UNIVERSIDAD NACIONAL DE INGENIERIA

FACULTAD DE ECONOMÍA, ESTADÍSTICA Y CIENCIAS SOCIALES

ESCUELA PROFESIONAL DE INGENIERÍA ESTADÍSTICA

TESIS

Comparison Between Neural Networks and Decision Trees Methods for Predicting the Criminal Behavior of Young People Using Information from the National Census 2016

EJECUTOR: ADAMA ESPERILLA JHONNY IVAN

ASESOR: xxx

Lima, 2016

ÍNDICE

RESUMEN

El Perú actualmente presenta problemas con la delincuencia de los jóvenes, para ello este estudio está dirigido a determinar factores que perfilan a los jóvenes infractores mediante técnicas de árbol de decisión y redes neuronales, la información teórica y la base de datos se recibió del primer censo nacional a centros juveniles realizado el 14 de agosto de este año, que contaba con 5 módulos. El tratamiento de las variables se realizo mediante métodos de imputación de datos, se utilizo la metodología de over-sampling para el tratamiento de la asimetría, y para verificar que los parámetros no estén sobreestimados se realizó el corss-validation, de los resultados el árbol de decisiones fue l que presento mejor estabilidad de los errores trabajados en el coss-validation, además de su aproximación de los indicadores que tuvo con las redes neuronales, finalmente se determinó los indicadores para compararlas técnicas y así elegir la técnica más adecuada, la cual fue4 la técnica de árbol de decisiones por sus indicadores

ABSTRAC

Peru currently presents problems with juvenile delinquency, for this study is aimed at determining the factors that are part of the young offenders using decision tree techniques and neural networks, the theoretical information and the database received from the first census National to youth centers held on August 14 this year, which had 5 modules. The treatment of the variables was performed using data imputation methods, the sampling methodology was used for the treatment of asymmetry, and to verify that the parameters were not overestimated the corss-validation was performed, the results of the tree of the Which has been proved the best stability of the errors worked in the validation of the coss, in addition to its approximation of the indicators that had with the neural networks, finally the indicators for the techniques and the selection of the most adequate technique were determined It was the tree-making technique for its indicators

DEDICATORIA

Este trabajo va dedicado a mi madre Lidia Esperilla Rivera, la mujer que me dio la vida y gracias a ella puede hacer mis sueños realidad

A mi hermano Richard Adama que me apoyo y no desistió hasta que ingrese a la universidad

A mis hermanas Jessenia, Nidia, Olivia, Margot, Yomira y Milagros por estar siempre con migo e incentivarme a no rendirme

A mis amigos en general que gracias a ellos pude aprender muchas cosas y desarrollarme como profesional.

AGRADECIMIENTOS

Director de la Unidad Estadística del INPE
Ing. Marcos lujan del Carpio

Profesor de la Escuela Profesional de Estadística - UNI
Lic. Huamanchumo de la Cuba Luis

Profesor de la Escuela Profesional de Estadística - UNI
Ing. Richard Fernandez

UNIVERSIDAD NACIONAL DE INGENIERÍA

MINISTERIO DE JUSTICIA Y DERECHOS HUMANOS

CAPITULO I

ANTECEDENTES

a. Dr. Emilio Octavio de Toledo y Ubieto y Dra. María Martín Lorenzo en su investigación denominada: **LOS MENORES DE EDAD INFRACTORESDE LA LEY PENAL**

Previamente a aventurarnos a definir, apoyados en la dogmática, lo que debe entenderse por un “menor infractor”, consideramos de trascendental relevancia hacer un breve estudio de la evolución que ha tenido esta noción a lo largo de la historia jurídica, tanto en el mundo como particularmente en España y en México. Ello, con el fin de aclarar, de inicio, el camino para el estudio de la actual legislación de la materia, y determinar no sólo la naturaleza y el contenido del polémico ‘derecho de menores’, sino emitir, en la medida de lo posible, una valoración sobre ésta con miras a su mejoramiento futuro; objetivo prioritario de este trabajo.

A nadie escapa que las ideas con relación a la significación del derecho penal y sus fines, han mutado notoriamente a lo largo de los siglos. Como señala RÍOS ESPINOSA: “las respuestas han oscilado entre las justificaciones retribucionistas absolutas que justifican la aplicación de una pena, en atención a la pretendida disolución del mal producido por el delito

por el correspondiente de la pena, y las justificaciones de orden utilitarista, que atienden no al fin de la pena como legítimo en sí mismo, sino a fines extrapunitivos asociados a ella.”4

Dichas concepciones utilitaristas, predominantes durante la modernidad, han permitido en múltiples ocasiones la extralimitación de la intervención punitiva del Estado, aduciendo la necesidad de garantizar el bienestar de la mayoría no delincuente.

En la actualidad, se acepta la doble finalidad del derecho penal, según la cual “éste cumple también importantes funciones como herramienta de minimización de la violencia hacia los destinatarios de sus normas, cuando las personas caen en el supuesto de infracción a la ley penal.”

b. Manel Capdevila Capdevila, Marta Ferrer Puig y Eulàlia Luque Reina año 2015 en su estudio: **LA REINCIDENCIA EN EL DELITO EN LA JUSTICIA DE MENORES**

La investigación que se presenta a continuación aporta datos concretos y actualizados de la reincidencia en el delito protagonizada por los jóvenes infractores que han entrado en el circuito de la justicia de menores en el ámbito territorial de Catalunya, después de la entrada en vigor de la Ley Orgánica 5/2000, de 12 de enero, Reguladora de la Responsabilidad Penal de los Menores (a partir de ahora LORPM).

El periodo de estudio se sitúa entre enero de 2002 y diciembre de 2004. La población objeto de estudio la componen todos los jóvenes que han finalizado una medida judicial en el año 2002 (en caso de haber más de una, la primera) y el periodo de seguimiento finaliza en diciembre de 2004, para averiguar si se han producido nuevos contactos con el sistema penal de menores o de adultos y por tanto reincidencia.

Los motivos que fundamentan la necesidad de llevar a cabo este estudio han sido varios.

En primer lugar, era necesario actualizar los datos disponibles a Catalunya sobre la reincidencia de los menores. El único estudio anterior que aborde este tema para toda la población atendida desde la Dirección General de Justicia Juvenil es de 1996 (Funes, Luque y Ruiz) y recogía datos de 1993. Era pues, necesario, plantearse una revisión.

En segundo lugar, desde el anterior estudio, tenemos un nuevo marco legal, derivado de la entrada en vigor en el año 2001 de la LORPM y un cambio del perfil de los menores infractores (causado por la misma Ley y por otros factores suficientemente importantes como la inmigración). La LORPM implica un cambio substancial en la edad y en el nombre de menores que son objeto de intervención y también en la forma de intervenir des de los servicios de de ejecución penal del Departamento de Justicia. Se imponía una revisión urgente de los nuevos perfiles de jóvenes llegados a esta nueva Ley y de sus comportamientos criminológicos, incluida la reincidencia, que diera elementos nuevos a los profesionales para planificar su intervención.

En tercer lugar, hacen falta datos objetivos que permitan valorar de manera realista el estado actual de la delincuencia juvenil y la validez de las actuales respuestaspenales. Benito et alt (2004) 1 mencionan en un artículo la denuncia que hacen los magistrados de la Sección 4a de la Audiencia Provincial de Madrid en la cual exponen que, pese a que hace muy poco tiempo de la entrada en vigor de la LORPM, ya se esta argumentando su ineficacia, sin hacer un análisis serio sobre los factores que influyen en la conducta delictiva de los jóvenes y adolescentes y se demanda un endurecimiento de las sanciones y la primacía del castigo sobre el tratamiento reeducador.

Cristina Rechea (2001) 2 recoge como causa de la alarma social que la aplicación de la Ley supone en determinados sectores, la tendencia a fijarse en un caso llamativo que atrae la atención de los medios de comunicación y la generalización que se hace de este caso concreto, sin que haya datos fundamentados que confirmen estos miedos, ni existan

estudios que contrasten un crecimiento de la violencia juvenil o un endurecimiento de las conductas de los jóvenes infractores.

Esta investigación ha de contribuir modestamente a disponer de datos objetivos para debatir sobre esta y otras cuestiones controvertidas.

Las finalidades que persigue el estudio son básicamente tres:

- Obtener una tasa general de reincidencia y unas tasas parciales de cada una de las medidas o intervenciones que se han llevado a cabo en la Dirección General de Justicia Juvenil Estas tasas han de contribuir en el futuro a la avaluación del sistema de justicia juvenil en Catalunya.
- Identificar los factores o variables que mejor expliquen y predigan el riesgo de reincidencia, para que la DGJJ pueda centrar el asesoramiento y la evaluación continua de los casos que permita obtener la información necesaria sobre estas variables más influyentes, y a la vez le permita centrar más los esfuerzos en la intervención sobre estos factores que contribuyen a reducir la reincidencia.
- Conocer el perfil de la población que esta llegando actualmente a la justicia de menores, una vez se llevan tres años de aplicación de la Ley 5/2000, para ofrecer datos cuantitativos y objetivos que faciliten a la DGJJ la creación de programas de intervención y de recursos adaptados a estos perfiles.

c. SENAME Ministerio de justicia de CHILE en el año 2015 presento su estudio sobre: **REINCIDENCIA DE LOS JOVENES INFRACTORES**

El presente estudio es una actualización de las tasas de reincidencia de jóvenes y/o adolescentes sometidos a sanciones privativas de libertad y medio libre, egresados desde el año 2009 hasta el año 2013. Desde el año 2012 estas tasas han sido calculadas periódicamente por el Departamento de Justicia Juvenil del Servicio Nacional de Menores. Esta vez, la publicación del Estudio de Reincidencia 2015 fue abordada por la Unidad de Estudios y pondrá en vigencia las tasas de las cohortes 2012 y

2013, pues existe un periodo de tiempo de uno y/o dos años para la observación y seguimiento del eventual comportamiento reincidente.

Esta actualización permitirá dilucidar un perfil sobre las características de los jóvenes envueltos en el circuito de la justicia de menores, así como también desarrollar un análisis comparativo del comportamiento entre las distintas cohortes, a ocho años de entrada en vigencia de la ley 20.084 y a cuatro años de la primera medición de las tasas de reincidencia a nivel nacional. Para llevar a cabo dicha investigación se contrastó la totalidad de población egresada de las sanciones de responsabilidad penal juvenil, con el catastro de nuevas condenas almacenadas en los registros de la Corporación Administrativa del Poder Judicial – CAPJ1. Tal comportamiento se analizó longitudinalmente con un seguimiento de 12 y 24 meses posteriores al egreso, por lo que en el actual estudio se han utilizado datos anteriores al año 2014.

