UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

FACULTAD DE CIENCIAS MATEMÁTICAS

EAP. DE ESTADÍSTICA

Comparación de modelos de clasificación: regresión logística y árboles de clasificación para evaluar el rendimiento académico

TESIS

Para optar el Título Profesional de Licenciada en Estadística

AUTOR

Mónica LIZARES CASTILLO

Lima - Perú

2017

"COMPARACIÓN DE MODELOS DE CLASIFICACIÓN: REGRESIÓN LOGÍSTICA Y ÁRBOLES DE CLASIFICACIÓN PARA EVALUAR EL RENDIMIENTO ACADÉMICO"

MÓNICA LIZARES CASTILLO

Tesina presentada a consideración del Cuerpo Docente de la Facultad de Ciencias Matemáticas de la Universidad Nacional Mayor de San Marcos, como parte de los requisitos para obtener el título profesional de Licenciada en Estadística .

Aprobado por:

.....

Mg.Grabiela Montes Quintana

.....

Mg. Geraldine Judith Vigo Chacón

Lima – Perú
2017

FICHA CATALOGRÁFICA

LIZARES CASTILLO MÓNICA

COMPARACIÓN DE MODELOS DE CLASIFICACIÓN: REGRESIÓN LOGÍSTICA Y ARBOLES DE CLASIFICACIÓN PARA EVALUAR EL RENDIMIENTO ACADÉMICO Lima 2017.

vii, 63p,29.7 cm (UNMSM, Licenciada, Estadística, 2017).
Universidad Nacional Mayor de San Marcos
Facultad de Ciencias Matemáticas
Estadística
UNMSM/FdeCM

A Dios quien supo guiarme por el buen camino, darme fuerzas para seguir adelante.

A mis padres Juan y María, porque ellos siempre estuvieron a mi lado brindándome su apoyo y sus consejos para hacer de mí una mejor persona.

A mis hermanos por estar siempre presentes, acompañándome para poderme realizar.

A mis amigos y compañeros que de una manera han contribuido para el logro de mis objetivos.

AGRADECIMIENTOS

Gracias Dios por bendecirme para llegar hasta donde he llegado, porque hiciste realidad este sueño anhelado.

A la Universidad Nacional Mayor de San Marcos por darme la oportunidad de estudiar y ser un profesional.

A mis padres quienes a lo largo de toda mi vida me han apoyado y motivado en mi formación académica y creyeron en mí en todo momento.

A mi asesora de tesina, por su paciencia y dedicación, quien con sus conocimientos y su experiencia ha logrado en mí pueda terminar con éxito, este trabajo de investigación.

A mis amigos; Marisol, Ruth, Yorgi y Marcos por confiar en mí y siempre alentándome para seguir adelante

De igual manera el agradecimiento a todos mis profesores que estuvieron presentes en mi formación profesional.

RESUMEN

"COMPARACIÓN DE MODELOS DE CLASIFICACIÓN: REGRESIÓN LOGÍSTICA Y ÁRBOLES DE CLASIFICACIÓN PARA EVALUAR EL RENDIMIENTO ACADÉMICO

Mónica Lizares Castillo
Setiembre 2017

En el trabajo de investigación se comparan dos modelos de clasificación: Regresión Logística Binaria y Arboles de clasificación (CHAID) para evaluar el rendimiento académico.

El comportamiento de estos modelos fue medido por cuatro indicadores: Sensibilidad, Curva ROC, Índice de GINI e Índice de Kappa en base al poder de clasificación y predicción de los modelos obtenidos sobre rendimiento académico.

Siendo Arboles de clasificación el mejor modelo por tener mayor poder de clasificación y predicción.

Para el análisis se utilizó una base de datos sobre estudiantes universitarios del primer semestre matriculado en el curso de Matemática, obtenido de un repositorio de Machine Learning.

PALABRAS CLAVES: Regresión Logística Binaria, Árboles de decisión, clasificación, Chaid.

ABSTRACT

"COMPARACIÓN DE MODELOS DE CLASIFICACION: REGRESIÓN LOGÍSTICA Y ÁRBOLES DE CLASIFICACIÓN PARA EVALUAR EL RENDIMIENTO ACADÉMICO

Mónica Lizares Castillo
Setiembre 2017

In the research work, two classification models are compared: Binary Logistic Regression and Classification Trees (CHAID) to evaluate academic performance.

The behavior of these models was measured by four indicators: Sensitivity, ROC Curve, GINI Index and Kappa Index based on the power of classification and prediction of the models obtained on academic performance.

Being Trees of classification the best model to have greater power of classification and prediction.

For the analysis, a database of first-year university students enrolled in the Mathematics course, obtained from a Machine Learning repository, was used.

KEY WORDS: Binary Logistic Regression, Decision Trees, CHAID.

INDICE

CAPÍTULO I

INTRODUCCIÓN

El rendimiento académico en la vida universitaria han de ser observadas cuidadosamente, pues las asignaturas de matemáticas en los primeros ciclos suelen ser las que menos gustan y las que menos entienden los estudiantes, lo que les ocasiona un rechazo y en consecuencia un posible abandono en un corto período de tiempo.

Teniendo en cuenta el gran avance en los sistemas de minería de datos las entidades educativas han buscado maneras de explotar al máximo la información existente en sus sistemas de información, esto basándose en técnicas estadísticas de clasificación y software especializados que permiten interpretación fácil y real de los resultados.

Es así como para dar apoyo en la toma de decisiones en el ambiente educativo se realiza comparación de modelos con las técnicas de Regresión Logística y Árboles de decisiones para evaluar el rendimiento académico, con la finalidad de predecir la clasificación (Desaprobado) usando una base de datos sobre rendimiento académico de alumnos universitarios en término de factores sociodemográficos y académicos.

Se usa la tabla de clasificación para evaluar la precisión de clasificadores, los resultados indican que la técnica de Árboles de decisión obtuvo el mayor porcentaje de buena clasificación, siendo el mejor modelo bajo la curva ROC, Sensibilidad, Índice de Gini e Índice de Kappa.

El presente trabajo de investigación comprende Cinco Capítulos: el primer Capítulo se describe el planteamiento del problema, formulación del problema, justificación del problema y objetivos de la investigación.

En el Capítulo II comprende el marco teórico de las técnicas de Regresión Logística y Árboles de clasificación.

En el capítulo III se presenta la Metodología donde se describe el diseño y tipo de investigación, la unidad de análisis, el tamaño de la muestra (entrenamiento y comprobación), la selección y descripción de las variables.

En el capítulo IV se presenta el análisis de resultados de las técnicas de clasificación: Regresión Logística y Árboles de clasificación y finalmente se presentan en el capítulo V las conclusiones y recomendaciones.

1.1. Situación Problemática

Hoy en día frente a diferentes situaciones de éxito o fracaso académico, el término más común usado para medir es el rendimiento académico, es por eso que las entidades educativas optan por instrumentos estadísticos que les permite clasificar de manera adecuada a un conjunto de estudiantes, en función de ciertas características que garanticen su desempeño académico, por consiguiente será necesario identificar cuáles son los factores que forman estos perfiles.

Por ello se considera las técnicas de clasificación de minería de datos: Regresión Logística y Arboles de clasificación que nos permitan etiquetar a estudiantes aprobados o desaprobados, en el cual se encontrara una técnica adecuada con la finalidad de obtener un modelo con un alto poder de clasificación y predicción para analizar el rendimiento académico.

1.2. Formulación del Problema

1.2.1 Problema General

¿Qué modelo de clasificación: Regresión Logística o Arboles de clasificación permiten evaluar el rendimiento académico adecuadamente?

1.2.2 Problemas Específicos

- ¿Regresión Logística permite evaluar el rendimiento académico adecuadamente?
- ¿Árboles de clasificación permite evaluar el rendimiento académico adecuadamente?

1.3. Justificación del problema

Cada día se generan grandes cantidades de datos que necesitan ser tratados mediante metodologías que generan información, que nos pueden ayudar a investigar, predecir o tomar decisiones, además el desarrollo tecnológico nos permite no solo almacenar la información sino procesarlo y general conocimiento .

Hay un gran interés por aplicar las técnicas de minerías de datos para evaluar el rendimiento académico, debido a la preocupación de las entidades educativas de identificar las características de dichos estudiantes universitarios con la finalidad de predecir correctamente la probabilidad de desaprobados en el curso de Matemática.

Bajo este enfoque se desarrolla el presente trabajo, específicamente en la toma de decisiones sobre rendimiento académico usando las técnicas de clasificación: Regresión Logística y Árboles de clasificación.

1.4. Objetivos

1.4.1 Objetivo General

Comparar los modelos de clasificación: Regresión Logística y Árboles de Clasificación para evaluar el rendimiento académico adecuadamente.

1.4.2 Objetivos Específicos

- Realizar el modelo de clasificación: Regresión Logística para evaluar el rendimiento académico adecuadamente.

- Realizar el modelo de clasificación: Arboles de clasificación para evaluar el rendimiento académico adecuadamente

CAPÍTULO II

2. MARCO TÉORICO

2.1. Antecedentes de investigación

Un estudio realizado por Eduardo Adolfo Porcel. (2010) En su artículo titulado "Predicción del rendimiento académico de alumnos de primer año de la FACENA (UNNE) en función de su caracterización socioeducativa. Argentina-Corrientes". En este trabajo se analiza la relación del rendimiento académico de los alumnos ingresantes a la Facultad de Ciencias Exactas y Naturales y Agrimensura de la Universidad Nacional del Nordeste (FACENA-UNNE) en Corrientes, Argentina, durante el primer año de carrera con las características socioeducativas de los mismos.

El rendimiento fue medido por la aprobación de los exámenes parciales o finales de la primera materia de Matemática que los alumnos cursan. Se ajustó un modelo de regresión logística binaria, en el cual el modelo ajustado clasifica correctamente el rendimiento académico del 74.8% de los ingresantes en el período 2004-2005.

Las variables que resultaron estadísticamente significativas fueron las siguientes: año de ingreso, carrera, tenencia de mail, título secundario, cobertura de obra social y nivel educacional de los padres, y las que no resultaron significativas fueron: sexo y dependencia del establecimiento secundario.

Se observa que, de las variables que resultaron estadísticamente significativas en la predicción del rendimiento académico, la más destacable es la que tiene que ver con el título secundario del alumno, que representa la orientación de la formación recibida por el mismo en el nivel medio preuniversitario.

Respecto de los métodos utilizados, el modelo de regresión logística binaria adoptado presenta un porcentaje elevado de predicción, quienes concluyen que "los modelos de regresión logística demuestran ser una herramienta muy poderosa para indagar sobre la explicación de variables categóricas de respuesta, en las que el investigador tenga especial interés".

