+0.197*Si consumía drogas antes de ingresar al penal+0.391*Si consumió alcohol seis horas antes de que ocurra el delito-0.383 Si considera ser de una comunidad campesina o nativa+0.123*Si estuvo en un centro juvenil+0.174*Tenía compañeros de secundaria que cometían delitos+0.091*Si consumía cigarro antes de ingresar al penal-0.184*Si trabajo alguna vez antes de ingresar al penal.

Después se evalúo la bondad de ajuste global del modelo logístico, obteniendo el siguiente estadístico de prueba ($G = 3079.79 > \chi^2_{(14,0.95)} = 23.68$) por lo que se concluye que es un modelo significativa porque por lo menos una de las variables independientes propuestas tiene asociación con la probabilidad de que un interno peruano cometa el delito de robo agravado en Lima Metropolitana. En conformidad con lo mencionado anteriormente la prueba de significancia individual de los parámetros, la estadística de Wald confirma la importancia de las catorce variables incluidas en el modelo como se observa en tabla 10.

Tabla 10

Resultado del modelo de regresión logística para los internos peruanos que cometieron el delito de robo agravado, año2016

Variable	$\hat{\beta}$	SE	p-valor	$Exp(\hat{\beta})$	I.C(95%) $Exp(\hat{\beta})$	
					L.Inferior	L.Superior
Edad(Si)	-0.055	0.002	0.000	0.947	0.943	0.950
Amigos que cometían delitos(Si)	0.178	0.042	0.000	1.195	1.100	1.298
Sexo(Hombre)	1.103	0.104	0.000	3.012	2.468	3.706
Uso arma(Si)	0.357	0.044	0.000	1.429	1.310	1.558
Hubo cómplices en el delito(Si)	0.869	0.037	0.000	2.383	2.219	2.561
Grado de Instrucción(Hasta Secundaria incompleta)	0.293	0.038	0.000	1.341	1.245	1.444
Presencia de bandas o pandillas en su barrio(Si)	0.168	0.043	0.000	1.183	1.087	1.288
Consumió Drogas(Si)	0.197	0.040	0.000	1.218	1.126	1.318
Consumir alcohol horas antes del delito(Si)	0.391	0.042	0.000	1.478	1.362	1.605
Pertenece a alguna comunidad campesina(Si)	-0.383	0.088	0.000	0.681	0.573	0.808
Permanecer internado en un centro juvenil(Si)	0.123	0.055	0.025	1.131	1.015	1.260
Compañeros que cometían delitos(Si)	0.174	0.050	0.000	1.190	1.079	1.312
Consumió Cigarro(Si)	0.091	0.038	0.018	1.095	1.016	1.181
Alguna vez trabajo(Si)	-0.184	0.078	0.019	0.832	0.713	0.970

Nota. Elaboración propia

De la tabla 10, obtenemos las siguientes interpretaciones: Aquellos internos que tenían amigos con problemas con la ley, siendo menores de edad tienen 1.2 más posibilidades de cometer el delito de robo agravado frente a los que no tuvieron amigos que cometían delitos. Los internos que son hombres tienen el triple de chance que cometan el delito de robo agravado respecto a las mujeres.

Los internos con que estudiaron hasta secundaria incompleta antes de ingresar al establecimiento penitenciario, tienen el 1.3 de mayor posibilidad que los internos que estudiaron por lo menos secundaria completa. Así también los internos que vivieron en un barrio con presencia de bandas delictivas y pandillas antes de cumplir los dieciocho años son 1.18 más probable que cometan el delito de robo agravado, frente los internos que no tenían pandillas en el barrio donde vivían antes de los dieciocho años.

El interno que consumía drogas antes de ingresar al establecimiento penitenciario es 1.21 veces más probable de cometer el delito de robo agravado, frente a los internos que no consumieron drogas entes de ingresar al penal. El interno que estuvo internado en algún centro juvenil tiene 1.13 más posibilidad de cometer el delito de robo agravado que el interno que no estuvo internado en un centro juvenil. El interno tiene el doble de chance de cometer el delito de robo agravado junto a otras personas que estando solo. Otros factores de riesgo son que el interno consumiera seis horas antes del delito alcohol, y consumiera cigarro antes de ingresar al centro penitenciario.

Como factores protectores tenemos a la edad, si se considera parte de una comunidad nativa o campesina y si trabajo alguna vez antes de ingresar al establecimiento penitenciario.