El informe buscó analizar las características de los sujetos y de las cohortes mencionadas, conjuntamente se explotaron ciertas nociones del desempeño de los programas de intervención, tomando en cuenta los fundamentos de responsabilización y reinserción social que están detrás de ellos y los orientan.

Si bien los estudios de este tipo son ampliamente utilizados por los sistemas de justicia para evaluar intervenciones que buscan disminuir la criminalidad y fomentar la reintegración. Cabe considerar que las tasas de reincidencia se ven también influidas por una serie de complejos procesos sociales, psicológicos, valóricos y jurídicos que pueden escapar al ámbito de análisis de este indicador cuantitativo.

CAPITULO II

PLANTEAMIENTO DEL PROBLEMA

2.1. Descripción del problema

En la sociedad la mala formación de los jóvenes conlleva a que estos tengan una vida desordenada, mayormente los jóvenes pertenecen a grupos sociales que no dan buena influencia en su vida como pandillas, etc.Estos mismos jóvenes cometen delitos al estar desorientados por los numerosos problemas que se presentan en el hogar, con los amigos, su barrio, el colegio, etc. La falta de recursos económicos influye también a que estos jóvenes cometan delitos constantemente para poder satisfacer sus necesidades.La mala formación de estos conlleva a que cuando sean adultos cometan delitos más graves, y más aún que solo se dediquen a delinquir.

Actualmente los jóvenes infractores están intervenidos en centros juveniles a nivel nacional, pero a nivel general estos centros juveniles cada día se están convirtiendo en una preparatoria para que cuando salguen ya estén listos para ingresar a un centro penitenciario.

Para poder definir bien estos acontecimientos, de manera general nos preguntamos ¿porque los jóvenes cometen delitos?, ¿Qué factores influyen en ellos para que sucedan este tipo de situaciones?, son preguntas que varios países ya están resolviendo, un ejemplo es el caso de España- Catalunya que relazaron “estudio sobre los factores que influyen en la reincidencia en los jóvenes infractores - 2013”, este estudio se realizo por el centro de estudios de Catalunya en el año 2013. Este presente estudio tendrá como base al primer censo nacional 2016, que se realizo a los centros juveniles.

2.2. FORMULACIÓN DEL PROBLEMA

En la actualidad existen centros juveniles para recluir a los jóvenes que cometen delitos graves como robo, hurto, homicidio, asesinato, narcoactividad, violaciones, abusos deshonestos, secuestro, los mismos carecen de una administración adecuada, y sobre todo de espacios requeridos para el adecuado funcionamiento de sus instalaciones. Al mismo tiempo la descomposición social y de personas hace que el gobierno implante operativos para reducir la criminalidad que impera en la actualidad, causan el arresto mayoritario de estos jóvenes.

Así mismo para dar una posible solución al problema se buscara predecir la reincidencia de los jóvenes infractores utilizando información del censo 2016, por eso se plantea lo siguiente.

2.2.1. Problema general

¿Cuál es la técnica más adecuada entre redes neuronales y árbol de decisiones para la predicción de reincidencia de jóvenes infractores usando información del censo 2016?

2.2.2. Problemas específicos

- ✓ ¿Cuáles son los indicadores de especificidad, sensibilidad, exactitud y curva ROC para la predicción de reincidencia de jóvenes infractores aplicando redes neuronales?

- ✓ ¿Cuáles son los indicadores de especificidad, sensibilidad, exactitud y curva ROC para la predicción de reincidencia de jóvenes infractores aplicando redes neuronales?

2.3. OBJETIVOS DE INVESTIGACIÓN

2.3.1. Objetivo General

Elegir la técnica más adecuada entre redes neuronales y árbol de decisiones para la predicción de reincidencia de jóvenes infractores usando información del censo 2016.

2.3.2. Objetivos Específicos

- ✓ Determinar los indicadores de especificidad, sensibilidad, exactitud, y curva ROC para la predicción de reincidencia de jóvenes infractores aplicando redes neuronales.

- ✓ Determinar los indicadores de especificidad, sensibilidad, exactitud, y curva ROC para la predicción de reincidencia de jóvenes infractores aplicando Árbol de Decisiones.

2.4. PLANTEAMIENTO DE HIPÓTESIS

2.4.1. Hipótesis general

La técnica más adecuada es el Árbol de decisión, ya que presenta una variación porcentual mayor al 3%, con respecto a sus indicadores.

2.4.2. Hipótesis especifica:

- ✓ Los indicadores de especificidad, sensibilidad, exactitud, y curva ROC son mayores en un 60 %, para la predicción de reincidencia de jóvenes infractores aplicando redes neuronales.
- ✓ Los indicadores de especificidad, sensibilidad, exactitud, y curva ROC son mayores en un 60 %, para la predicción de reincidencia de jóvenes infractores aplicando árbol de decisiones.

2.5. JUSTIFICACIÓN.

Los jóvenes en la actualidad para ganar reconocimiento, ser aceptados en grupos no necesariamente buenos (pandillas), etc., tienden a cometer imprudencias. La mala formación de los padres, la falta de comunicación, la desorientación hacen que los jóvenes cometan delitos constantemente, aumentando la tasa de victimización en el país.

Una de las necesidades más importantes de la sociedad es reducir la cantidad de victimización que va en aumento año tras año, los diferentes factores que influyen en la recurrencia de un delito en los jóvenes, traerá un panorama más amplio para poder llegar a la raíz del problema, aplicando políticas sociales para poder recuperar a estos jóvenes, al mismo tiempo ayudar a los jóvenes que no están internados en los centros juveniles pero pueden estar a punto de cometer un delito.

Muchos estudios de otros países encuentran la solución en base a todas las investigaciones realizadas con respecto a esta problemática, por ejemplo España en su estudio de “tasas de reincidencia 2011 de justicia juvenil” año 2012, autores: Área de investigación y formación social y criminológica de Catalunya, plantearon políticas sociales ya que la cantidad de jóvenes que cometían delitos estaba en aumento, además año tras año eran jóvenes de menor edad (de 10 a 14 años) los que cometían delitos graves.

Este trabajo se está desarrollando como una de las primeras investigaciones realizadas en nuestro país, ya que reciénse presentó el primer censo realizado en julio del 2016. A base de esto, este trabajo será una de las bases para tratar más a fondo problemas sociales aplicados a centros juveniles.

El poder judicial por su parte está realizando sus propias investigaciones que serán publicadas a fines de noviembre del 2016, ante esto, esté trabajo será un aporte que ellos tomaran como referencia para futuras investigaciones, con la finalidad de rehabilitar a nuestros jóvenes y disminuir la taza de victimización en nuestro país.

Además servirá para los investigadores que retomaran este tema, reforzando mas este estudio para hallar nuevas teorías para poder combatir estos problemas sociales.

2.6. MATRIZ DE CONSISTENCIA:

CUADRO 01

MATRIZ DE CONSISTENCIA

Formulación del Problema	Formulación de los objetivos	Formulación de la Hipótesis	
Problema General	**Objetivo General**	**Hipótesis General**	**Variable**
¿Cuál es la técnica más adecuada entre redes neuronales y árbol de decisiones para la predicción de reincidencia de jóvenes infractores usando información del censo 2016?	Elegir la técnica más adecuada entre redes neuronales y árbol de decisiones para la predicción de reincidencia de jóvenes infractores usando información del censo 2016.	La técnica más adecuada es el Árbol de decisión ya que presenta una predicción superior en 3%, con respecto a sus indicadores.	Identificación de las variables Variable dependiente: Reincidencia Variable independiente: factores que influyen en la reincidencia
Problemas Específicos	**Objetivos Específicos**	**Hipótesis Específicas**	**indicadores**
¿Cuáles son los indicadores de especificidad, sensibilidad, exactitud y curva ROC para la predicción de reincidencia de jóvenes infractores aplicando redes neuronales?	Determinar los indicadores de especificidad, sensibilidad, exactitud, y curva ROC para la predicción de reincidencia de jóvenes infractores aplicando redes neuronales.	Los indicadores de especificidad, sensibilidad, exactitud, y curva ROC son mayores en un 60 %, para la predicción de reincidencia de jóvenes infractores aplicando redes neuronales.	Identificación de las variables Variable dependiente: Reincidencia Variable independiente: características del delito, condiciones sociales y familiares, estadía en el penal, discriminación
¿Cuáles son los indicadores de especificidad, sensibilidad, exactitud y curva ROC para la predicción de reincidencia de jóvenes infractores aplicando redes neuronales?	Determinar los indicadores de especificidad, sensibilidad, exactitud, y curva ROC para la predicción de reincidencia de jóvenes infractores aplicando Árbol de Decisiones.	Los indicadores de especificidad, sensibilidad, exactitud, y curva ROC son mayores en un 60 %, para la predicción de reincidencia de jóvenes infractores aplicando árbol de decisiones.	variable dependiente reincidencia Variable independiente: escaparon de casa antes de los 15 años, consume tabaco, consume drogas

ELABORACION: PROPIA

CAPÍTULO III

MARCO TEÓRICO

3.1. DEFINICIONES GENERALES:

La delincuencia como tal, refleja una construcción cultural e histórica de una sociedad y tiempo determinado. La delincuencia ha sido definida como el fenómeno social constituido por el conjunto de las infracciones, contra las normas fundamentales de convivencia, producidas en un tiempo y lugar determinados (Herrero, 1997).

Concordamos con lo expuesto anteriormente, los jóvenes infractores en los últimos años están en aumento en la población juvenil, creciendo en un 7% aproximadamente año tras año (INEI censo 2016).

Para el estudio se está utilizando la **definición de reincidencia**, siendo esta definición tomada del Código penal: El art. 46-B del CP recoge un supuesto de reincidencia genérica y real. Es genérica, por cuanto el legislador no exige que el segundo delito sea de igual o semejante naturaleza, bastará con que se trate de un delito doloso. Es real, por cuanto se exige que se haya cumplido en todo o en parte la pena impuesta por el primer delito. En este punto, hemos de

criticar el hecho de que la Ley 30076 amplíe el ámbito de aplicación de la reincidencia, pues si antes el legislador tomaba como presupuesto el cumplimiento [total o parcial] de una condena a pena privativa de libertad, hoy este se extiende a cualquier tipo de pena. En la misma lógica de aplicar esta agravante cualificada a las faltas, el legislador sigue extendiendo esta figura a delitos de bagatela

Técnicas previas:

En un estudio deslizado por alumnos de la Universidad Nacional de Ingeniería para el año 2016, tomaron la variable reingreso a un centro penitenciario como variable respuesta, este estudio estuvo dirigido como población objetivo a todos los penitenciarios de lima y callao. Las técnicas que utilizaron fueron regresión logística para variables respuesta balanceada y no balanceada, punto de corte optimo, y adicional a ello presentaron árbol de decisiones para encontrar el perfil de los penitenciarios re ingresantes en lima y callao.