El autor(a) Liliana Vitola Garrido, (2014) publicó un artículo de investigación "Una aplicación en la identificación de variables que inciden en el rendimiento académico, en el área de matemáticas", la investigación estuvo motivada por el bajo rendimiento que los estudiantes del nivel básico secundario registran en dicha área, tanto a nivel institucional como local.

Aplicando un modelo de regresión logística, se analizaron las variables estructurales mencionadas, en el que se indagó sobre aspectos socioeconómicos, demográficos, familiares, personales y físicos de la vivienda en la cual habitan los estudiantes y que pueden estar relacionados con su rendimiento académico.

Como se puede observar, las variables número de hermanos, sexo, estado civil, raza, ingresos económicos de los padres, personas con quien vive e interés por el estudio tiene un valor p menor que 0,05, lo que lleva a concluir que dichas variables influyen sobre la variable dependiente rendimiento académico en el área de matemáticas con una confianza del 95%.

Concluyendo que el rendimiento académico de los estudiantes de la IE Santa Rosa de Lima, en el área de matemáticas, está influenciado por variables tales como el número de hermanos que el estudiante tiene, el sexo del estudiante, el estado civil de los padres, la raza del estudiante, los ingresos económicos de los padres, con quién vive estudiante y si el alumno tiene interés por estudiar.

En la Universidad Nacional de Misiones (Argentina) se realizó una investigación sobre deserción estudiantil utilizando las técnicas de minería de datos. Su objetivo principal fue maximizar la calidad que los modelos tienen para clasificar y agrupar a los estudiantes, de acuerdo a sus características académicas, factores sociales y demográficos, que han desertado de la Carrera Analista en Sistemas de Computación de la Facultad de Ciencias Exactas, Químicas y Naturales analizando los datos de las cohortes entre los años 2000 al 2006 (Pautsch, 2009) (Pautsch et al., 2010).

Se interpretan los patrones descubiertos con el fin de consolidar el conocimiento descubierto e incorporarlo en otro sistema para posteriores acciones o para confrontarlo con conocimiento previamente descubierto.

Regla 1: El promedio de notas es menor que 2.4 el estudiante deserta. El 19% del total de estudiantes (2.136) que ingresaron a la Universidad de Nariño y la Institución Universitaria CESMAG entre los años 2004 y 2006 se clasifican de esta manera y el 34,8 % del total de estudiantes desertores (1.165), cumplen con este patrón. De igual manera, si el promedio de notas esta entre 2,4 y 3,1 entonces el estudiante deserta. El 18% de los 2.136 estudiantes que ingresaron en las cohortes estudiadas tienen este perfil y el 32,8% del total de desertores cumplen este patrón.

Regla 2: El 100% de los estudiantes que desertan realizaron su bachillerato en un colegio público, son solteros, su promedio de notas es menor que 2.4, han perdido materias en los primeros semestres (1 a 4) y todas las materias las han perdido una sola vez. El 11.3% del total de estudiantes (2.136) que ingresaron a la Universidad de Nariño y la Institución Universitaria CESMAG entre los años 2004 y 2006 cumplen con este patrón.

Regla 3: El 100% de los estudiantes que desertan son hombres solteros, su promedio de notas es menor que 2.4 y son de una universidad privada, para este caso IUCESMAG. El 8.5% del total de estudiantes (2.136) que ingresaron a la Universidad de Nariño y la Institución Universitaria CESMAG entre los años 2004 y 2006 cumplen con este patrón.

Como parte de la conclusión :Los perfiles de deserción estudiantil obtenidos a través de las técnicas de minería de datos: clasificación, asociación y agrupamiento indican que éstas son capaces de generar modelos consistentes con la realidad observada y el respaldo teórico, basándose únicamente en los datos que se encontraron almacenados en las bases de datos de la Universidad de Nariño y de la Institución Universitaria CESMAG, complementados con fuentes externas de datos pertenecientes principalmente a SISBEN, Sistema de Prevención y Análisis de la Deserción en las Instituciones de Educación Superior (SPADIES),Alcaldía Municipal de Pasto (Estratificación), Departamento Administrativo Nacional de Estadística (DANE), Instituto Colombiano.

2.2. Regresión Logística

En muchas investigaciones se está interesado en relacionar una variable dependiente dicotómica y variables independientes de tipo categórico y cuantitativo. Recordemos que una variable dicotómica, es una variable que puede tomar sólo uno de dos valores mutuamente excluyentes, por lo general se codifican como Y = 1 para éxito e Y= 0 para fracaso.

El análisis de regresión es uno de los métodos estadísticos más útiles y utilizados. El objetivo de los modelos de regresión es describir la relación entre una variable respuesta y una o más variables explicativas. Entre los diferentes modelos de regresión, la regresión logística juega un papel particular. El concepto básico, sin embargo, es universal.

El análisis de regresión logística es una técnica estadística multivariante que nos permite estudiar la relación entre una o más variables independientes y una variable dependiente de tipo dicotómica, representa la ocurrencia o no de un suceso, como por ejemplo, muerte o vida, sano o enfermo, fumador o no fumador, madre adolescente o madre no adolescente, hipertenso o no hipertenso, etc.

El análisis de Regresión logística tiene la misma estrategia que el Análisis de Regresión Lineal Múltiple, el cual se diferencia esencialmente del Análisis de Regresión Logística porque la variable dependiente es métrica; en la práctica el uso de ambas técnicas tienen mucha semejanza, aunque sus enfoques matemáticos son diferentes .

La variable dependiente o respuesta no es continua sino discreta, sino discreta (generalmente toma valores 1,0). Las variables explicativas pueden ser cuantitativas o cualitativas; y la ecuación del modelo no es una función lineal de partida, sino exponencial, si bien, por sencilla transformación logarítmica, puede finalmente presentarse como una función lineal. Así pues el modelo será útil en frecuentes situaciones prácticas de investigación en que la respuesta puede tomar únicamente dos valores: 1, presencia (con probabilidad p); y 0, ausencia (con probabilidad 1-p).

El modelo será de utilidad puesto que, muchas veces, el perfil de variables puede estar formado por caracteres cuantitativos cualitativos; y se pretende hacer participar a todos ellos en una única ecuación conjunta.

El modelo puede acercarse más a la realidad ya que muchos fenómenos, como los del campo epidemiológico, se asemejan más a una curva que a una recta. Además la curva exponencial elegida como mejor ajuste, puede ser transformada logarítmicamente en una ecuación lineal de todas las variables, siendo así que el aparato matemático estudiado para la regresión lineal múltiple será aplicable; aunque el investigador tenga, al final, que deshacer la transformación para interpretar sus conclusiones.

2.2.1 El Modelo de Regresión Logística

En el modelo de regresión logística, la variable respuesta (Y) es dicotómica, posee valores 1 o 0, con probabilidad π_i para $_{Y_i} = 1$ y probabilidad $1 - \pi_i$ para $Y_i = 0$, según Hosmer y Lemeshow (2000).

El análisis de Regresión Logística comprende la estimación de la probabilidad de que ocurra un evento (variable de respuesta dicotómica; con valores 0 y 1) como función de los valores de p variables independientes (predictoras).

Consideremos Y una variable respuesta y una colección de p variables independientes expresado por el vector $X' = (x_1, x_2, \dots \dots, x_p)$.
La forma específica del modelo logístico con p variables predictoras está representado por:

$$\pi = \pi(x) = P(Y = 1|x) \;=\; \frac{e^{x'\beta}}{1 + \; e^{x'\beta}} \qquad (2.2.1)$$

Que representa que la probabilidad condicional de que el evento Y=1 ocurra dada la ocurrencia de un conjunto de variables X. (Probabilidad de éxito).

Una transformación de $\pi(x)$ que es fundamental en el estudio de la regresión logística es la transformación logit. Esta transformación se define en términos de $\pi(x)$, como:

$$g(x) = \ln\left[\frac{\pi(x)}{1 - \pi(x)}\right] = \beta_0 + \; \beta_1 x_1 + \; \dots + \; \beta_p x_p \qquad (2.2.2)$$

Donde $\boldsymbol{\beta_0}$ es la constantes y los $\boldsymbol{\beta_i}$ son los coeficientes de los predictores $\boldsymbol{x_i}$ del modelo .La importancia de esta transformación es que g(x) posee muchas de las propiedades deseables de un modelo de regresión lineal. La función logit es lineal en sus parámetros, puede ser continuo y variar de -∞ a +∞, dependiendo del rango de x.

El modelo logístico puede expresarse en términos de (ODDS) de ocurrencia de eventos. Esta razón se define como el cociente de la probabilidad de que el evento ocurra a la probabilidad de que el evento no ocurra.

Entonces sí:

$\pi(x) =$ Probabilidad de que el evento ocurra

$1 - \pi(x) =$Probabilidad de que el evento no ocurra

$$\text{ODDS} = \frac{\pi(x)}{1-\pi(x)} \qquad (2.2.3)$$

2.2.2 Estimación de los Coeficientes del Modelo

Debido a que la distribución de Y dado un conjunto de variables X= $(x_{1,}, x_{1,}, \dots, x_p)$ no es normal y no existe homocedasticidad en los errores, las estimación del vector $\hat{\beta} = (\widehat{\beta_1}, \widehat{\beta_2}, .., \widehat{\beta_p})$ por el métodos de mínimos cuadrados no tiene propiedades optimas, y en su lugar emplearemos el método de máxima verosimilitud para obtener los valores de los parámetros desconocidos que maximizan la probabilidad de obtener el conjunto observado de datos.

Para estimar los parámetros $\hat{\beta} = (\widehat{\beta_1}, \widehat{\beta_2}, .., \widehat{\beta_p})$ del modelo se utiliza el método de máxima verosimilitud, con lo cual encontramos el valor de β que maximiza $l(\beta)$

Así, la función de verosimilitud puede ser escrita:

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i}$$

Aplicando logaritmo neperiano, la expresión L(β) se define como:

$$\text{L}(\beta) = \ln[\, \text{l}(\beta)] = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\}$$

Para encontrar el valor de β se deriva L(β) con respecto a $\beta_0, \beta_1, \dots, \beta_p$ y se iguala al valor cero, obteniéndose:

$$\sum_{i=1}^{n}[y_i - \pi(x_i)] = 0$$

$$\sum_{i=1}^{n} x_{ij}[y_i - \pi(x_i)] = 0$$

Para j=1,.., p

Para encontrar la solución de este conjunto de ecuaciones se utiliza el método iterativo de Newton-Raphson. Hoy en día existen paquetes estadísticos para estimar estos parámetros $\hat{\beta}$. Denota la solución de estas ecuaciones.

El método de estimación de la varianza y covarianza de los coeficientes estimados siguen desde la teoría bien desarrollada de la estimación de máxima verosimilitud.