Tabla 11

Matriz de los internos que cometió el delito de robo agravado y la predicción basado en la regresión logística

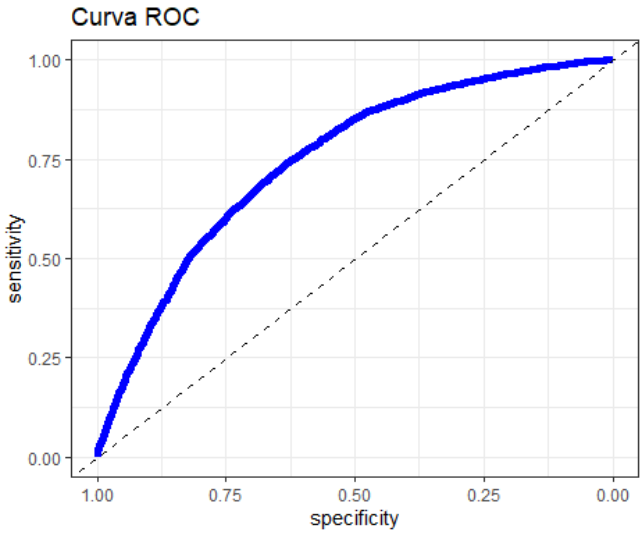
Predicción	Cometió el delito de robo agravado	
	No	Si
No	2834	615
Si	1748	1850

Nota. Elaboración propia

De la tabla 11, observamos que lo falsos negativos es de 1748 internos y los falsos positivos es de 615 internos, en comparación con los verdaderos positivos y verdaderos negativos es menor. Por lo resulta una sensibilidad del 75%, la cual nos indica la proporción de internos peruanos que cometieron el delito de robo agravado en Lima Metropolitana que fueron clasificados correctamente.

Figura 17

Curva característica de operación del modelo de regresión logística.



Nota. Elaboración propia

El área bajo la curva es de 0.75, lo cual es el estadístico por excelencia para medir la capacidad discriminante de la prueba.

4.1.5 Evaluación

En esta etapa se compara los dos modelos ejecutados en la etapa anterior obteniendo los siguientes resultados:

Tabla 12

Métricas de evaluación de modelos

Indicadores	CART	Logístico
Sensibilidad	44%	75%
Especificidad	84%	62%
AUC	72%	75%

Nota. Elaboración propia

Como se puede observar en la tabla 12, el modelo logístico discrimina mejor a los que cometieron el delito de robo agravado (sensibilidad 75%) en comparación al árbol de clasificación CART. Aunque, el algoritmo CART predice mejor a los que no cometieron el delito de robo agravado (especificidad 84%), se optó por considerar los resultados del modelo de regresión logística, porque este estudió tiene como objetivo principal determinar las variables más importantes asociadas a las personas que cometieron el delito de robo agravado.

4.1.6 Implementación

Se procedió a realizar la implementación de la demo en la plataforma de la empresa, mostrando las capas de información según distrito.

Análisis de los Resultados

Después de comparar los modelos de clasificación, nos quedarnos con los resultados obtenidos por el modelo de regresión logística; en las características demográficas identificamos como factor de riesgo; que el interno sea hombre, si el interno estudio hasta secundaria incompleta y como factor protector; que por cada año más que tenga el interno disminuye la posibilidad de cometer el delito de robo agravado.

Dentro de las características sociales encontramos los siguientes factores de riesgo: El interno tenía amigos con problemas con la ley siendo menores de edad, el interno vivió en un barrio con presencia de bandas delictivas y pandillas antes de cumplir los dieciocho años, el interno que consumía drogas y cigarro antes de ingresar al establecimiento penitenciario, el interno que estuvo internado en algún centro juvenil, el interno tenia compañeros con problemas con la ley cuando estaba en la secundaria y como factor protector tenemos que; el interno considera que pertenece a una comunidad nativa o campesina y si trabajó alguna vez antes de ingresar al establecimiento penitenciario.

Para las características referentes al delito las variables que se encuentran muy asociadas son uso arma cuando ocurrió el delito, consumió alcohol seis horas antes de cometer el delito alcohol y tuvo cómplices para cometer el delito.

V. CONCLUSIONES

Se consolido la información proveniente de los módulos 860, que corresponde a información de datos personales del interno, 861 contiene características sociales y familiares del interno y 862 tiene información de las características relacionadas a la tipificación del delito.

Se realizó la comparación del modelo de regresión logística y el modelo proveniente del algoritmo CART, determinando como mejor modelo a la regresión logística con una sensibilidad del 75% y un AUC 75%, el cual indica que es un modelo que discrimina bien a un nuevo interno.

Por último, se identifica 1) las características sociales tales como: si sus mejores amigos antes de cumplir los 18 años cometían delitos, si había presencia de bandas o pandillas en su barrio donde creció y si sus compañeros de secundaria tenían problemas con la ley. 2) Las características relacionadas a la tipificación del delito: si uso arma durante el homicidio y si había otras personas involucradas en el delito del que fue acusado y 3) Las variables

sociodemográficas: edad y el grado de instrucción; como variables importantes para describir a un interno que cometa el delito de robo agravado.

VI. RECOMENDACIONES

Se recomienda incluir información de la zona de residencia del interno, para próximos cuestionarios; puesto que esto describir el lugar de procedencia de los internos.

Recomiendo seguir investigando en este problema social, considerando otras regiones del Perú para poder comparar los resultados por regiones.

VII. BIBLIOGRAFIA

Ancco, A. (2017). Factores asociados al rendimiento académico en los cursos de matemática básica y cálculo I de los alumnos ingresantes de la FCM-UNMSM utilizando regresión logística binaria [Tesis de pregrado, Universidad Nacional Mayor de San Marcos].