Técnica a usar:

Como se menciono en los antecedentes, la técnica para hallar el perfil es árbol de decisiones, pero aplicaremos la técnica de redes neuronales para poder comparar y al final poder trabajar con la mejor técnica de predicción.

Terminología básica:

3.2. PRUEBAS PRELIMINARES

Correlación de pearson:

Este grado o intensidad de relación entre dos variables continuas, se resume mediante un coeficiente de correlación que se conoce como “r de Pearson” en honor del matemático Karl Pearson (el mismo del coeficiente que mide la asimetría). Dicha técnica es válida solamente si es posible establecer los siguientes supuestos:

$$r_s = \frac{\sum_{i=1}^{n} XY - \frac{\sum_{i=1}^{n} X \sum_{i=1}^{n} Y}{n}}{\sqrt{\left(\sum_{i=1}^{n} X^2 - \frac{\left(\sum_{i=1}^{n} X\right)^2}{n}\right)\left(\sum_{i=1}^{n} Y^2 - \frac{\left(\sum_{i=1}^{n} Y\right)^2}{n}\right)}} =$$

Prueba de normalidad

Cuando la prueba Kolmogorov-Smirnovkolmogorov se aplica para contrastar la hipótesis de normalidad de la población, el estadístico de prueba es la máxima diferencia:

$$D = máx\,|F_n(x) - F_0(x)|$$

siendo Fn(x) la función de distribución muestral y Fo(x) la función teórica o correspondiente a la población normal especificada en la hipótesis nula.

La distribución del estadístico de Kolmogorov-Smirnov es independiente de la distribución poblacional especificada en la hipótesis nula y los valores críticos de este estadístico están tabulados. Si la distribución postulada es la normal y se estiman sus parámetros, los valores críticos se obtienen aplicando la corrección de significación propuesta por Lilliefors.

3.3. ÁRBOL DE DECISIONES

El procedimiento Árbol de decisión crea un modelo de clasificación basado en árboles y clasifica casos en grupos o pronostica valores de una variable (criterio) dependiente basada en valores de variables independientes (predictores). El procedimiento proporciona herramientas de validación para análisis de clasificación exploratorios y confirmatorios.

El procedimiento se puede utilizar para:

Segmentación. Identifica las personas que pueden ser miembros de un grupo específico.

Estratificación. Asigna los casos a una categoría de entre varias, por ejemplo, grupos de alto riesgo, bajo riesgo y riesgo intermedio.

Predicción. Crea reglas y las utiliza para predecir eventos futuros, como la verosimilitud de que una persona cause mora en un crédito o el valor de reventa potencial de un vehículo o una casa.

Reducción de datos y clasificación de variables. Selecciona un subconjunto útil de predictores a partir de un gran conjunto de variables para utilizarlo en la creación de un modelo paramétrico formal.

Identificación de interacción. Identifica las relaciones que pertenecen sólo a subgrupos específicos y las especifica en un modelo paramétrico formal.

Fusión de categorías y discretización de variables continuas. Vuelve a codificar las variables continuas y las categorías de los predictores del grupo, con una pérdida mínima de información.

Ejemplo. Un banco desea categorizar a los solicitantes de créditos en función de si representan o no un riesgo crediticio razonable. Basándose en varios factores, incluyendo las valoraciones del crédito conocidas de clientes anteriores, se puede generar un modelo para pronosticar si es probable que los clientes futuros causen mora en sus créditos.

Un análisis basado en árboles ofrece algunas características atractivas:

- Permite identificar grupos homogéneos con alto o bajo riesgo.
- Facilita la creación de reglas para realizar predicciones sobre casos individuales.

Consideraciones de los datos

Datos. Las variables dependientes e independientes pueden ser:

- Nominal. Una variable puede ser tratada como nominal cuando sus valores representan categorías que no obedecen a una clasificación intrínseca. Por ejemplo, el departamento de la compañía en el que trabaja un empleado. Algunos ejemplos de variables nominales son: región, código postal o confesión religiosa.
- Ordinal. Una variable puede ser tratada como ordinal cuando sus valores representan categorías con alguna clasificación intrínseca. Por ejemplo, los niveles de satisfacción con un servicio, que abarquen desde muy insatisfecho hasta muy satisfecho. Entre los ejemplos de variables ordinales se incluyen escalas de actitud que representan el grado de satisfacción o confianza y las puntuaciones de evaluación de las preferencias.
- Escalas. Una variable puede tratarse como escala (continua) cuando sus valores representan categorías ordenadas con una métrica con significado, por lo que son adecuadas las comparaciones de distancia entre valores. Son ejemplos de variables de escala: la edad en años y los ingresos en dólares.

Ponderaciones de frecuencia Si se encuentra activada la ponderación, las ponderaciones fraccionarias se redondearán al número entero más cercano; de esta manera, a los casos con un valor de ponderación menor que 0,5 se les asignará una ponderación de 0 y, por consiguiente, se verán excluidos del análisis.

Supuestos. Este procedimiento supone que se ha asignado el nivel de medición adecuado a todas las variables del análisis; además, algunas características suponen que todos los valores de la variable dependiente incluidos en el análisis tienen etiquetas de valor definidas.

- Nivel de medición. El nivel de medición afecta a los tres cálculos; por lo tanto, todas las variables deben tener asignado el nivel de medición

adecuado. De forma predeterminada, se supone que las variables numéricas son de escala y que las variables de cadena son nominales, lo cual podría no reflejar con exactitud el verdadero nivel de medición. Un icono junto a cada variable en la lista de variables identifica el tipo de variable.

Puede cambiar de forma temporal el nivel de medición de una variable; para ello, pulse con el botón derecho del ratón en la variable en la lista de variables de origen y seleccione un nivel de medición del menú emergente.

- Etiquetas de valor. La interfaz del cuadro de diálogo para este procedimiento supone que, o todos los valores no perdidos de una variable dependiente categórica (nominal, ordinal) tienen etiquetas de valores definidas, o que ninguno de ellos las tienen. Algunas características no estarán disponibles a menos que como mínimo dos valores no perdidos de la variable dependiente categórica tengan etiquetas de valor. Si al menos dos valores no perdidos tienen etiquetas de valor definidas, todos los demás casos con otros valores que no tengan etiquetas de valor se excluirán del análisis.

Métodos de crecimiento

Los métodos de crecimiento disponibles son:

- *CHAID*. Detección automática de interacciones mediante chi-cuadrado (CHi-squareAutomaticInteractionDetection). En cada paso, CHAID elige la variable independiente (predictora) que presenta la interacción más fuerte con la variable dependiente. Las categorías de cada predictor se funden si no son significativamente distintas respecto a la variable dependiente.

- *CHAID exhaustivo*. Una modificación del CHAID que examina todas las divisiones posibles de cada predictor.

- *CRT*. Árboles de clasificación y regresión (Classification and RegressionTrees). CRT divide los datos en segmentos para que sean lo más homogéneos que sea posible respecto a la variable dependiente. Un nodo terminal en el que todos los casos toman el mismo valor en la variable dependiente es un nodo homogéneo y "puro".

- *QUEST*. Árbol estadístico rápido, insesgado y eficiente (Quick, Unbiased, EfficientStatisticalTree). Método rápido y que evita el sesgo que presentan otros métodos al favorecer los predictores con muchas categorías. Sólo puede especificarse QUEST si la variable dependiente es nominal.

MÉTRICAS

Los algoritmos para la construcción de árboles de decisión suelen trabajar de manera top-down, escogiendo en cada paso la variable que mejor divide el conjunto de elementos.11 Diferentes algoritmos utilizan diferentes métricas para medir el "mejor". Estos miden generalmente la homogeneidad de la variable de destino dentro de los subconjuntos. Algunos ejemplos se dan a continuación. Estas métricas se aplican a cada subconjunto candidato, y los valores resultantes se combinan (por ejemplo, un promedio) para proporcionar una medida de la calidad de la división.

REDUCCIÓN DE LA VARIANZA

Introducido en ACR,3 la reducción de la varianza se emplea a menudo en los casos en que la variable de destino es un árbol de regresión continuo, lo que significa que el uso de muchas otras métricas requeriría primero discretización antes de ser aplicada. La reducción de la varianza de un nodo N se define como la reducción total de la varianza de la variable destino “x” debido a la partición en este nodo.

IMPUREZA DEL GINI

No debe confundirse con el coeficiente de Gini. Utilizado por el algoritmo de ACR (Árboles de Clasificación y Regresión) , la impureza de Gini es una medida de cuán a menudo un elemento elegido aleatoriamente del conjunto sería etiquetado incorrectamente si fue etiquetado de manera aleatoria de acuerdo a la distribución de las etiquetas en el subconjunto. La impureza de Gini se puede calcular sumando la probabilidad de cada elemento siendo veces elegido la probabilidad de un error en la categorización de ese elemento. Alcanza su mínimo (cero) cuando todos los casos del nodo corresponden a una sola categoría de destino.

$$I_G(f) = \sum_{i=1}^{m} f_i(1 - f_i) = \sum_{i=1}^{m} (f_i - {f_i}^2) = \sum_{i=1}^{m} f_i - \sum_{i=1}^{m} {f_i}^2 = 1 - \sum_{i=1}^{m} {f_i}^2$$

Ganancia de información: entropía

Intenta maximizar la ganancia de información conseguida por el uso del atributo Ai para ramificar el árbol de decisión mediante la minimización de la función I:

$$I(A_i) = \sum_{j=1}^{M_i} p(A_{ij}) H(C|A_{ij})$$

Donde Ai es el atributo utilizado para ramificar el árbol, Mi es el número de valores diferentes del atributo Ai , p(Aij) es la probabilidad de que el atributo Ai tome su j-esimo valor y H(C|Aij) es la entropía de clasificación del conjunto de ejemplos en los que el atributo Ai toma su j-ésimo valor. Esta entropía de clasificación se define como

$$H(C|A_{ij}) = -\sum_{k=1}^{J} p(C_k|A_{ij}) \log_2 p(C_k|A_{ij})$$

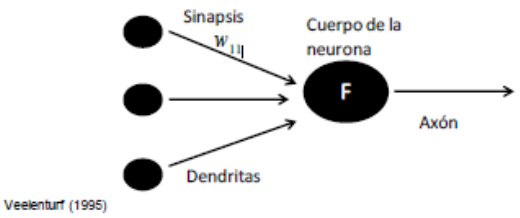
3.4. REDES NEURONALES:

Definición

La mayoría de investigadores definen a las redes neuronales como modelos artificiales y simplificados del cerebro humano, capaces de aprender a través de sus sistemas interconectados, y que tienen como unidades elementales a los nodos que vendrían a simbolizar las unidades básicas del cerebro humano, **las neuronas**.