Esta teoría establece que los estimadores son obtenidos de la matriz de segunda derivada parcial de la función de logaritmo de verosimilitud. Esta derivación parcial tiene la siguiente forma general:

$$\frac{\partial^2 l(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i(1-\pi_i) \quad (2.2.4)$$

$$\frac{\partial^2 l(\beta)}{\partial \beta_j^2 \partial \beta_l} = -\sum_{i=1}^{n} x_{ij}\, x_{jl} \pi_i(1-\pi_i) \quad (2.2.5)$$

Para j, l=0, 1,2,..., p donde π_i representa $\pi(x_i)$. Dado la matriz (p+1)*(p+1) contiene los negativos de los términos dados en la ecuación (2.2.4) y (2.2.5) se denota como I(β). Esta matriz se llama matriz de información observada. La varianza y covarianza de los coeficientes estimados se obtienen de la inversa de la matriz el cual se denota como Var (β)=I-1(β).

Estas ecuaciones son llamadas ecuaciones verosímiles.Las ecuaciones verosímiles no son lineales en los parámetros β y esto requiere métodos especiales para su solución, estos son de naturaleza iterativos y han sido programados por muchos paquetes estadística.

2.2.3 Prueba para la significancia del modelo

Después de haber ajustado el modelo de regresión logística Múltiple, comenzaremos el proceso de evaluación. Es decir se examina si las variables están significativamente relacionadas con la variable dependiente Y.

Estadística G

La prueba de razón de verosimilitud para estudiar la significación total de los parámetros de las variables independientes en el modelo se basa en la estadística G (Hosmer e Lemeshow ,2000).

Para la evaluación de la significancia del modelo, es decir determinar si las variables independientes son significativas o no, se plantea las siguientes hipótesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \ldots = \beta_P = 0$$

$$H_1 : \beta_i \neq 0 \quad para\ algún\ \beta_i$$

El Estadístico que se plantea para la evaluación de la significancia del modelo es la diferencia del valor de la desvianza del modelo con y sin la variable, basando el contraste en la diferencia entre razones de verosimilitud, medida por el estadístico.

G=D (modelo sin las variables) – D (modelo con las variables)

$$G = -2\ln\left[\frac{verosimilitud\ del\ modelo\ sin\ la\ variable}{verosimilitud\ del\ modelo\ con\ la\ variable}\right]$$

Este estadístico sigue una distribución Chi – Cuadrada con p grados de libertad. El estadístico G es útil cuando se desea evaluar si la adición de una variable(o conjunto de variables) tiene un efecto significativo sobre la variable de respuesta.

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

Donde:

$$n_1 = \sum y_i \quad , \quad n_0 = \sum(1-y_i)$$

La estadística G tiene una distribución Chi-Cuadrado con p grados de libertad χ_p^2 , bajo la hipótesis nula, Rechazamos Ho a un nivel de significancia α, si

$$G > \chi^2_{1-\alpha}(p)$$

Y concluímos que por lo menos uno de los parámetros es diferente de cero**.**

Estadística de Wald

Para evaluar la significancia de cada uno de los parámetros del modelo formulamos las siguientes hipótesis, Hosmer y Lemeshow (2000)

$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0 \quad para\ algún\ \beta_i$$

La estadística de Wald es:

$$W = \hat{\beta}'\left[\hat{Var}(\hat{\beta})\right]^{-1}\hat{\beta} = \hat{\beta}'\left(X'VX\right)\hat{\beta}$$

Donde:

V es una matriz diagonal de dimensión nxn, esto es:

$$V = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}$$

y X es una matriz de dimensión n x (p+1)

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

La estadística tiene distribución de una variable aleatoria chi-cuadrado con p+1 grados de libertad, χ^2_{p+1}, bajo la hipótesis nula, luego H_0 es rechazada a un nivel de significancia α , si :

$$W > \chi^2_{1-\alpha}(p+1)$$

Donde concluimos que por lo menos uno de los parámetros es diferente cero.

2.2.4 Pseudo Estadísticas R^2

El R cuadrado de Cox y Snell: es un coeficiente de determinación generalizado que se utiliza para estimar la proporción de la varianza de la variable dependiente explicada por las variables independientes. Se basa en la comparación del logaritmo de la verosimilitud para

el modelo respecto al logaritmo de la verosimilitud para un modelo de línea base. Los valores oscilan entre 0 y 1.

El R cuadrado de Nagelkerke: es una versión corregida de la R cuadrado de Cox y Snell, la R cuadrado de Cox y Snell tiene un valor máximo inferior a 1, incluso para un modelo "perfecto". La R cuadrado de Nagelkerke corrige la escala del estadístico para cubrir el rango completo de 0 a 1.

2.2.5. Evaluación de la bondad del ajuste del modelo

Test de Hosmer y Lemeshow

Para evaluar la bondad de ajuste del modelo, Hosmer–Lemeshow utiliza una estrategia de agrupamiento para obtener la estadística de bondad de ajuste, obtenida por el cálculo de la estadística Chi-Cuadrado de Pearson de una tabla de frecuencias observadas y frecuencias esperadas estimadas, Hosmer e Lemeshow (2000).

Hosmer–Lemeshow prueba las siguientes hipótesis:

H_0: No existen diferencias entre los valores observados y predichos

H_1: Existen diferencias entre los valores observados y predichos

Si rechazamos Ho, implica que el modelo ajustado no es el adecuado.

Se dividen todos los casos en deciles basados en las probabilidades predichas, el primer decil se cuentan los casos con las probabilidades más altas, siendo el estadístico:

La estadística de prueba es:

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n_k \overline{\pi}_k)^2}{\overline{\pi}_k n_k (1 - \overline{\pi}_k)}$$

$O_k = \sum_{j=1}^{c_k} y_j$: Número de respuestas entre las covariables

n_k : Número de covariables en el k-esimo decil

c_k **:** Total de individuos en el k-esimo grupo

$\bar{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j \hat{\pi}_j}{n_k}$ **:** Probabilidad media estimada.

La estadística $\hat{C}$ tiene aproximadamente una distribución Chi-cuadrado con g – 2 grados de libertad, bajo la hipótesis nula. A un nivel de significancia α, rechazamos Ho si

$$\hat{C} > \chi^2_{1-\alpha}(g-2)$$

Y concluimos que el modelo no es el adecuado.

2.3. Árboles de decisión

Un árbol de decisión es una forma gráfica y analítica de representar todos los eventos (sucesos) que pueden surgir a partir de una decisión asumida en cierto momento. Nos ayudan a tomar la decisión más "acertada", desde un punto de vista probabilístico, ante un abanico de posibles decisiones. Estos árboles permiten examinar los resultados y determinar visualmente cómo fluye el modelo.

El procedimiento de Árboles de decisión crea un modelo de clasificación basado en árboles y clasifica casos en grupos o pronostica valores de una variable dependiente basada en valores de una variable independiente. El procedimiento proporciona herramientas de validación para análisis de clasificación exploratorios y confirmatorios.

Cabe resaltar que las variables en estudios pueden se definidas como valores discretos o continuos, donde se tiene una mayor representación de valores discretos debido al fácil acceso.

Dentro de los métodos basados en árboles se pueden distinguir dos tipos:

- Los árboles de clasificación, se emplea para variables categóricas, tanto nominales como ordinales.

- .Los árboles de regresión, este tipo de discriminación se aplica a variables continuas.

La característica más importante es que se asume que los grupos son disjuntos. Dado que la clasificación trata con grupos disjuntos ,un árbol de decisión conducirá un objeto hasta una y solo una hoja ,asignando por lo tanto, un único grupo a aun objeto .Para ello es necesario que las particiones existentes deben ser disjuntas.

2.3.1 Algoritmos

Existe una gran variedad de algoritmos de árboles de decisión que ayudan a construir un árbol:

CART (Arboles de clasificación y Regresión), es un algoritmo binario completo que hace particiones de datos y produce subconjuntos homogéneos precisos, utiliza el criterio del "índice de Gini" para seleccionar atributos.

CHAID, es un algoritmo estadístico rápido y multidireccional que explora rápida y eficientemente datos, y construye segmentos y perfiles en función de la variable de respuesta establecida.

CHAID Exhaustivo, es una modificación del CHAID que examina todas las posibles particiones de la variable predictora.

QUEST, es un algoritmo estadístico que selecciona variables de manera no sesgada y construye árboles binarios precisos rápidos y eficientes.

Árboles Bayesianos, es un algoritmo basado en la aplicación de métodos Bayesianos a árboles de decisión. Buntine (1992).

2.3.2 Técnica Chaid

Detección automática de interacciones mediante chi-cuadrado (Chi-saquea Automatic Interaction Detection).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.

Esta técnica CHAID divide el conjunto de datos en subconjunto que son mutuamente excluyentes y exhaustivos, que describen la mejor manera el comportamiento de la variable dependiente, es una técnica cuyo propósito es el de obtener tipologías y perfiles, el CHAID sirve para realizar segmentación de mercado.

Los árboles de decisión son una técnica estadística para la segmentación, la estratificación, la predicción, la reducción de datos y el filtrado de variables, la identificación de interacciones, la fusión de categorías y la discretización de variables continuas.

Es una especie de regresión múltiple para variables nominales, ordinales, categóricas, discretas, discontinuas, como por ejemplo, sexo, nivel socioeconómico, religión, ocupación,

ciudad, distrito, provincia; en la que existe una variable dependiente y al menos una variable independiente, que trata de predecir la variable de respuesta a través de las variables predictoras.

El CHAID ahorra tiempo al investigador, evitando realizar múltiple "tabulaciones cruzadas, divide a la población en dos o más grupos distintos basados en categorías del mejor predictor de una variable dependiente. Luego divide cada uno de estos en grupos más pequeños basados en variables de otros predictores.

Este proceso de división continua termina hasta que no se encuentren más predictores estadísticamente significativos(o hasta que se cumpla una regla de paro).

2.3.3. Condiciones y procedimientos para aplicar la segmentación Chaid

El algoritmo CHAID se caracteriza por realizar particiones n-binarias. Los datos con que trabajan corresponden a un conjunto de individuos u objetos N,$O_1, O_{2,}, \ldots, O_s$; denominando prototipos.

Individuo	x_1	x_2	…	x_i	……..	x_d	**Clase**
O_1	x_{11}	x_{12}	…	x_{1j}		x_{1d}	1
O_2	x_{21}	x_{22}	…	x_{2j}		x_{2d}	1
			…				
O_I	x_{I1}	x_{i2}	…	x_{ij}		x_{id}	1
			…				
O_S	x_{s1}	x_{s2}	…	x_{sj}		x_{sd}	j

La manera de trabajar mediante este método es el que se muestra a continuación:

- **Preparación de los predictores:** Dividiendo las respectivas distribuciones continuas en un número de categorías con un número aproximadamente igual de observaciones

.Para variables predictoras categóricas, las categorías (clases) son naturalmente definidas.