Aquino, J. C. (2019). Variables que explican los rangos remunerativos del primer empleo de los egresados universitarios del Perú aplicando regresión logística ordinal [Tesis de pregrado, Universidad Nacional Agraria La Molina].

Breiman, L., Friedman, J., Stone, C., y Olshen, R. (1984). Classification and regression trees. CRC press.

Caballero, K. (2021). Caracterización del trabajo infantil en el Perú 2019, usando árboles de decisión [Trabajo de investigación, Universidad Nacional Mayor de San Marcos].

Chavez, V. (2018,17 de agosto). Arboles Decisión. https://rpubs.com/elfenixsoy/arbol-veronica

Coronado, L. (2019). Factores asociados al problema de la delincuencia y propuesta de solución en el distrito de Castilla-Piura. https://repositorio.unp.edu.pe/bitstream/handle/UNP/1770/IND-COR-NIZ-18.pdf?sequence=1&isAllowed=y

Defensoria del Pueblo. (22 de abril de 2020). Frente al riesgo de masificación del contagio, Defensoría del Pueblo plantea recomendaciones para deshacinar las cárceles. https://www.defensoria.gob.pe/frente-al-riesgo-de-masificacion-del-contagio-defensoria-del-pueblo-plantea-recomendaciones-para-deshacinar-las-carceles/

Escaff, E., Alfaro, R., Ledezma, C., y González, M. (2013). Factores asociados a la reincidencia en delitos patrimoniales, según sexo: estudio desde la perspectiva personal de

condenados(as) en dos penales de Santiago de Chile. Revista Criminalidad, 55 (2), 79-98. http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S1794-31082013000200005

Gallarday, S. (2016). Factores de reincidencia de los internos en el delito de robo agravado del Centro Penitenciario San Pedro-Lurigancho-2016. https://repositorio.ucv.edu.pe/bitstream/handle/20.500.12692/7539/Molocho_VLE.pdf?sequence=1&isAllowed=y

Gervilla García, E., Jiménez López, R., Montaño Moreno, J. J., Cajal Blasco, B., & Palmer Pol, A. (2008). The methodology of Data Mining. An application to alcohol consumption in teenagers. 65-80.

Gironés, J., Casas, J., & Minguillon, J. (2017). Minería de datos: modelos y algoritmos. En J. Girones, J. Casas, & A. Minguillón, Minería de datos: modelos y algoritmos (págs. 209 - 228). Barcelona: UOC.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining Concepts and Techniques. ELSEVIER.

Hand, D., Mannila, H., & Smyth, P. (2001). Principles of Data Mining. Cambridge: Massachusetts Institute of Technology.

Hilbe, J. (20009). Logistic Regression Models. CRC Press.

Hosmer, D., Lemeshow, S., & Sturdivant, R. (2013). Applied Logistic Regression. New Jersey: John Wiley & Sons.

Instituto Nacional Penitenciario. (2019). Informe Estadístico. Lima.

Kantardzic, M. (2011). Data mining: concepts, models, methods, and algorihms.

Kimball, R. (2002). The Data Warehouse. Second ed.

Límaco,W y Solano,O (2019). Factores asociados a la violencia conyugal hacia la mujer en el Perú, utilizando Regresión Logística.PESQUIMAT,107-118. http://dx.doi.org/10.15381/pesquimat.v22i2.17237

Lizares Castillo, M. (2017). Comparación de modelos de clasificación: regresión logística y árboles de clasificación para evaluar el rendimiento académico. Lima.

Payam, A., Saman, M., Reza, S., & Omid, H. (2017). Prevalence and determinants of preterm birth in Tehran, Iran: a comparison between logistic regression and decision tree methods. Public Health and research perspectives, 195-200. https://doi.org/10.24171/j.phrp.2017.8.3.06

Przemyslaw, J., y Krzysztof, K. (2019). Classification of the Symbolic Financial Data on the Forex Model. Computational Collective Intelligence [ICCCI] (págs. 122-132). Hendaye, France: Springer.

Urquizo, J. (2010). Codigo Penal. Lima: Moreno S.A.

Valderrama, M. (2013). Factores que influyen en la reincidencia del delito por robo agravado de los adolescentes infractores de la ley del centro juvenil de diagnóstico y rehabilitación Trujillo en el período 2012-2013. https://dspace.unitru.edu.pe/bitstream/handle/UNITRU/4271/VALDERRAMA%20FERNANDEZ%20MARIA%20YNES(FILEminimizer).pdf?sequence=1&isAllowed=y

Vigo, G. (2010). Método de clasificación para evaluar el riesgo crediticio: una comparación. [tesis de pregrado, Universidad Nacional Mayor de San Marcos], Lima. https://hdl.handle.net/20.500.12672/16835

Weiss, G., & Davison, B. (2010). Data Mining. *To appear in the Handbook of Technology Management, H. Bidgoli* (págs. 1-17).