En la figura 2.1 se puede visualizar una analogía entre los componentes de una neurona y un modelo de red neuronal. Las **dendritas** en una neurona son las encargadas de recibir la información proveniente de otras neuronas, luego esta información en procesada en el **cuerpo de la neurona** y la respuesta resultante es enviada hacia otras neuronas a través del **axón**, este traspaso de información se hace a través de un impulso eléctrico que determina el grado de excitación de la neurona, denominado sinapsis. En la red neuronal el proceso es similar, para explicar esto se ha mostrado una red simple, donde las dendritas son las conexiones que se dan desde los valores de entrada (*inputs*) hacia el nodo de la capa oculta, este nodo viene hacer el cuerpo de la neurona que por medio de una función matemática procesa la información la cual finalmente da una salida (*output*) , esto a través de la conexión del nodo hacia afuera, esto sería equivalente al axón, finalmente lafuerza con que se traspasa la información (sinapsis) es representada por los pesos de la red neuronal.

FIGURA 01
PROCESO DE LA RED



FUENTE: Departamento de Ciencias de la Computación e Inteligencia Artificial, Paris

Arquitectura de Redes Neuronales

Giudici (2003) define al término arquitectura como la organización de la red neuronal: el número de capas, el número de de unidades (neuronas) que siguen en cada capa, y la manera en que éstas son conectadas.

Respecto al tipo de capas que puede tener una red neuronal, existen tres tipos:

- Input: Son las encargadas de recibir sólo la información del ambienteexterno, cada neurona en este caso corresponde a una variableexplicatoria, en esta capa no se realiza ningún tipo de cálculo.
- Output: Es la capa que produce los resultados finales, los cuales sonenviados al ambiente exterior.
- Oculta: Son las capas que se encuentran entre las capas *input* y *output*,y reciben este nombre porque no tienen contacto con el ambienteexterno, son capas utilizadas exclusivamente para el análisis.

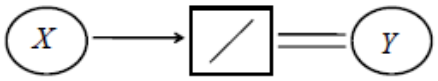
La arquitectura de la red neuronal por lo general es presentadagráficamente, por esta razón en ocasiones recibe el nombre de “topología de lared”. Sarle (1994) señala como poder graficar redes neuronales, las indicaciones que señala son:

- Círculos representan variables observadas: Con el nombre mostrado dentro del círculo.
- Cuadrados representan valores calculados con una función de uno o más argumentos (Ver 2.1.3). El símbolo dentro de la caja indica el tipo de función. La mayoría de cajas también disponen del parámetro bias(Ver 2.1.4).

- Las flechas indican de donde provienen los recursos que servirán de argumentos para la función, por lo general están asociados a pesos que deben ir ajustándose conforme la red vaya aprendiendo (Ver 2.1.4).
- Dos líneas paralelas indican que el resultado final ha sido calculado poralgún método de estimación.

Un ejemplo de topología, se muestra en la figura 2.2, esta red neuronalsimboliza o equivale a una regresión lineal simple:

FIGURA 02
REPRESENTACION DE REGRESION LINEAL SIMPLE EN UN ESQUEMA DE REDES



FUENTE: Departamento de Ciencias de la Computación e Inteligencia Artificial, Paris

Al trabajar con redes neuronales los términos de función de transferencia, pesos, bias, tipos de aprendizaje, ratio de aprendizaje son comunes, por ello a continuación se explicará cada uno con el fin de tener unamejor comprensión.

Función de entrada

La neurona trata a muchos valores de entrada como si fueran uno solo; esto recibe el nombre de entrada global. Por lo tanto, ahora nos enfrentamos al problema de cómo se pueden combinar estas simples entradas (ini1, ini2, ...) dentro de la entrada global, gini. Esto se logra a través de la función de entrada, la cual se calcula a partir del vector entrada. La función de entrada puede describirse como sigue:

$$input_i = (in_{i1} \bullet w_{i1}) * (in_{i2} \bullet w_{i2}) * \ldots (in_{in} \bullet w_{in})$$

Donde: * representa al operador apropiado (por ejemplo: máximo, sumatoria, productora, etc.), n al número de entradas a la neurona Ni y wi al peso.

Algunas de las funciones de entrada más comunes son:

- **Sumatoria de las entradas pesadas**: es la suma de todos los valores de entrada a la neurona, multiplicados por sus correspondientes pesos.

$$\sum_j (n_{ij}\, w_{ij}), \quad \text{con } j = 1, 2, ..., n$$

- **Productoria de las entradas pesadas:** es el producto de todos los valores de entrada a la neurona, multiplicados por sus correspondientes pesos.

$$\prod_j (n_{ij}\, w_{ij}), \quad \text{con } j = 1, 2, ..., n$$

- **Máximo de las entradas pesadas:** solamente toma en consideración el valor de entrada más fuerte, previamente multiplicado por su peso correspondiente.

$$\underset{j}{\text{Max}}(n_{ij}\, w_{ij}) \quad \text{con } j = 1, 2, ..., n$$

Función de Activación

Una neurona biológica puede estar activa (excitada) o inactiva (no excitada); es decir, que tiene un “estado de activación”. Las neuronas artificiales también tienen diferentes estados de activación; algunas de ellas solamente dos, al igual que las biológicas, pero otras pueden tomar cualquier valor dentro de un conjunto determinado. La función activación calcula el estado de actividad de una neurona; transformando la entrada global (menos el umbral, Θi) en un valor (estado) de activación, cuyo rango normalmente va de (0 a 1) o de (–1 a 1). Esto es así, porque una neurona puede estar totalmente inactiva (0 o –1) o activa (1). La función activación, es una función de la entrada global (gini) menos el umbral (Θi). Las funciones de activación más comúnmente utilizadas se detallan a continuación:

- **Función lineal**

$$f(x) = \begin{cases} -1 & x \leq -1/a \\ a * x & -1/a < x < 1/a \\ 1 & x \geq 1/a \end{cases}$$

$$\text{con } x = gin_i - \Theta_i, \text{ y } a > 0.$$

Los valores de salida obtenidos por medio de esta función de activación serán: a·(gini - Θi), cuando el argumento de (gini - Θi) esté comprendido dentro del rango (-1/a, 1/a). Por encima o por debajo de esta zona se fija la salida en 1 o –1, respectivamente. Cuando a = 1 (siendo que la misma afecta la pendiente de la gráfica), la salida es igual a la entrada.

- **Función sigmoidea**

$$f(x) = \frac{1}{1 + e^{-g\,x}}, \text{ con } x = gin_i - \Theta_i.$$

Los valores de salida que proporciona esta función están comprendidos dentro de un rango que va de 0 a 1. Al modificar el valor de g se ve afectada la pendiente de la función de activación.

- **Función tangencial hiperbólica**

$$f(x) = \frac{e^{g\,x} - e^{-g\,x}}{e^{g\,x} + e^{-g\,x}}, \text{con } x = gin_i - \Theta_i.$$

Los valores de salida de la función tangente hiperbólica están comprendidos dentro de un rango que va de -1 a 1. Al modificar el valor de g se ve afectada la pendiente de la función de activación.

Para explicar porque se utilizan estas funciones de activación se suele emplear la analogía a la aceleración de un automóvil. Cuando un auto inicia su movimiento necesita una potencia elevada para comenzar a acelerar. Pero al ir tomando velocidad, este demanda un menor incremento de dicha potencia para mantener la aceleración. Al llegar a altas velocidades, nuevamente un amplio

incremento en la potencia es necesario para obtener una pequeña ganancia de velocidad. En resumen, en ambos extremos del rango de aceleración de un automóvil se demanda una mayor potencia para la aceleración que en la mitad de dicho rango.

Función de salida

El último componente que una neurona necesita es la función de salida. El valor resultante de esta función es la salida de la neurona i (outi); por ende, la función de salida determina que valor se transfiere a las neuronas vinculadas. Si la función de activación está por debajo de un umbral determinado, ninguna salida se pasa a la neurona subsiguiente. Normalmente, no cualquier valor es permitido como una entrada para una neurona, por lo tanto, los valores de salida están comprendidos en el rango [0, 1] o [-1, 1]. También pueden ser binarios {0, 1} o {-1, 1}.

Dos de las funciones de salida más comunes son:

- Ninguna: este es el tipo de función más sencillo, tal que la salida es la misma que la entrada. Es también llamada función identidad.
- Binaria $\begin{cases} 1 & \text{si act}_i \geq \xi_i \\ 0 & \text{de lo contrario} \end{cases}$, donde ξ_i es el umbral.

3.5. SENSIBILIDAD Y ESPECIFICIDAD

Generalmente, la exactitud diagnóstica se expresa como sensibilidad y especificidad diagnósticas. Cuando se utiliza una prueba dicotómica (una cuyos resultados se puedan interpretar directamente como positivos o negativos), la sensibilidad es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea el definido como positivo respecto a la condición que estudia la prueba, razón por la que también es denominada fracción de verdaderos positivos (FVP). La especificidad es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea el definido como negativo. Es igual al

resultado de restar a uno la fracción de falsos positivos (FFP). Cuando los datos de una muestra de pacientes se clasifican en una tabla de contingencia por el resultado de la prueba y su estado respecto a la enfermedad, es fácil estimar a partir de ella la sensibilidad y la especificidad de la prueba (tabla 1). Conviene insistir –ya que esta distinción aparecerá repetidamente en lo sucesivo– en que lo que realmente obtenemos son estimaciones de los verdaderos valores de sensibilidad y especificidad para una población teórica de la que suponemos que nuestro grupo de pacientes constituye una muestra aleatoria. Por tanto, un tratamiento estadístico correcto de cantidades como las calculadas por el método descrito por la tabla 1 exigiría incluir medidas de su precisión como estimadores, y, mejor aún, utilizarlas para construir intervalos de confianza para los verdaderos valores de sensibilidad y especificidad.