- **La fusión de las categorías:** Consiste en recorrer los predictores para determinar para cada predictor el par de categorías diferentes con respecto a la variable dependiente, para problemas de clasificación, se calculara la prueba chi-cuadrado, para problemas de regresión se toma las pruebas F.

- Si la prueba correspondiente para un determinado par de categorías de predicción no es estadísticamente significativa según la definición de un α valor, entonces se fusionaran las categorías respectivas de predicción y se repite este paso(es decir encontrar el siguiente par de categorías , que ahora pueden incluir fusionada categorías anteriores).Si la significación estadística para el par de las categorías respectivas del predictor es significativo entonces, calculara un ajuste de Bonferroni p-valor para el conjunto de categorías para el predictor respectivo.

- **Selección de la variable dividida:** Consiste en elegir la división de la variable explicativa con los ajustados p-valor , es decir la variable que producirá la división más importante , si el más pequeño (Bonferroni) ajustadas p-valor para cualquier predictor es mayor que el 5% que cierto a dividir el valor α, a continuación ,no más divisiones se llevaran a cabo, y el nodo correspondiente es un nodo terminal .Cabe recalcar las perspectivas más resaltantes de este método como se muestra a continuación :

Este método de segmentación trabaja con variables nominales, ordinales, intervalo y de frecuencias, el algoritmo CHAID identifica variables predictores y agrega grupos o clases en cada iteración.

Lo procedimientos que elabora están condicionados por lo siguiente.

a) Examina la relación entre una variable categórica con un número de variables predictoras, mediante árboles de decisión

b) Es un procedimiento basado en el estadístico chi-cuadrado para resolver problemas de predicción y clasificación, determinando variables y que grupos pueden obtenerse tal que se produzca la mayor discriminación posible entre grupos.

c) Los datos son agrupados, utilizando a un gran número de variables predictoras, de tal manera que se pueda mejorar la predicción o clasificación de acuerdo a una variable dependiente.

2.3.4 Pruebas Estadísticas

- **Prueba Chi-cuadrado de Independencia**

Supongamos que se realizan h experimentos independientes, cada uno, compuesto de k sucesos $A_1, A_2, \ldots, A_k$ con:

$$P_{IJ} = \mathrm{P}\,(A_k);\ i = 1,2,\ldots,k \quad ; \quad \mathrm{j}=1, 2, \mathrm{h}$$

Supongamos que el experimento l se repite n_i veces .Sean además $A_1, A_2, \ldots, A_k$ las variables aleatorias que describen el número de veces que se observa los sucesos A_i en los n_i repeticiones.

- **Evaluación del modelo**

$$H_O: Variables\ no\ asociadas\ al\ modelo$$
$$H_1: Variables\ asociadas\ al\ modelo$$

Con un nivel de significación α.

- **Estadístico de Prueba**

Se plantea para la evaluación del modelo, el estadístico que sigue una distribución Chi-cuadrado con k grado de libertad

$$X^2 = \sum_{I=1}^{m} \frac{(o_{ij}-e_{ij})^2}{e_{ij}} \sim X^2_{(r-1)x(k-1)}$$

Donde:

o_{ij}=Frecuencia Observada de la i-ésima categoría de la variable X y la j-ésima categoría de la variable Y.

$e_{ij=}$Frecuencia esperada de la i-ésima categoría de la variable X y la j-ésima categoría de la variable Y

$Frecuencia\ esperada\ e_{ij}$=$(\frac{Marginal\ f_i * Marginal\ f_j}{N})$

	Y_1	y_2	…..	Y_k	$f_{i.}$
x_1	O_{11} e_{11}	O_{12} e_{12}	…..	O_{1k} e_{1k}	$f_{1.}$
x_2	O_{21} e_{21}	O_{22} e_{22}		O_{2k} e_{2k}	$f_{2.}$
x_3	O_{31} e_{31}	O_{32} e_{32}	…..	O_{3k} e_{3k}	$f_{3.}$
….	….	…..	…..	…..	…..

x_r	O_{r1} e_{r1}	O_{r2} e_{r2}	…..	O_{rk} e_{rk}	$f_{r.}$
f_j	$f_{.1}$	$f_{.2}$	…..	$f_{.k}$	N

Marginal f_i. : Total de observaciones de la categoría X_i de la variable X

Marginal $f_{.j}$: Total de observaciones de la categoría Y_J de la variable Y

Cuanto mayor sea el valor de $X^2_{(g)}$, menos verosímil es que la hipótesis sea correcta. Los grados de libertad g.l vienen dados por:

$$G=(r-1)\,(k-1)$$

Donde:

r = Numero de categorías de la variable X.
k=Numero de categorías de la variable Y.

- **Regla de decisión**

i) Si $X^2 > X^2_{(g,1-\alpha)}$, rechaza Ho
ii) Si $X^2 < X^2_{(g.,1-\alpha)}$, acepta Ho

En caso de rechazar Ho, el modelo considera que las variables están asociadas, caso contrario, desestima la variable considerada.

2.3.5 Componente de un Análisis Chaid

Un análisis CHAID tiene los siguientes componentes básicos:

- Una o más variables predictivas cuyos valores se utilizan para definir los segmentos
- Podemos utilizar cualquier tipo de variable categórica incluyendo las demográficas, de estilo de vida, psicográficas y conductuales.

- El criterio (variable dependiente) para construir el modelo de segmentación.
- Este criterio está controlado por la elección de una o solo una variable (que debe ser categórica u ordinal).

2.3.6 El Algoritmo Chaid

El Algoritmo utilizado en Chaid tiene tres etapas: fusión, división y paro.

Etapa 1: Fusión

Para cada predicción x_1, x_2,... , x_k une categorías por medio de estos pasos:

1. Forma una tabulación cruzada de dos vías con una variable dependiente.
2. Por cada par de categorías que se pueden fusionar, mide estadísticas chi-cuadradas para probar la independencia entre el par de categorías y la variable dependiente se utilizan todas las variables de las categorías dependientes.
3. Calcula el valor p por cada par perfecto de ji-cuadrada.
4. Para cualquier variable conjunta que contenga tres o más categorías, prueba si la que es predictora se debe separar utilizando el nivel de importancia de las estadísticas Ji-cuadrada.
5. Une cualquier categoría que tenga pocas observaciones.

Etapa 2: División

Para las predicciones con valores p ajustados importantes estadísticamente hablando, divide el grupo en la predicción que tenga el valor p más bajo .Cada una de las categorías fusionadas

de la predicción se convierte en un nuevo subgrupo del grupo principal .Si ninguna predicción tiene un valor p importante, no divide el grupo.

Etapa 3: Paro

Regresa al paso 1 para analizar el siguiente subgrupo que contenga por lo menos tantas observaciones como especificaciones del tamaño mínimo del subgrupo (antes de dividirlo).Se detiene cuando haya analizado todos los subgrupos o cuando contengan demasiados casos.

2.3.7 Ventajas y desventajas de Árbol de Decisión

Los árboles de decisión crean un modelo de clasificación basado en diagramas de flujo. Clasifican casos en grupos o pronostican valores de una variable dependiente (criterio) basada en valores de variables independientes (productoras).

Las ventajas de un árbol de decisión son (Pérez, 2011):

- Facilita la interpretación de la decisión adoptada.
- Facilita la comprensión del conocimiento utilizado en la toma de decisiones.
- Explica el comportamiento respecto a una determinada decisión.
- Reduce el número de variables independientes.
- Permite al usuario reconocer segmentos del mercado.

Dentro de las desventajas que se plantean, las más representativas entre estos métodos de clasificación son:

- La reglas de asignación son bastantes sensibles a pequeñas perturbaciones en los datos (inestabilidad).
- Dificultad para elegir el árbol óptimo.

2.4. Técnicas de Evaluación de Clasificadores

La evaluación de las técnicas de clasificación, es importante porque permite validar la bondad de ajuste del modelo sobre el conjunto de entrenamiento. Así mismo, permiten comparar entre varias técnicas de clasificación y seleccionar la que tenga la mayor precisión. Para la evaluación de las técnicas de minería de datos: Regresión Logística y Árboles de decisión, se propone usar las tablas de clasificación, para determinar la Sensibilidad, área bajo la curva ROC, y el coeficiente Kappa.

2.4.1. Curvas ROC

Una forma de medir la bondad de ajuste es a través de la representación gráfica de la Curva ROC que compara la tasa de negativos verdaderos (Specificity) frente a 1 - tasa de positivos verdaderos (1 - Sensitivity) para varios puntos de corte.
El análisis de performance del modelo se realiza comparando el área por debajo de la curva ROC (denominado como AUC) con un área de 0,5 que resulta si el modelo clasifica aleatoriamente los casos.

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.

Las Hipótesis nula y alternamente son:

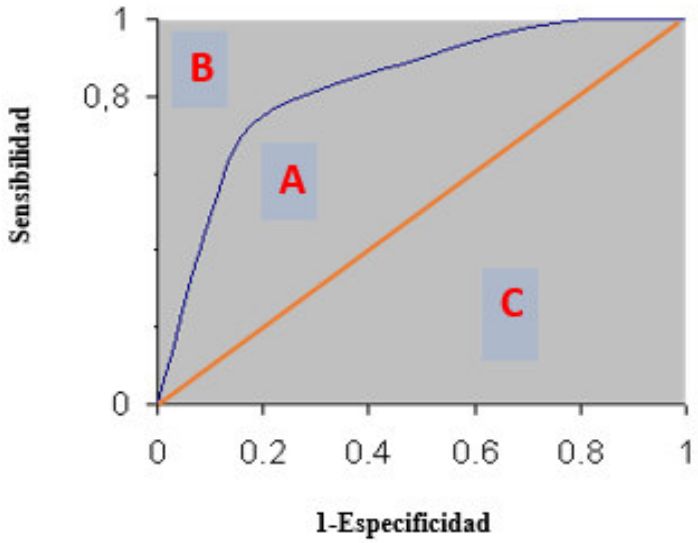
$$\boldsymbol{H_0}\textbf{ = El área bajo la curva ROC es igual a 0.5}$$

$$\boldsymbol{H_1}\textbf{ = El área bajo la curva ROC no es igual a 0.5}$$

Si rechazamos la $\boldsymbol{H_0}$ asociado a un p-valor, implica que el modelo ajustado es el adecuado

La curva ROC permite cuantificar la precisión discriminatoria de un modelo, mediante el área bajo la curva (AUC).