CUADRO 02
CUADRO DE INDICADORES

		Verdadero Diagnóstico	
		Enfermo	**Sano**
Resultado de la Prueba	**Prueba Positiva**	Verdadero Positivo (VP)	Falso Positivo (FP)
	Prueba Negativa	Falso Negativo (FN)	Verdadero Negativo (VN)
		VP + FN	**VN + FP**
Sensibilidad	**= VP/(VP + FN) = FVP (fracción de verdaderos positivos)**		
Especificidad	**= VN/(VN + FP) = FVN (fracción de verdaderos negativos) = 1 - FFP (fracción de falsos positivos)**		

ELABORACION: PROPIA

3.6. LA CURVA ROC

La limitación principal del enfoque hasta ahora expuesto estribaría en nuestra exigencia de que la respuesta proporcionada por la prueba diagnóstica sea de tipo dicotómico, por lo que en principio quedaría excluida la amplia gama de pruebas diagnósticas cuyos resultados se miden en una escala (nominalmente) continua o, al menos, discreta ordinal. Piénsese, por ejemplo, respecto al primer tipo en la determinación de la glucosa sérica por el laboratorio o, respecto al segundo, en una prueba realizada por el Servicio de Radiología en que los resultados se expresen empleando las categorías "seguramente normal",

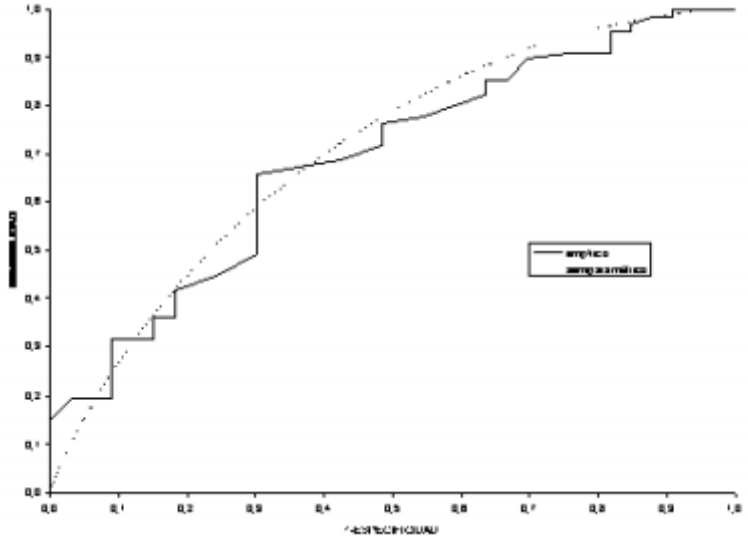
"probablemente normal", "dudoso", "probablemente anormal" y "seguramente anormal". La generalización a estas situaciones se consigue mediante la elección de distintos niveles de decisión o valores de corte que permitan una clasificación dicotómica de los valores de la prueba según sean superiores o inferiores al valor elegido. La diferencia esencial con el caso más simple es que ahora contaremos no con un único par de valores de sensibilidad y especificidad que definan la exactitud de la prueba, sino más bien con un conjunto de pares correspondientes cada uno a un distinto nivel de decisión. Este procedimiento constituye la esencia del análisis ROC, una metodología desarrollada en el seno de la Teoría de la Decisión en los años 50 y cuya primera aplicación fue motivada por problemas prácticos en la detección de señales por radar (aunque el detalle pueda parecer anecdótico, la equivalencia entre el operador que interpreta los picos en la pantalla del radar para decidir sobre la presencia de un misil y el médico que emplea el resultado de una prueba diagnóstica para decidir sobre la condición clínica del paciente, es completa 1). La aparición del libro de Swets y Pickett 2 marcó el comienzo de su difusión en el área de la Biomedicina, inicialmente en Radiología, donde la interpretación subjetiva de los resultados se recoge en una escala de clasificación, pero de modo creciente en relación con cualquier método diagnóstico que genere resultados numéricos.

3.6.1. Métodos de cálculo de la curva ROC

Un primer grupo de métodos para construir la curva ROC lo constituyen los llamados métodos no paramétricos. Se caracterizan por no hacer ninguna suposición sobre la distribución de los resultados de la prueba diagnóstica. El más simple de estos métodos es el que suele conocerse como empírico, que consiste simplemente en representar todos los pares (FFP, FVP) – es decir todos los pares (1-especificidad, sensibilidad) – para todos los posibles valores de corte que se puedan considerar con la muestra particular de que dispongamos. Desde un punto de vista técnico, este método sustituye las funciones de distribución teóricas por una estimación no paramétrica de ellas, a

saber, la función de distribución empírica construida a partir de los datos. Informalmente, es como si en la figura 1 sustituyéramos las funciones de densidad por histogramas obtenidos a partir de la muestra de pacientes sanos y enfermos y construyéramos la curva ROC a partir de ellos. En la figura 4 se representa la curva ROC obtenida por el método empírico para un conjunto de datos obtenidos en un grupo de pacientes investigados con el fin de establecer un diagnóstico de anemia ferropénica mediante la determinación del volumen corpuscular medio (ver apartado a) del apéndice). El verdadero diagnóstico se establece empleando como gold standard el examen de la médula ósea. La representación obtenida por este método tiene forma aproximadamente en escalera. En efecto, para cada variación mínima del valor de corte que produzca cambios en sensibilidad o especificidad, al menos un caso pasa a ser considerado bien como verdadero positivo, lo que se corresponde con un trazo vertical, bien como falso positivo, lo que da lugar a un trazo horizontal. Existe aún otra posibilidad, derivada de la posibilidad de que se produzcan empates, es decir, dos o más casos con el mismo valor de la prueba: si el empate ocurre entre un caso del grupo enfermo y otro del grupo sano aparecerá un trazo diagonal en la representación.

FIGURA 12
CURVA ROC



FUENTE: Departamento de Ciencias de la Computación e Inteligencia Artificial, Paris

3.7. ERROR CUADRÁTICO MEDIO

En estadística, el error cuadrático medio (ECM) de un estimador mide el promedio de los errores al cuadrado, es decir, la diferencia entre el estimador y lo que se estima. El ECM es una función de riesgo, correspondiente al valor esperado de la pérdida del error al cuadrado o pérdida cuadrática. La diferencia se produce debido a la aleatoriedad o porque el estimador no tiene en cuenta la información que podría producir una estimación más precisa.1

El ECM es el segundo momento (sobre el origen) del error, y por lo tanto incorpora tanto la varianza del estimador así como su sesgo. Para un estimador insesgado, el ECM es la varianza del estimador. Al igual que la varianza, el ECM tiene las mismas unidades de medida que el cuadrado de la cantidad que se estima. En una analogía con la desviación estándar, tomando la raíz cuadrada del ECM produce el error de la raíz cuadrada de la media o la desviación de la raíz cuadrada media (RMSE o RMSD), que tiene las mismas unidades que la cantidad que se estima; para un estimador insesgado, el RMSE es la raíz cuadrada de la varianza, conocida como la desviación estándar.

Si $\hat{Y}$ es un vector de n predicciones y Y es el vector de los verdaderos valores, entonces el (estimado) ECM del predictor es:

$$\mathrm{ECM} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2.$$

3.8. OVER-SAMPLING

Este método replica las observaciones de la clase minoritaria con el fin de que se tenga una proporción similar a la otra clase y así obtener un conjunto de datos balanceados. También es conocido como upsampling.

CAPÍTULO IV

METODOLOGÍA

4.1. Tipo de investigación:

El tipo de investigación es exploratoria ya que es considerada como primer acercamiento científico a un problema de la realidad peruana, ya que recién se cuenta con el primer censo nacional 2016 realizado a los jóvenes infractores.

4.2. Nivel de investigación:

El nivel de la investigación tiene alcance descriptivo y predictivo

4.3. Diseño de la investigación:

El tipo de diseño es no experimental de carácter transversal

4.4. Población en estudio:

La población de estudio a tratar consta de 2,023 jóvenes infractores que están internados en 10 centros juveniles a nivel nacional.

4.5. Unidad de análisis

Son las variables tomadas en el censo nacional a jóvenes infractores 2016, en los 10 centros juveniles a nivel nacional

4.6. Fuentes de información:

CENSO NACIONAL DE POBLACIÓN EN LOS CENTROS JUVENILES DE DIAGNÓSTICO Y REHABILITACIÓN, 2016

OBJETIVOS

Obtener información estadística sobre la población de los (las) adolescentes con medida de internamiento en los Centros Juveniles de Diagnóstico y Rehabilitación que sirva para elaborar políticas públicas de prevención de las infracciones y conductas de riesgo, orientadas a la reeducación, rehabilitación y reincorporación de esta población vulnerable a la sociedad.

CARACTERÍSTICAS TÉCNICAS

- MODALIDAD DE RECOJO DE INFORMACIÓN
  El levantamiento de la información se realizara a través de entrevistas en el centro juvenil donde sea ubicada.
- INSTRUMENTO DE RECOLECCIÓN
  Cédula censal impreso.
- PERIODO DE EJECUCIÓN
  El levantamiento de la información se tiene programado realizarlo del 28 de Marzo al 01 de abril de 2016.
- COBERTURA DEL CENSO

  Cobertura Geográfica

  El estudio se realizará en los departamentos del país donde se encuentran ubicados los Centros Juveniles de Diagnóstico y Rehabilitación.

  Cobertura Temática

El estudio se centra en las variables que permitirán conocer las condiciones por las cuales el interno(a) procedió a cometer infracciones. Para ello se ha tomado en cuenta los siguientes acápites y preguntas:

4.7. Diseño de muestreo y preparación de datos:

Se trabajara con la población a nivel nacional de los 10 centros juveniles.

CAPITULO V

RESULTADOS

Procedimiento estadístico:

5.1. Interpretación de variables influyentes:

Para la interpretación de las variables se utilizo las definiciones del Código Penal, además se tomaron las variables que tienen mayor influencia teórica con nuestra variable respuesta que es Reincidencia, Ejm: se descartóla variable tiempo de espera hasta que llego el abogado para revisar su caso, entre otros.

Consultando con especialistas, como abogados penalistas, sociólogos, etc. se descartaron variables por relevancia teórica ya que teniendo estudios anteriores como el de Catalunya “Estudio de reincidencia a los jóvenes infractores” – España 2015, no influían en la reincidencia.

5.2. Limpieza de datos

La limpieza de los datos se desarrollo de la siguiente manera:

- Los datos con demasiados valores perdidos se dicotomizó para poder trabajar estas variables y no perder información.

  Ejm: la variable; cuantas veces fue agredido por las personas que trabaja en los centros juveniles. Fue transformada a: Lo agredían en los centros juveniles.
- Se aplico imputación de datos a las variables que presentaban valores perdidos menores al 10%, y dependiendo a la naturaleza de las variables se reemplaza con la media, moda o mediana.
- Se re categorizo variables que tenían más de 10 categorías, reagrupándolas para poder tener una mejor visión de los resultados.
- Las variables de tipo cadena se eliminaron por la complejidad de frecuencias que presentan y además por la poca cantidad de datos que presenta (un total de 48 variables)

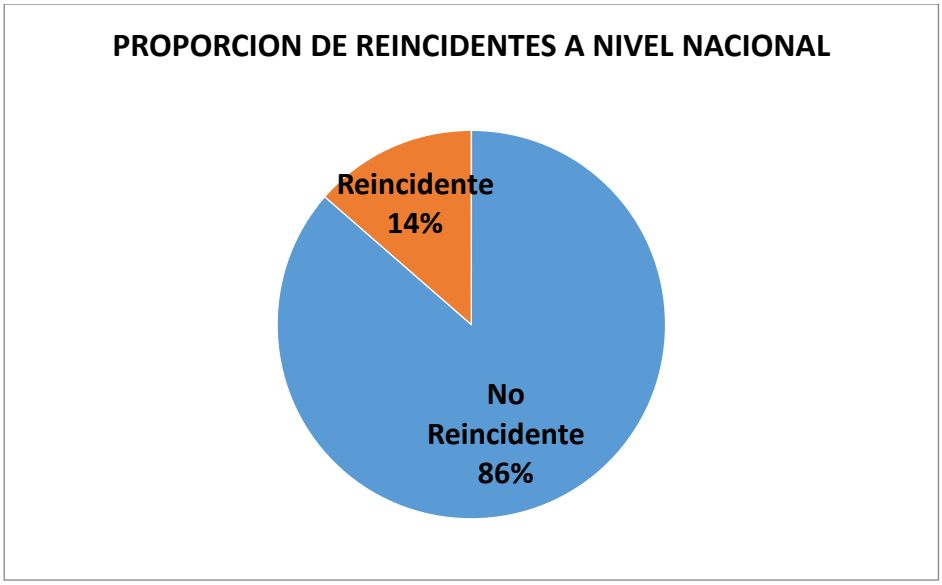
5.3. Análisis descriptivo:

Analizamos nuestra variable respuesta:

CUADRO xxx

Target	Frecuencia	Porcentaje
No Reincidente	1698	17.0%
Reincidente	267	2.7%
Total	1965	19.7%

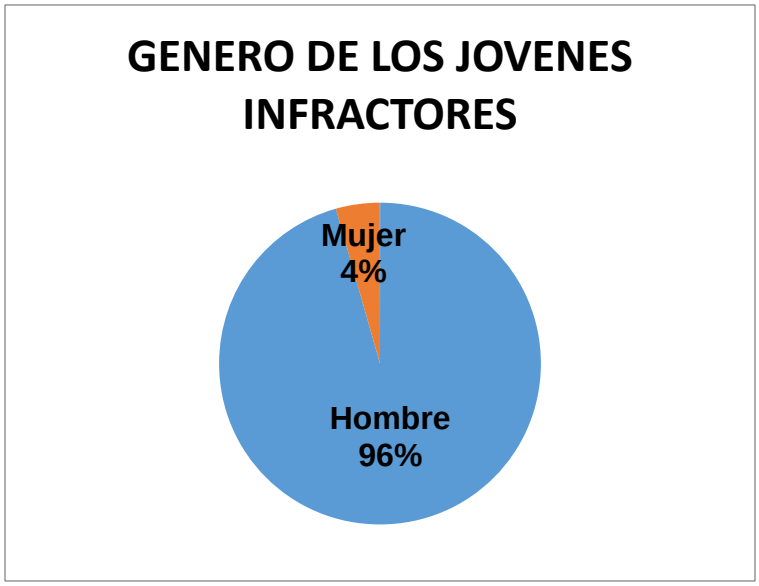
Del cuadro anterior se puede visualizar la cantidad de reincidentes a nivel nacional, el 13,6% son jóvenes reincidentes.



Fuente: INEI
Elaboración: Propia

Analizando por género

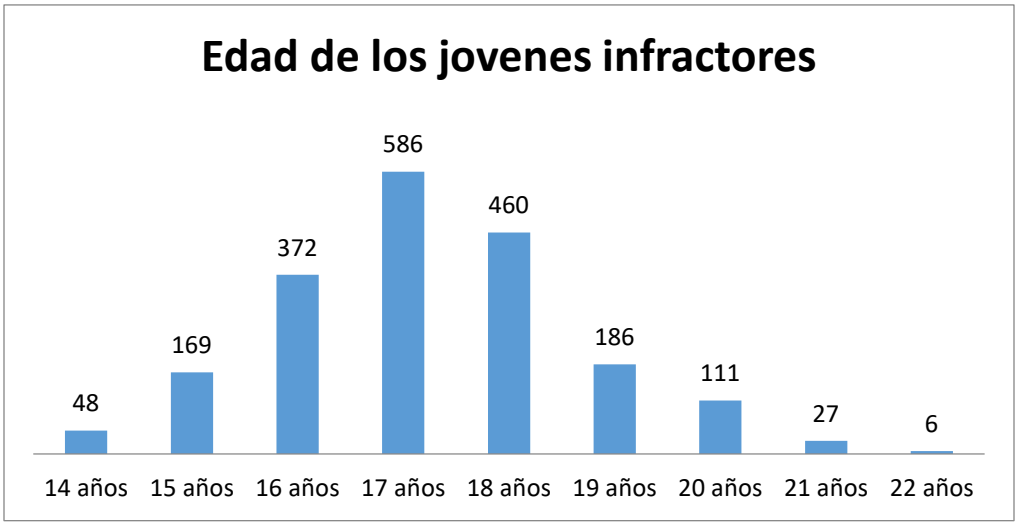
Genero	Frecuencia	Porcentaje
Hombre	1878	95.6%
Mujer	87	4.4%
Total	1965	100.0%



Del grafico podemos observar que la mayor cantidad de jóvenes infractores pertenecen al género merculino

Analizando por edad

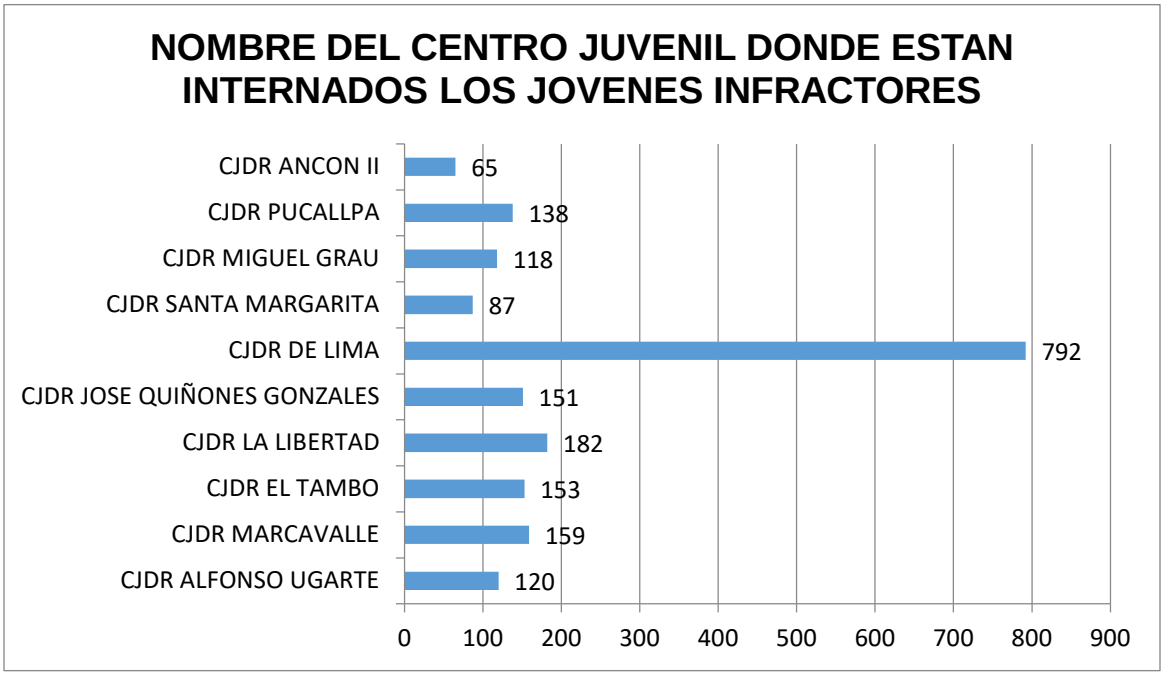
EDAD	Frecuencia	Porcentaje
14 años	48	2.4%
15 años	169	8.6%
16 años	372	18.9%
17 años	586	29.8%
18 años	460	23.4%
19 años	186	9.5%
20 años	111	5.6%
21 años	27	1.4%
22 años	6	0.3%
Total	1965	100.0%



Del grafico mostrado, la mayor concentración de jóvenes infractores internos en los centros juveniles tiene 17 años de edad y comprenden el 30% de la población de jóvenes infractores

Analizando por centro juvenil donde se encuentra internado el joven infractor

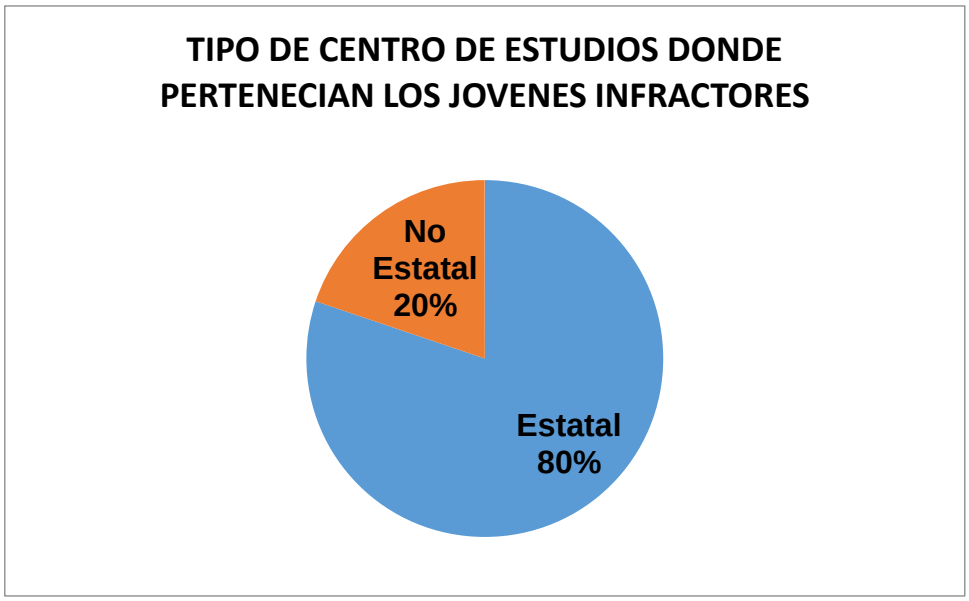
NOMBRE DEL CENTRO JUVENIL	Frecuencia	Porcentaje
CJDR ALFONSO UGARTE	120	6.1%
CJDR MARCAVALLE	159	8.1%
CJDR EL TAMBO	153	7.8%
CJDR LA LIBERTAD	182	9.3%
CJDR JOSE QUIÑONES GONZALES	151	7.7%
CJDR DE LIMA	792	40.3%
CJDR SANTA MARGARITA	87	4.4%
CJDR MIGUEL GRAU	118	6.0%
CJDR PUCALLPA	138	7.0%
CJDR ANCON II	65	3.3%
Total	1965	100.0%



Del grafico observado tenemos que la mayor cantidad de jóvenes infractores está internado en el centro juvenil de lima con el 40% de la población a nivel nacional.