Figura 2.4.1.1.Gráfico de la Curva ROC



Arriba se muestra el ejemplo de la parcela ROC para el modelo de clasificación. Recuerde que el ROC se utiliza para evaluar el rendimiento de un modelo clasificador. Lo primero que hay que hacer antes de crear la trama ROC es, por supuesto, crear el modelo, mientras que el AUC es el área bajo la curva ROC. A modo de guía para interpretar las curvas ROC se han establecido los siguientes intervalos para los valores de AUC:

Tabla 2.4.1.1. Escala AUC

AUC	Scores
[0.5 , 0.6>	Test malo
[0.6 , 0.75>	Test regular
[0.75 , 0.9>	Test bueno
[0.9 , 0.97>	Test muy bueno
[0.97 , 1>	Test excelente
Fuente: Elaboración Propia	

2.4.2 Índice GINI

El coeficiente de GINI es el área que hay entre la curva ROC y la línea diagonal, representado por A en el gráfico anterior, expresado como un porcentaje del área triangular formado por la línea diagonal del modelo aleatorio.

El área ROC es el área, expresado como un porcentaje de toda el área cuadrangular.
Según las áreas representadas en el gráfico anterior tenemos:

$$\boldsymbol{COEFICIENTE\ GINI} = {}^{\boldsymbol{A}}\!/_{(\boldsymbol{A}+\boldsymbol{B})}$$

$$\boldsymbol{AREA\ ROC} = {}^{(\boldsymbol{A}+\boldsymbol{C})}\!/_{(\boldsymbol{A}+\boldsymbol{B}+\boldsymbol{C})}$$

El cálculo de Gini está estrechamente relacionado con el cálculo del AUC y puede ser calculado por:

Gini = 2 * AUC-1

Gini=2(ROC-50%)

AUC será entre 0 y 1. Cuanto mayor sea el valor de AUC, por lo general mejor es el modelo.

2.4.3 Índice de Kappa

Es un coeficiente estadístico propuesto originalmente por (Cohen, 1960) que permite medir la concordancia entre los resultados de dos o más variables cualitativas. El índice k, aplicado a la tabla de confusión permite evaluar si la clasificación observada es similar (concordante) con la clasificación predecida por el clasificador.

Para dos categorías, el coeficiente de Kappa se calcula:

$$k=[(P0-Pe)/(1-Pe)],$$

$$0 \leq k \leq 1 \text{ con } P0 = [(VP+VN)/N] \text{ y } Pe = [(a*c+b*d)/N2]$$

$$\text{Siendo: } a=VP+FP,\ b=FN+VN,\ c=VP+FN,\ d=FP+VN$$

Donde: Po, es la proporción de aciertos. Pe, es la proporción de aciertos esperados bajo la hipótesis de independencia entre las dos variables. En la Tabla se presenta la valoración del valor de k que utiliza la escala propuesta por (Landis and Koch, 1977).

Tabla 2.4.3.1.Kappa

Clasificación observada	Clasificación Predecida		Total (Observado)
	Positiva	Negativa	
Positiva	VP	FN	VP+FN
Negativa	FP	VN	FP+VN
Total (Predecido)	VP+FP	FN+VN	N

Fuente: Elaboración Propia

Donde: N=VP+VN+FP+FN

El VP (verdaderos positivos) y El VN (verdaderos negativos), es el número de observaciones que predice correctamente el clasificador como la clase positiva y negativa. El FP (falsos positivos) y El FN (falsos negativos), es el número de observaciones que se predice incorrectamente como la clase positiva siendo de la clase negativa y como la clase negativa siendo de la clase positiva respectivamente.

Tabla 2.4.3.1.Escala Kappa

Valor de k	Concordancia
< 0.20	Pobre
0.21 - 0.40	Débil
0.41 - 0.60	Moderada
0.61 - 0.80	Buena
0.81 - 1.00	Muy buena
Fuente: Elaboración Propia	

S= [(*VP*+*VN*)/*N*] La tasa de buena clasificación. Mide la proporción de observaciones que el clasificador predice correctamente la clase positiva y negativa *e*= [(*FN*+*FP*)/N] La tasa de mala clasificación. Es la proporción de observaciones que el clasificador predice incorrectamente.

2.4.4 Sensibilidad

La capacidad de que nuestro modelo estime el suceso de interés de cuyo valor es 1,se denomina sensibilidad.

Es la probabilidad de clasificar correctamente a un alumno desaprobado, capacidad del test para detectar el bajo rendimiento académico

$$S= {}^{VP}/_{(VP+FN)}$$

CAPITULO III

3. METODOLOGÍA

3.1. Tipo de investigación

El tipo de estudio es descriptivo y transversal porque se va a realizar en un solo momento específico del tiempo.

3.2. Diseño de Investigación

Se empleó un diseño observacional, No experimental; debido a que se usó datos a los cuales no se aplicó ningún tipo de tratamiento.

3.3. Método

Modelo no paramétrico de Árboles de decisión (CHAID) y Modelo de Regresión Logística para la comparación.

3.4. Materiales e Instrumentos

Se trabaja con el Software Estadístico IBM SPSS 21.

3.5. Base de datos

Se cuenta con una base de datos sobre el rendimiento académico (factores sociodemográficos y factores académicos) de 3600 registros, en el cual dividiremos la muestra; donde se considera 70% de la muestra para entrenamiento (2520 registros) y el 30% de prueba (1080 registros).

La base de datos se obtuvo de un repositorio de Machine learning sobre estudiantes universitarios del primer semestre matriculados en el curso de Matemática. Los cuales han sido obtenidos de la página https://archive.ics.uci.edu/ml/datasets.html.

Estos datos extraídos para un estudio que se realizó "Técnicas de clasificación aplicadas al estudio del rendimiento de ingresantes universitarios de la Universidad Nacional del Nordeste (UNNE), Argentina. Por (Dapozo.G y Porcel. E).

3.5.1 Muestra de entrenamiento

El conjunto de entrenamiento (para construir el modelo) está conformada por el 70% de estudiantes universitarios del primer semestre matriculados en el curso de Matemática.

3.5.2 Muestra de comprobación

El conjunto de prueba (para evaluar el modelo) está conformada por el 30% de estudiantes universitarios del primer semestre matriculados en el curso de Matemática

3.6. Selección de variables

Variable Dependiente:

Resultado Final del curso

Es la información respectiva de los promedios finales de los alumnos matriculados en los cursos de Matemática respectivamente, pudiendo ser que el alumno aprobó o desaprobó el curso, codificados como 0: Aprobado y 1: Desaprobado.

Variables Independientes: Para establecer las variables independientes se utilizaron diferentes variables sociodemográficas y académicas

$$X' = (x_1, x_2, \dots\dots, x_p).$$

3.7. Descripción de las variables

Cuadro 3.7.1. Descripción de Variables

Factores	Variable	Tipo	Escala de medición	Categorías
Variable dependiente	Resultado Final del curso	Cualitativa	Nominal	0: Desaprobado
				1: Aprobado
Factores sociodemográficos	Sexo	Cualitativa	Nominal	0 : Hombre
				1: Mujer
	Edad	Cualitativa	Ordinal	0: Menos de 25 años
				1: 25 años o Más
	Con quien vive el alumno	Cualitativa	Nominal	0: Con padres y hermanos
				1: Sólo con padre / madre
				2: Sólo
				3:Otros familiares
Factores académicos	Tipo de Colegio	Cualitativa	Nominal	0:Nacional
				1: Particular
	Lugar de preparación	Cualitativa	Nominal	0: Academia
				1: Pre
				2: Casa
				3: Grupo de Estudio
	Satisfacción por el curso	Cualitativa	Nominal	0: Satisfecho
				1: No satisfecho
	Condición laboral	Cualitativa	Nominal	0: Si trabaja
				1: No trabaja
	Tiempo de estudio en el día después de clase	Cualitativa	Ordinal	0: Menos de 3 horas
				1: De 3 horas a más
	Asistencia a clases	Cualitativa	Nominal	0:Si
				1:No
	Nota de examen parcial	Cualitativa	Nominal	0: Menor a 11
				1: Mayor a 11
	Nivel académico del padre	Cualitativa	Ordinal	0: Sin nivel
				1: Primaria
				2: Secundaria
				3: Técnico
				4: Universitario
	Nivel académico de la madre	Cualitativa	Ordinal	0: Sin nivel
				1: Primaria
				2: Secundaria
				3: Técnico
				4: Universitario

Fuente: Elaboración propia

CAPITULO IV

4. RESULTADOS

4.1 ANÁLISIS EXPLORATORIO UNIVARIADO

Cuadro 4.1.1.Analisis Univariado

				Porcentaje (%)
Variable dependiente		Resultado Final del curso (Y)	Desaprobado	51,4
			Aprobado	48,6

Fuente: Elaboración propia

Interpretación:

En el cuadro (4.1.1) se observa que del total de estudiantes, el 51.40% desaprobó el curso de Matemática básica, mientras que el 48.6% aprobó el curso.

Cuadro 4.1.2.Analisis Univariado

			Porcentaje (%)
Factores sociodemográficos	Sexo (X1)	Hombre	54,0
		Mujer	46,0
	Edad (X2)	Menos de 25 años	72,7
		25 años o Más	27,3
	Con quien vive el alumno (X3)	Con padres y hermanos	54,2
		Sólo con padre / madre	24,3
		Sólo	16,1
		Otros familiares	5,4

Fuente: Elaboración propia

Interpretación:

- En el cuadro (4.1.2) el 54% de la base de datos está conformado por varones y el 46% son mujeres.

Cuadro 4.1.3.Analisis Univariado

			Porcentaje (%)
Factores académicos	Tipo de Colegio (X4)	Nacional	71,1
		Particular	28,9
	Lugar de Preparación (X5)	Academia	52,4
		Pre	35,6
		Casa	10,1
		Grupo de Estudio	1,9
	Satisfacción por el curso (X6)	Satisfecho	41,1
		No satisfecho	58,9
	Condición laboral (X7)	Si trabaja	65,3
		No trabaja	54,7
	Tiempo de estudio (X8)	Menos de 3 horas	65,9
		De 3 horas a más	34,1
	Asistencia a clases (X9)	Si	64,5
		No	35,5
	Nota de examen Parcial (X10)	Menor a 11	63,7
		Mayor o igual a 11	36,3
	Nivel académico del padre (X11)	Sin nivel	30,9
		Primaria	19,0
		Secundaria	19,1
		Técnico	19,8
		Universitario	10,2
	Nivel académico de la madre (X12)	Sin nivel	20,7
		Primaria	19,6
		Secundaria	20,4
		Técnico	20,3
		Universitario	19,0

Fuente: Elaboración propia

Interpretación:

- Los estudiantes que provienen de colegios nacionales representan un 71.10%.
- El 65.9% de los estudiantes estudia menos de tres horas luego de clases.