Analizando por tipo de de centro de estudios donde pertenecieron los jóvenes infractores

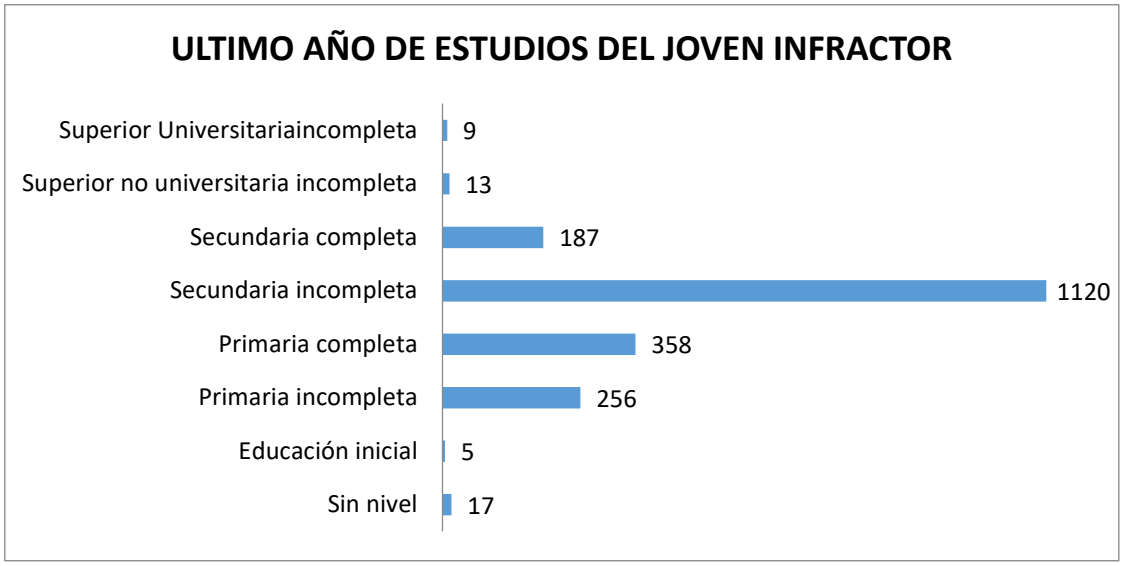
Tipo de centro de estudios	Frecuencia	Porcentaje
Estatal	1576	80.2%
No Estatal	389	19.8%
Total	1965	100.0%



Del grafico anterior el 80% de jóvenes infractores estuvieron estudiando en un centro estatal antes de entrar al centro juvenil.

Analizando por ultimo año o grado de estudios a los jóvenes infractores

Nivel de estudios	Frecuencia	Porcentaje
Sin nivel	17	0.9%
Educación inicial	5	0.3%
Primaria incompleta	256	13.0%
Primaria completa	358	18.2%
Secundaria incompleta	1120	57.0%
Secundaria completa	187	9.5%
Superior no universitaria incompleta	13	0.7%
Superior Universitaria incompleta	9	0.5%
Total	1965	100.0%



Del grafico podemos observar que el 57% de los jóvenes infractores no termino su secundaria.

5.4. Balanceo over-sampling:

Este método replica las observaciones de la clase minoritaria con el fin de que se tenga una proporción similar a la otra clase y así obtener un conjunto de datos balanceados. También es conocido como upsampling. Trabajando la base de datos, se elegirán aleatoriamente los valores

Del cuadro xxx anterior se pudo observar que nuestra variable respuesta estaba des balanceada, aplicando esta metodología se llego a balancear la base de datos, el cuadro xxx nos mostrara la nueva distribuciónsimétrica de nuestra variable reincidente:

Cuadro xx

Target	Frecuencia	Porcentaje
No Reincidente	1698	51.45%
Reincidente	1602	48.55%
Total	3300	100.00%

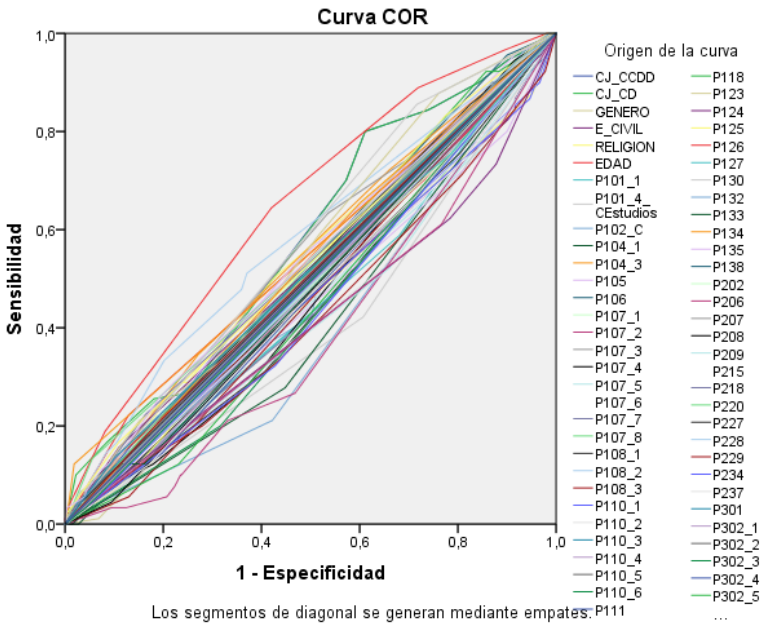
Elaboración: Propia

La limitación que tiene esta metodología se da cuando replicas la cantidad de valores positivos en nuestra variable, y con esto nuestro modelo puede estar sobreestimado, para poder verificar que nuestros parámetros no están sobreestimados se aplicara el crossvalidation para comprobar la estimación de parámetros.

5.5. Modelado:

Para la selección de variables utilizaremos la curva COR con las 123 variables y tomaremos las que presentan mayor al 5%.

Gráfico xxx



Elaboración: Propia

CUADRO xx

Variables renombradas	variable	Área	Límite inferior	Límite superior	gini
EDAD	1. DEPARTAMENTO	.645	.586	.705	29.1%
P206	5. NOMBRE DEL CENTRO JUVENIL	.393	.334	.451	21.5%
P132	7. SEXO	.397	.337	.456	20.7%
P130	8. ESTADO CIVIL	.406	.343	.470	18.8%
P133	9. RELIGIÓN	.414	.353	.475	17.1%
P332	11. EDAD DEL (LA) ADOLESCENTE INFRACTOR	.416	.348	.484	16.7%
P325	EL ÚLTIMO AÑO O GRADO DE ESTUDIOS Y NIVEL QUE USTED APROBÓ	.579	.513	.644	15.7%

Elaboración: Propia

Del cuadro xxx se puede observar una pequeña muestra de las variables que presentan mayor influencia con nuestra variable respuesta.

De esta selección se escogieron 50 variables.

Redes Neuronales:

Resumen del modelo

Capa oculta

Número de unidades: 5

Función de activación: softmax

Capa de salida

Función de activación: Identidad

Función de error: Suma de cuadrados

Entrenamiento	Error de suma de cuadrados	146.680
	Porcentaje de pronósticos incorrectos	27.3%
	Tiempo de entretamiento	0:01:03.10
Pruebas	Error de suma de cuadrados	68,159[a]
	Porcentaje de pronósticos incorrectos	32.8%

Tabla de clasificación de entrenamiento y validación

Ejemplo		Pronosticado		
		No Reincidente	Reincidente	Porcentaje correcto
Entrenamiento	No Reincidente	268	125	68.2%
	Reincidente	80	278	77.7%
	Porcentaje global	46.3%	53.7%	72.7%
Pruebas	No Reincidente	95	49	68.0%
	Reincidente	58	124	68.1%
	Porcentaje global	46.9%	53.1%	67.2%

Del cuadro xxx se obtiene los indicadores que se usaran para comprobar las técnicas estadísticas utilizadas.

Importancia de variables

variables	Importancia	Importancia normalizada
11. EDAD DEL (LA) ADOLESCENTE INFRACTOR	.182	100.0%
SEGÚN LO DICHO POR LAS AUTORIDADES, CUANDO OCURRIÓ LA INFRACCIÓN ¿USTED LLEVABA ALGÚN ARMA?	.064	35.0%
ANTES DE INGRESAR AL CENTRO JUVENIL, ¿ALGUN(OS) DE SU(S) MEJOR(ES) AMIGO(S) COMETIERON O COMETÍA(N) ALGUNAS INFRACCION(ES) CONTRA LA LEY?	.042	23.1%
EN EL BARRIO/LUGAR DONDE USTED VIVÍA ANTES DE INGRESAR AL CENTRO JUVENIL, ¿HABÍAN PANDILLAS O BANDAS DELICTIVAS?	.041	22.7%
Drogas	.038	21.1%
¿ALGÚN MIEMBRO DE SU FAMILIA ESTUVO INTERNADO EN UN ESTABLECIMIENTO PENITENCIARIO ALGUNA VEZ?	.038	20.8%
¿USTED PERTENECIÓ O PERTENECE A ALGUNA BANDA CRIMINAL?	.038	20.7%

Árbol de decisiones:

Resumen del modelo

Método de crecimiento: CHAID

Variable dependiente: Reincidencia

Máxima profundidad del árbol: 3

Casos mínimos en nodo padre: 100

Casos mínimos en nodo hijo: 50

Resultados del árbol

Numero de nodos: 19

Numero de nodos terminales: 11

Profundidad: 3

Tabla de calcificación

Ejemplo		Pronosticado		
		No Reincidente	Reincidente	Porcentaje correcto
Entrenamiento	No Reincidente	837	374	69.1%
	Reincidente	371	728	66.2%
	Porcentaje global	52.3%	47.7%	67.7%
Prueba	No Reincidente	329	158	68.0%
	Reincidente	155	348	68.1%
	Porcentaje global	46.6%	53.1%	67.2%

Del cuadro xxx se obtiene los indicadores que se usaran para comprobar las técnicas estadísticas utilizadas.

Variables más influyentes según árbol de decisiones

Tuvo amigos que cometieron crímenes

Edad

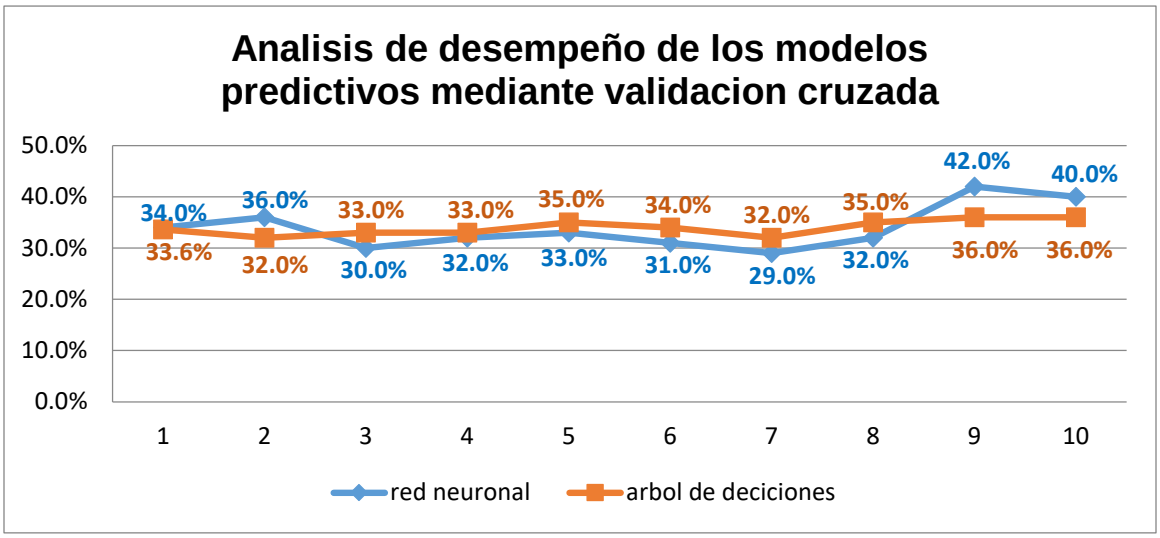
Violencia en el barrio donde vivía

Pasatiempos dentro del centro penitenciario

Castigos sobre suspensión de visitas

5.6. Cross validation:

Utilizando una partición de 10 muestras, vemos como el error del árbol de decisiones tiene una tendencia constante, en cambio la red neuronal no presenta una estacionalidad en sus errores.

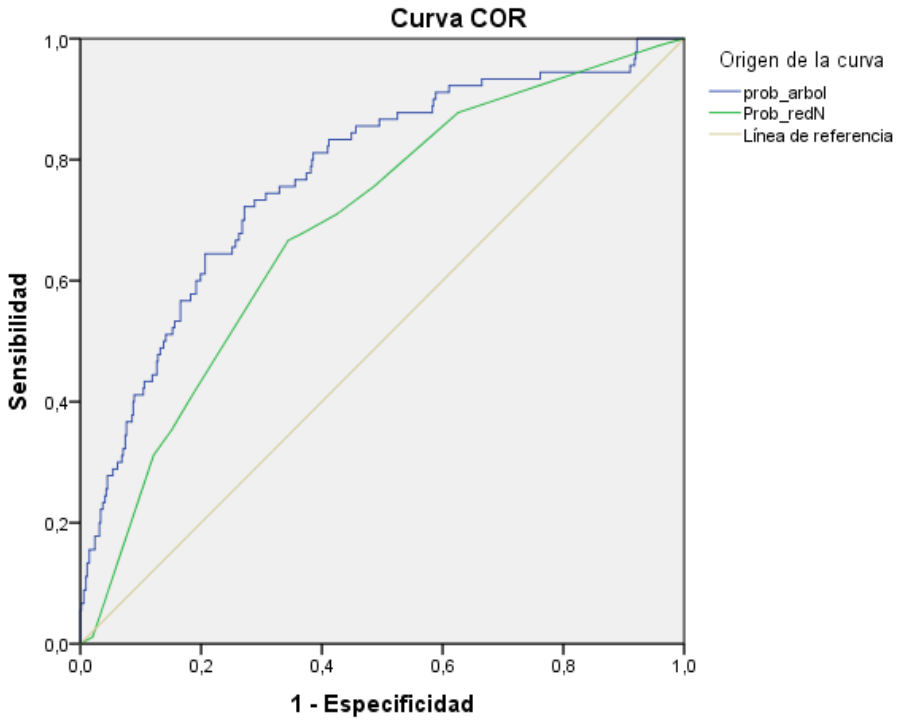


5.7. Comparación de técnicas

INDICADORES	RED NEURONAL	ARBOL DE DECISIÓN
sensibilidad	66.2%	68.1%
especificidad	68.1%	68.0%
exactitud	66.7%	67.2%
VP	52.3%	53.1%
VN	47.7%	46.9%

Por la inestabilidad de los parámetros de las redes neuronales, se escoge como mejor modelo al árbol de decisiones también por la proximidad de sus parámetros con la red y su estabilidad que presenta el error en el corssvalidation

Analizando por la curva ROC



Variables de resultado de prueba	Área	GINI
prob_arbol	.772	54.5%
Prob_redN	.691	38.2%

Del cuadro anterior se puede observar que el mayor GINI presentado es el árbol de decisiones

5.9. Perfil de los jóvenes infractores reincidentes

variables	categorías	propensión alta	propensión baja
genero	masculino	97%	97%
edad	12 a 16 años	58%	55%
tipo de colegio	estatal	68%	50%
grado de instrucción	primaria	54%	51%
consumo de drogas	si	64%	60%
consumo de bebidas alcohólicas	si	64%	52%
consumo de cigarrillos	si	67%	58%
lugar del delito	vía publica	59%	57%
programas en los centros juveniles	si	13%	12%

De cuadro anterior se puede observar que se tomo las probabilidades de propensión alta y baja para la interpretación de porcentajes de reincidencia.

CONCLUSIONES:

- ✓ Para la obtención de la técnica estadística se obtuvo que la más adecuada es el árbol de decisión por su estabilidad de parámetros en el cros validation y sus indicadores
- ✓ Los indicadores hallados en la red neuronal fueron:

INDICADORES	RED NEURONAL
sensibilidad	66.2%
especificidad	68.1%
exactitud	66.7%
VP	52.3%
VN	47.7%
GINI	38.2%

- ✓ Los indicadores hallados el árbol de decisiones fueron:

INDICADORES	ARBOL DE DECISIÓN
sensibilidad	68.1%
especificidad	68.0%
exactitud	67.2%
VP	53.1%
VN	46.9%
GINI	54.5%

RECOMENDACIONES

- ✓ Para el tratamiento de la asimetría se pudo utilizar una regresión logística también por la parsimonia, pero como en los antecedentes el árbol de decisión explica bien este tipo de estudios se tomo prioridad a esta técnica
- ✓ Para elevar la potencia del trabajo se pudo dicotomizar las variables y ver el impacto de cada categoría.
- ✓ El estudio se realiza para todos los departamentos, pero se debe realizar un estudio por centro juvenil para detectar mejor las soluciones adecuadas.

REFERENCIAS BIBLIOGRÁFICAS

[1] Manel Capdevila Capdevila, Marta Ferrer Puig, Eulàlia Luque Reina; *La reincidencia en el delito en la justicia de menores*, España 2005.

[2] Unidad de Estudios Servicio Nacional de Menores, *Reincidencia de jóvenes infractores de ley RPA,* Chile 2015.

[3] Naciones unidas: Oficina contra la droga y el delito; *La Relación droga y delito en adolescentes infractores de la ley la experiencia de Bolivia, Chile, Colombia, Perú Y* Uruguay*,* Argentina 2010.

[4] Condori Ingaroga Luis Julios, *Funcionamiento familiar y situaciones de crisis de adolecentes infractores y no infractores en lima metropolitana*, Perú 2002.

[5] Claudia Sandoval Ibarra; *Relatos de vida de jóvenes infractores de ley: una aproximación a sus procesos de reinserción social y comunitaria*, Chile 2007

[6] Esther F.J.C. van Ginneken, Alex Sutherland, ToonMolleman; *Ecological analysis of* prison *overcrowding and suicide rates in England and Wales, 2000–2014*, Inglaterra 2015.

[7] Patrick Ramos Chavarría; *Sobrepoblación y hacinamiento carcelario: los casos de los* Centros *de Atención Institucional La Reforma, El Buen Pastor y San Sebastián*, Costa Rica 2008.

ANEXOS

Cobertura Temática
El estudio se centra en las variables que permitirán conocer las condiciones por las cuales el interno(a) procedió a cometer infracciones. Para ello se ha tomado en cuenta los siguientes acápites y preguntas:

TOTAL DE PREGUNTAS	**179 Preguntas**
Ubicación Geográfica del Establecimiento	5 Preguntas
Identificación del Interno	11 Peguntas
CAPITULO 100: Condiciones Sociales y Familiares del (la) Adolescente Infractor	**55 Preguntas**
a) Educación del Interno	07
b) Salud del Interno	14
c) Discapacidad	02
d) Empleo	08
e) Etnicidad	01
f) Entorno Familiar	20
g) Discriminación	03
CAPÍTULO 200: Situación de la Infracción Penal	**42 Preguntas**
a) Descripción de la infracción penal	30
b) Situación jurídica del (la) adolescente infractor	12
CAPÍTULO 300: Condiciones de Vida del (la) Adolescente Infractor en el Centro Juvenil	**46 Preguntas**
a) Condiciones de vida en el Centro Juvenil	06
b) Educación en el Centro Juvenil	08
c) Salud en el Centro Juvenil	07
d) Entretenimiento	04
e) Servicios del Centro Juvenil	06
f) Visitas Familiares	07
g) Discriminación	04
h) Seguridad, Violencia y Consumo de Drogas	04
CAPÍTULO 400: Rol de las Instituciones	**18 Preguntas**

a) Durante su Detención	04
b) Rol de la Comisaría o Sede Policial	10
c) Rol de la Fiscalía	02
d) Juzgado de Familia	01
e) Medida de internamiento	01

CAPÍTULO 500: Expectativas del (la) Adolescente Infractor **02 Preguntas**