4.2 ANÁLISIS EXPLORATORIO BIVARIADO

Cuadro 4.2.1

Distribución de estudiantes, según su Rendimiento académico y Sexo

			Resultado final del curso		
			Aprobado	Desaprobado	*Total*
Sexo	Hombre	Recuento	899	937	1836
		%	49,0%	51,0%	100,0%
	Mujer	Recuento	851	913	1764
		%	48,2%	51,8%	100,0%

Interpretación:

En el *cuadro* (4.2.1) observamos que del grupo de los desaprobados el 51,0% son de sexo masculino, mientras que el 51,8% son de sexo femenino. En el grupo de los aprobados existe relativamente paridad porcentual entre hombres y mujeres.

Cuadro 4.2.2

Distribución de estudiantes, según su Rendimiento académico y Grupo etario

			Resultado final del curso		
			Aprobado	Desaprobado	Total
Edad	< 25 años	Recuento	1340	1278	2618
		%	51,2%	48,8%	100,0%
	25 a +	Recuento	410	572	982
		%	41,8%	58,2%	100,0%

Interpretación:

En el cuadro (4.2.2) observamos que de los alumnos que desaprueban el curso de Matemática el 48.8% tienen menos de 25 años, este porcentaje aumenta a un 51.2% en el grupo de los aprobados.

Cuadro 4.2.3

Distribución de estudiantes, según su Rendimiento académico y Colegio de procedencia

			Resultado final del curso		
			Aprobado	Desaprobado	Total
CP	Nacional	Recuento	998	1561	2559
		%	39,0%	61,0%	100,0%
	Particular	Recuento	752	289	1041
		%	72,2%	27,8%	100,0%

Interpretación:

De los alumnos que desaprueban el curso de Matemática, el 27.8% estudió en un colegio particular (ver cuadro 4.2.3), de los alumnos aprobados el 72.2% estudio en colegio particular. Observamos claramente la diferencia porcentual en el grupo de los desaprobados siendo en su mayoría (61.0%) alumnos provenientes de colegios nacionales.

Cuadro 4.2.4

Distribución de estudiantes, según su Rendimiento académico y Lugar de preparación

Lugar de Preparación		Resultado final del curso		
		Aprobado	Desaprobado	Total
Academia	Frecuencia	902	983	1885
	%	47,9%	52,1%	100,0%
Pre	Frecuencia	599	683	1282
	%	46,7%	53,3%	100,0%
Grupo de estudio	Frecuencia	203	160	363
	%	55,9%	44,1%	100,0%
Casa	Frecuencia	46	24	70
	%	65,7%	34,3%	100,0%

Interpretación:

En el cuadro (4.2.4) se observa que de los alumnos que no aprueban el curso de Matemática el 53.3% se preparó en una pre, en cambio en el grupo de los aprobados el porcentaje es de 46.7%.

Cuadro 4.2.5

Distribución de estudiantes según su Rendimiento académico y Condición Laboral

			Resultado final del curso		
			Aprobado	Desaprobado	Total
CL	Si trabaja	Recuento	275	439	714
		%	38,5%	61,5%	100,0%
	No trabaja	Recuento	1475	1411	2886
		%	51,1%	48,9%	100,0%

Interpretación:

En el cuadro (4.2.5) se observa que de los alumnos que aprueban el curso de Matemática el 51.1% no trabajan, así mismo de los alumnos que no aprueban el curso Matemática el 61.5% si trabaja.

Cuadro 4.2.6

Distribución de estudiantes según su Rendimiento académico y Satisfacción por el curso

			Resultado final del curso		
			Aprobado	Desaprobado	Total
satisfacción	Si	Recuento	989	490	1479
		%	66,9%	33,1%	100,0%
	No	Recuento	761	1360	2121
		%	35,9%	64,1%	100,0%

Interpretación:

En el cuadro (4.2.6) se observa que de los alumnos que no aprobaron el curso de Matemática el 64.1% no tenía satisfacción por el curso que estudió, así mismo de los que aprueban el curso de Matemática el 66.9% si tenía satisfacción por el curso que estudió.

Cuadro 4.2.7

Distribución de estudiantes según su Rendimiento académico y Nota del Examen Parcial

			Resultado final del curso		
			Aprobado	Desaprobado	Total
Nota EP	< 11	Recuento	1331	1057	2388
		%	55,7%	44,3%	100,0%
	≥ 11	Recuento	419	793	1212
		%	34,6%	65,4%	100,0%

Interpretación:

En el cuadro (4.2.7) se observa que de los alumnos que desaprueban el curso de Matemática el 44.3% desaprobaron el examen parcial con nota menor a 11.

Cuadro 4.2.8

Distribución de estudiantes según su Rendimiento académico y Tiempo de estudio

			Resultado final del curso		
			Aprobado	Desaprobado	Total
Tiempo de estudio	<=3	Recuento	1336	1037	2373
		%	56,3%	43,7%	100,0%
	>3	Recuento	414	813	1227
		%	33,7%	66,3%	100,0%

Interpretación:

En el cuadro (4.2.8) se observa que de los alumnos que desaprueban el curso de Matemática el 66.3% estudia el curso más de tres horas después de clases, este porcentaje disminuye a un 33.7% en el grupo de los que aprobaron el curso de Matemática.

Cuadro 4.2.9

Distribución de estudiantes según su Rendimiento académico y convivencia del alumno

			Resultado final del curso		
			Aprobado	Desaprobado	Total
familia_alumno	Con padres y hnos.	Recuento	426	480	906
		%	47,0%	53,0%	100,0%
	Sólo con padre / madre	Recuento	410	461	871
		%	47,1%	52,9%	100,0%
	Sólo	Recuento	456	483	939
		%	48,6%	51,4%	100,0%
	Otros familiares	Recuento	458	426	884
		%	51,8%	48,2%	100,0%

Interpretación:

En el cuadro (4.2.9) se observa que de los alumnos que aprueban el curso de Matemática el 47.0% convive con padres y hermanos. Así mismo de los alumnos que desaprobaron el curso el 51.4% vive sólo.

Cuadro 4.2.10

Distribución de estudiantes de la FCM-UCI según su Rendimiento académico y Asistencia a clases

			Resultado final del curso		
			Aprobado	Desaprobado	Total
Asistencia	Si	Recuento	1092	1229	2321
		%	47,0%	53,0%	100,0%
	No	Recuento	658	621	1279
		%	51,4%	48,6%	100,0%

Interpretación:

En el cuadro (4.2.10) se observa que de los alumnos que aprueban el curso de Matemática, el 47,0% asiste diario a clases, así mismo el de los alumnos que desaprobaron el curso de Matemática el 53.0% asistió a clases.

Cuadro 4.2.11

Distribución de estudiantes, según su Rendimiento académico y Nivel académico del padre

			Resultado final del curso		
			Aprobado	Desaprobado	Total
N.A.P	Sin nivel	Recuento	369	382	751
		%	49,1%	50,9%	100,0%
	Primaria	Recuento	353	365	718
		%	49,2%	50,8%	100,0%
	Secundaria	Recuento	332	354	686
		%	48,4%	51,6%	100,0%
	Técnico	Recuento	341	374	715
		%	47,7%	52,3%	100,0%
	Universitario	Recuento	355	375	730
		%	48,6%	51,4%	100,0%
Total		Recuento	1750	1850	3600
		%	48,6%	51,4%	100,0%

Interpretación:

En el cuadro (4.2.11) se observa que el 18,8% de los alumnos que aprueban el curso de Matemática, sus padres estudiaron una carrera Técnica, asimismo el 54,4% de los alumnos que desaprobaron el curso de Matemática, sus padres estudiaron el nivel de Secundaria.

Cuadro 4.2.12

Distribución de estudiantes de la FCM -UCI según su Rendimiento académico y Nivel académico de la madre

			Resultado final del curso		
			Aprobado	Desaprobado	Total
N.A.M	Sin nivel	Recuento	335	368	703
		%	47,7%	52,3%	100,0%
	Primaria	Recuento	348	356	704
		%	49,4%	50,6%	100,0%
	Secundaria	Recuento	343	390	733
		%	46,8%	53,2%	100,0%
	Técnico	Recuento	385	369	754
		%	51,1%	48,9%	100,0%
	Universitario	Recuento	339	367	706
		%	48,0%	52,0%	100,0%
Total		Recuento	1750	1850	3600
		%	48,6%	51,4%	100,0%

Interpretación:

En el cuadro (4.2.12) se observa que el 51,1% de los alumnos que aprueban el curso de Matemática, sus madres estudiaron una carrera Técnica, así mismo el 53.2% de los alumnos que desaprobaron el curso de Matemática sus madres alcanzaron el nivel Secundaria.

4.3. ANÁLISIS EXPLORATORIO DE INDEPENDENCIA

Para observar la dependencia de las variables optaremos por utilizar "el test de independencia de Chi-Cuadrado de Pearson"

4.3.1 Variables cualitativas con respecto a Resultado Final del curso

Cuadro 4.3.1.1

Variable / Resultado final del curso	CHI-CUADRADO	P-VALOR
Sexo	0.188	0.665
Edad	25,435	0,000
Tipo de Colegio	327,265	0,000
Lugar de preparación	18,229	0,000
Satisfacción por el curso	335,004	0,000
Condición laboral	36,339	0,000
Tiempo de estudio	164,772	0,000
Asistencia diaria	6,384	0,012
Nota del Examen Parcial	144,182	0,000
Con quien convive el alumno	5,366	0,147
Nivel académico del padre	0,425	0,980
Nivel académico de la madre	3.328	0,504

Fuente: Elaboración propia

4.4 TÉCNICA DE CLASIFICACIÓN: REGRESIÓN LOGÍSTICA

4.4.1. Análisis del Modelo de Regresión Logística Binaria

El método de Regresión Logística se aplicó con la finalidad de predecir, clasificar e identificar las variables que influyen en el bajo rendimiento académico. Para la estimación del modelo se procedió a utilizar solo la muestra de entrenamiento.

El método de selección de variables fue por pasos hacia adelante, el cual comienza con el modelo que incluye solo el término constante y se van añadiendo al mismo modelo las variables independientes según su grado de relación con la variable dependiente y su significación estadística. La que a continuación se presenta:

Cuadro 4.4.1.1. Variables del Modelo

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 1[a]	Edad(1)	-,237	,122	3,758	1	,073	,789
	TipoColegio(1)	-3,968	,178	498,647	1	,000	,019
	Lugar_P			7,210	3	,047	
	Lugar_P(1)	1,141	,450	6,420	1	,011	3,129
	Lugar_P(2)	1,117	,453	6,087	1	,014	3,057
	Lugar_P(3)	,966	,476	4,128	1	,022	2,628
	CondicionLab(1)	,560	,133	17,872	1	,000	1,752
	Satisfaccion(1)	-,451	,502	207,146	1	,000	,637
	NotaEP(1)	1,199	,505	5,638	1	,068	3,317
	Horas_estudio(1)	-3,903	,178	483,077	1	,000	,020
	Asistencia(1)	,071	,110	,411	1	,522	1,073
	Constante	-1,979	,680	8,459	1	,004	,138

De acuerdo al cuadro detallado anteriormente, se puede observar que las variables en mención resultaron significativas (p-valor < 0,05). Las variables: Tipo de colegio, Lugar de preparación, Horas de estudio, Condición laboral y Satisfacción por el curso están incluidas

en el modelo llegando a concluir que influyen sobre el rendimiento académico en el curso de Matemática Básica con una confianza del 95%.

Interpretación:

- Lugar de preparación en una pre tiene una relación positiva con la variable dependiente (1,141)
- Lugar_P(1): La chance o probabilidad de desaprobar el curso de Matemática con preparación en la PRE respecto a una preparación en academia es: 3,129 veces más.
- Horas_estudio (1): La chance o probabilidad de desaprobar el curso de Matemática con más de 3 horas de estudio respecto a menos de 3 horas de estudio disminuye su chance en (1-0,020 =0,980.)
- TipoColegio(1): La chance o probabilidad de desaprobar el curso de Matemática con procedencia de colegio particular respeto a procedencia de colegio nacional disminuye su chance en (1-0,019=0,981).

Entonces el modelo de regresión logística ajustado será:

$$p_i = \frac{1}{1 + e^z}$$

Dónde

Z= -1,979 + 0,071 Asistencia(1)- 3,903 Horas_estudio(1)+ 1,199 NotaEP(1) -0,451 Satisfaccion(1)+0,560 CondicionLab(1)+ ,966 Lugar_P(3)+ 1,117 Lugar_P(2)+ 1,141 Lugar_P(1) - 3,968 TipoColegio(1) -,237 Edad(1)

Del modelo logístico considerado podemos observar que el signo de los coeficientes de algunas variables es positivo, eso significa que la variable aumenta la probabilidad del suceso en estudio, lo que es lo mismo que aumenta la probabilidad de desaprobar el curso de Matemática.

4.4.2. Pruebas Significativas del Modelo

a) Significatividad del Modelo

$$H_{0:}\beta_i = 0$$

$$H_{0:}\beta_i \neq 0$$

Cuadro 4.4.2.1: Pruebas ómnibus de coeficientes de modelo

	Chi-cuadrado	gl	Sig.
Modelo	589,055	12	,000

Fuente: Elaboración Propia

Para contrastar la significatividad global del modelo, se utilizará el estadístico de Razón de Verosimilitud (Prueba Ómnibus).

Se muestra un valor de Chi- Cuadrado de 589.055 con un p-valor de 0,00 <0.05, lo que indica que hay una relación significativa entre las variables independientes y el resultado, es decir, el modelo es significativo.

b) Bondad de Ajuste

H_0: No existen diferencias entre los valores observados y predichos

H_1: Existen diferencias entre los valores observados y predichos

Cuadro 4.4.2.2. Prueba de Hosmer y Lemeshow

Escalón	Chi-cuadrado	gl	Sig.
1	52,091	6	0,221

Fuente: Elaboración Propia

Se muestra un valor de Chi-Cuadrado de 52.091 con un p-valor de 0.221, lo que indica que no existen diferencias entre los valores observados y valores estimados, por lo tanto se puede concluir que el modelo ajustado es significativo.

c) Pseudo Estadísticas R^2

Cuadro 4.4.2.3. Resumen del Modelo

Escalón	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	978,510a	,497	,530

Fuente: Elaboración Propia

Para nuestro caso, según el cuadro muestra un valor de 0.497 que indica que el 49,7% de la variación de la variable dependiente es explicada por las variables incluidas en el modelo (variables independientes) los cuales son: Tipo de colegio, Lugar de preparación, Horas de estudio, Condición laboral y Satisfacción por el curso.

4.4.3. Tabla de Clasificación de Regresión Logística

Cuadro 4.4.3.1.Tabla de clasificación

Observado			Pronosticado		
			Casos no seleccionados: Validación		
			Rendimiento académico		Corrección de porcentaje
			Aprobó	Desaprobó	
Paso 1	Resultado final del curso	Aprobó	448	113	79,9
		Desaprobó	124	395	76,1
	Porcentaje global				78.1

Fuente: Elaboración Propia

Aquellos alumnos que desaprueban el curso de Matemática se estimaron con una precisión de 76,1% (sensibilidad) y aquellos alumnos que aprobaron el curso se estimaron con una precisión de 79.9% (especificidad).

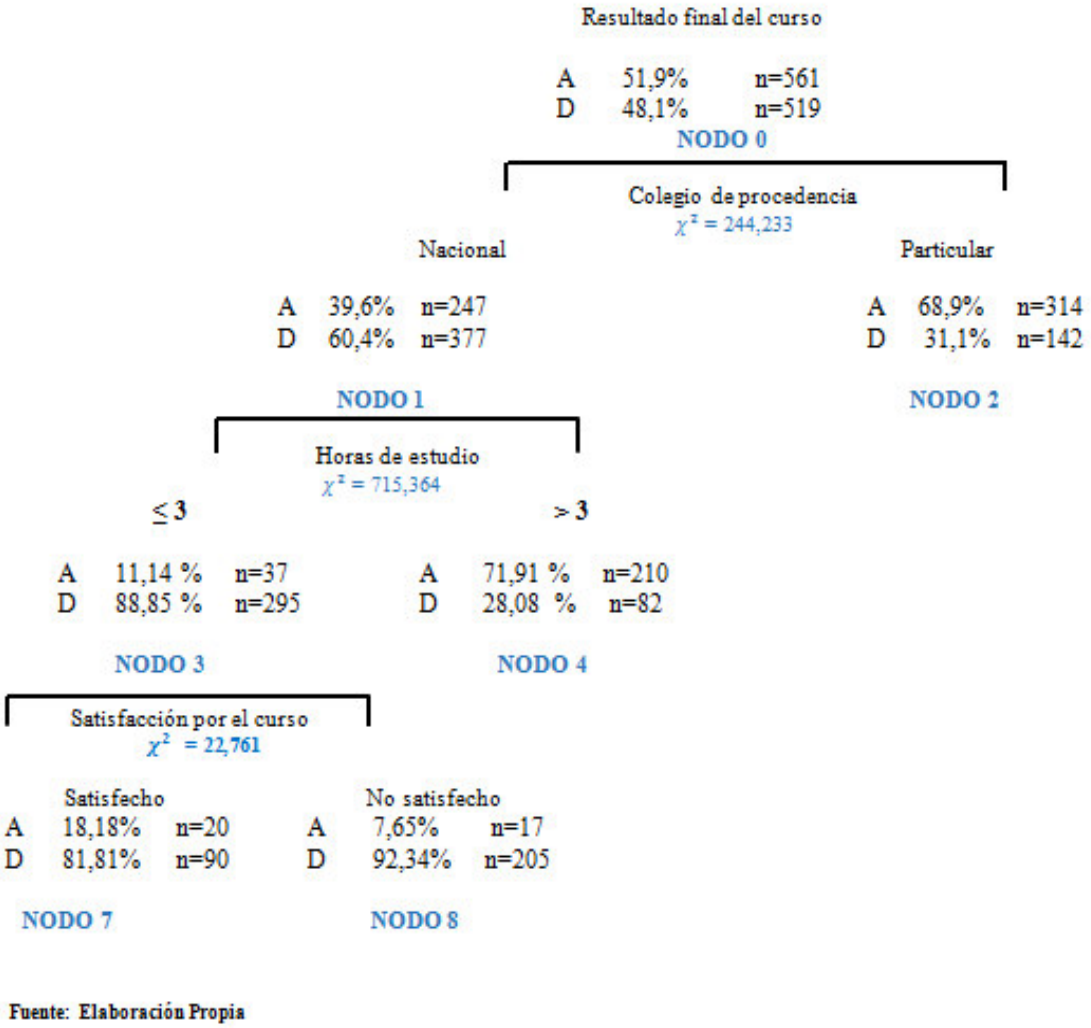
Para evaluar cuan bueno es el modelo para predecir, se utilizó la muestra de prueba y se obtuvo un 78,1% de buena clasificación lo cual indicó que el modelo es bueno para predecir, los alumnos aprobados y desaprobados están correctamente clasificados por el modelo.

4.5 TÉCNICA DE CLASIFICACIÓN: ARBOLES DE DECISIÓN

4.5.1. Análisis del Modelo de Árboles de decisión

De las 13 variables incluidas en el modelo, con el número de casos mínimo aceptable para un nodo padre es de 100 y de nodo hijo 50 y profundidad de 3.

Grafico N° 4.5.1.1: Árbol de Clasificación-Método CHAID (Muestra Prueba)



Fuente: Elaboración Propia

Interpretación

1. En primer lugar, nos fijamos en el nodo 0 que describe la variable dependiente: Rendimiento académico, nos muestra el porcentaje de los que aprueban y desaprueban el curso.

2. Seguidamente observamos que la variable dependiente se ramifica en dos nodos: Nodo 1 y Nodo 2 pertenecientes a la variable Colegio de procedencia, indicando que esta es la variable principal predictora.

3. A continuación, debemos fijarnos en el Nodo 1, ya que su Chi-Cuadrado es superior a la del Nodo 2. Además, nos interesa conocer el perfil de los estudiantes que no aprueban, por ser nuestro objetivo de investigación.

4. El Nodo 1 se vuelve a ramificar en los Nodos 3 y 4 pertenecientes a la variable: Tiempo de estudio. Observamos en el Nodo 3, de los que estudian el curso menos de tres horas, el 88,85%, desaprobó el curso de Matemática Básica frente a un 28,08% del nodo 4 que desaprueban el curso debido a que estudiaron más de tres horas.

5. El Nodo 3 se ramifica en los nodos 7 y 8, pertenecientes a la variable Satisfacción por la asignatura y aquí observamos que a un 81,81% de los estudiantes que no les satisface el curso de Matemática Básica desaprueban.

6. Por tanto, a modo resumen, los nodos que definen el perfil de los estudiantes que no aprueban (variables que influyen en Desaprobar) son: Nodo 0, Nodo 1, Nodo 4 y Nodo 10. Es decir, influyen las siguientes variables: Rendimiento académico, Tipo de colegio de procedencia, Hora de estudio, Satisfacción por el curso.

4.5.2. Tabla de Clasificación de Árboles de Clasificación-Chaid

Cuadro 4.5.2.1.Tabla de clasificación-Chaid

Observado			Pronosticado		
			Casos no seleccionados: Validación		
			Rendimiento académico		Corrección de porcentaje
			Aprobó	Desaprobó	
Paso 1	Resultado final del curso	Aprobó	456	105	81,3
		Desaprobó	116	403	77,6
	Porcentaje global				79.5

Fuente: Elaboración Propia

Aquellos alumnos que desaprueban el curso de Matemática se estimaron con una precisión de 77,6% (sensibilidad) y aquellos alumnos que aprobaron el curso se estimaron con una precisión de 81.3% (especificidad).

Para evaluar cuan bueno es el modelo para predecir este modelo, se utilizó la muestra de prueba y se obtuvo un 79,5% de buena clasificación lo cual indicó que el modelo es bueno para predecir, los alumnos aprobados y desaprobados están correctamente clasificados por el modelo.

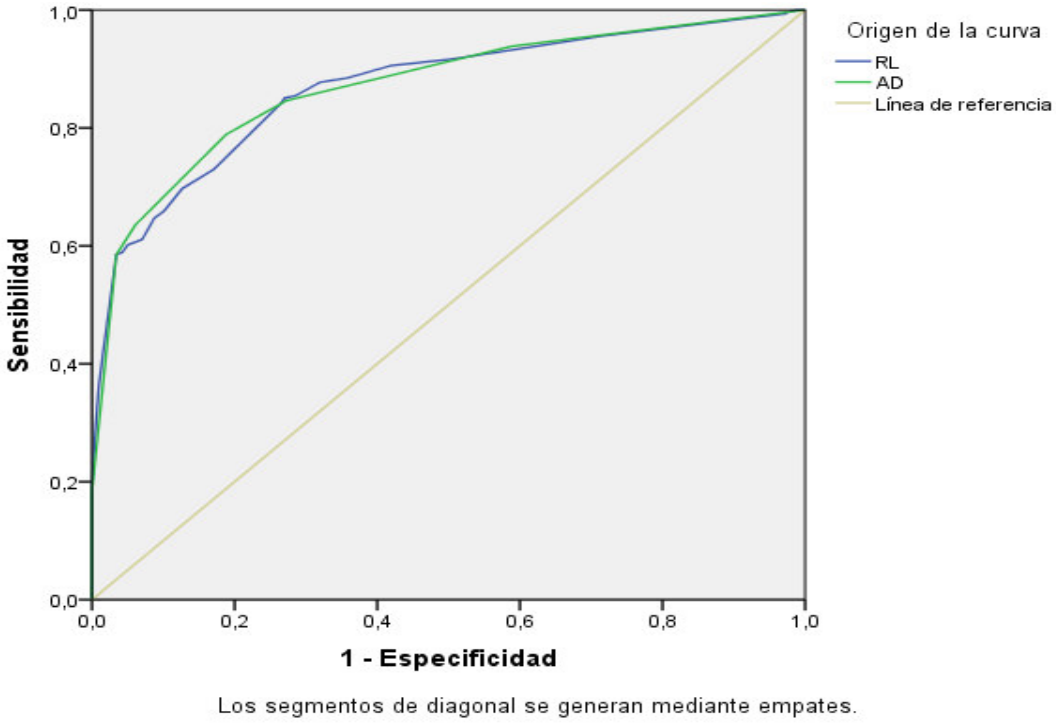
4.6 TÉCNICA DE EVALUACIÓN DE CLASIFICADORES

4.6.1. Comparativo de los Modelos de Regresión Logística y Árboles de Clasificación

Cuadro 4.6.1.1.Clasificadores

	Sensibilidad	AUC	GINI	KAPPA
RL	76,1%	86,7%	73,60%	0.5641244
AD	77,6%	90,1%	80,20%	0.5897954

Cuadro 4.6.1.2.Curva ROC



	Áreas bajo la curva
RL	86,7%
AD	90,1%

Interpretación:

El resultado muestra un AUC = 86,7% con el modelo de Regresión Logística y un AUC = 90,1% con el modelo de Árboles de clasificación siendo ambos buenos para el estudio del rendimiento académico. Por lo que el modelo de Árboles de clasificación tiene un poder de discriminación más alto. En el cuadro (4.6.1.1) observamos que el mejor modelo de clasificación es con la técnica de Árboles de clasificación por tener mayor Sensibilidad=77,6%, AUC=90,1%, GINI=80,2%y KAPPA=58,9%.

CAPITULO V

5. CONCLUSIONES Y RECOMENDACIONES

- Con la técnica de clasificación, Regresión Logística: mediante el estadístico de Wald en el rendimiento académico, las variables más influyentes son: “Tipo de colegio”, “Lugar de preparación”, “Horas de estudio”,”Condición laboral” y “Satisfacción por el curso”.Todas ellas tienen un p-valor asociado al estadístico de Wald menor al 5%.

- Así mismo con la técnica de clasificación, Árboles de decisión: en el caso de las variables influyentes en el rendimiento académico son: “Tipo de Colegio”, “Horas de estudio “y “Satisfacción por el curso”, de los cuales el “Tipo de colegio de procedencia” es la variables principal predictora.

- Las variables que no son influyentes en el rendimiento académico en ambos modelos son: sexo, nivel académico del padre y nivel académico de la madre.

- La variable “tipo de colegio” tiene coeficiente positivo en el modelo logístico, lo que indica el alumno aumenta sus posibilidad de aprobar el curso de Matemática Básica si proviene de un colegio Particular.

- Según la evaluación de la clasificación de los modelos optamos por la Técnica de Árboles de clasificación, siendo la más óptima por tener mayor Sensibilidad=77,6% AUC=90,1%, Gini =80,2% y Kappa=0,589.

- La probabilidad más alta de desaprobar un curso es un 92,34% se da entre los estudiantes que proceden de colegios nacionales, que estudian menos de tres horas y que no tienen satisfacción por la asignatura.

- El aumento de los valores mínimos tiende a generar árboles con menos nodos. La disminución de dichos valores mínimos generará árboles con más nodos. Para archivos de datos con un número pequeño de casos, es posible que, en ocasiones, los valores por defecto de 100 casos para nodos parentales y de 50 casos para nodos filiales den como resultado árboles sin ningún nodo por debajo del nodo raíz; en este caso, la disminución de los valores mínimos podría generar resultados más útiles.

- A medida del criterio que se considere fracaso en el curso (desaprobado) aun estudiante se puede observar que la sensibilidad va disminuyendo y la especificidad va aumentando, en la determinación del punto de corte.

- En estudios futuros se pueden incluir variables psicológicas en los modelos con el fin de aumentar su precisión de la predicción.

- En el presente estudio no se puedo incluir como variable, la performance del alumno en su etapa escolar en los cursos, por lo que se puede considerar esta variable en futuros estudios

- Una de las variables que influyen en el rendimiento académico es Satisfacción por el curso, por lo que se una vez dentro de la universidad se debería incentivar e informar a los alumnos sobre las metodologías de los cursos.

- La técnica de minería de datos Arboles de clasificación demuestra ser una herramientas eficaz para obtener un modelo que permitan predecir sobre el rendimiento académico debería ser empleado en estudios de ámbito educativo.

- Debería realizarse un estudio con estas características estudiadas en nuestra realidad.

BIBLIOGRAFÍA

[1] Álvarez, j. (2011). *Causas endógenas y exógenas del rendimiento académico de los estudiantes de matemática de la facultad de ciencias de la educación de la UNJBG de Tacna.* Ciencia y Desarrollo. Perú – Tacna.

[2] Barahona, p. (2012). *Factores determinantes del rendimiento académico de los estudiantes de la Universidad de Atacama. (Tesis de Maestría).* Universidad de Atacama. Chile.

[3] Hernández, D. (2011). *Impacto de la didáctica en el rendimiento académico universitario*. (Tesis de Licenciatura). Universidad Nacional Federico Villarreal. Perú.

[4] Hosmer. w. d. y lemeshow, *S. Applied Logistic Regression. Second Edition.* A Wiley-Interscience Publication. John Wiley & Sons, INC. 2000.

[5] Izar, j. (2011) .*Factores que afectan el desempeño académico de los estudiantes del nivel superior en Rioverde, San Luis de Potosí, México*. Universidad de Veracruz. México

[6] El Dim Ahmed, A, & Sayed, I. (2014*).Data mining: A prediction for Students Performance Using Classification* Method.World Journal Of computer Application and Technology.Vol 2(2):43-47.

[7] Martinez, H.; Ramirez, G. & Zalazar, L. (2011). *Análisis del bajo rendimiento en matemática de los ingresantes de la facultad de ciencias económicas*. (Tesis Maestría) Universidad Nacional del Nordeste. Argentina.

[8] Bacallao, C., Parapar de la Riestra, M., Roque, M., & Bacallao, J. (2004). *Árboles de regresión y otras opciones metodológicas aplicadas a la predicción del rendimiento académico.* Revista de Educación Médica Superior, Vol. 18, Nº 3.

[9] Porto, A.; Di Gresia, L., & López A, M. (2004). *Mecanismos de admisión a la Universidad y rendimiento de los estudiantes.* Departamento de Economía, Universidad Nacional de La Plata.

[10] Hernández Orallo, J., Ferri Ramírez, C. y Ramírez Quintana M. J. "*Introducción a la Minería de Datos"*. España: Prentice Hall. Pearson Education.2004

[11 Porcel, E., Dapozo, G. y López, M. (2010). *Predicción del rendimiento académico de alumnos de primer año de la FACENA (UNNE) en función de su caracterización socioeducativa. Revista Electrónica de Investigación Educativa, 12*(2). Consultado el 2 de Julio de 2017 en: http://redie.uabc.mx/vol12no2/contenido-porceldapozo.html

ANEXOS

Figura A.1.Parte de la Base de datos sobre rendimiento académico

	RF	Sexo	Edad	TipoColegio	Lugar_P	CondicionLab	Satisfaccion	NotaEP
1196	Desaprobado	Hombre	< 25 años	Nacional	Academia	No trabaja	Si	Mayor a 11
1197	Desaprobado	Hombre	26 a +	Nacional	Pre	No trabaja	Si	Mayor a 11
1198	Aprobado	Hombre	26 a +	Particular	Academia	Si trabaja	No	Mayor a 11
1199	Aprobado	Hombre	< 25 años	Nacional	Pre	No trabaja	Si	Menor a 11
1200	Aprobado	Hombre	< 25 años	Nacional	Academia	No trabaja	Si	Mayor a 11
1201	Aprobado	Hombre	< 25 años	Particular	Pre	No trabaja	No	Mayor a 11
1202	Desaprobado	Hombre	< 25 años	Nacional	Academia	Si trabaja	No	Menor a 11
1203	Desaprobado	Hombre	< 25 años	Nacional	Pre	Si trabaja	No	Menor a 11
1204	Aprobado	Hombre	26 a +	Particular	Academia	Si trabaja	No	Mayor a 11
1205	Desaprobado	Hombre	< 25 años	Particular	Academia	Si trabaja	No	Menor a 11
1206	Desaprobado	Hombre	< 25 años	Nacional	Academia	Si trabaja	No	Menor a 11
1207	Desaprobado	Hombre	< 25 años	Nacional	Academia	Si trabaja	No	Menor a 11
1208	Aprobado	Hombre	< 25 años	Particular	Pre	No trabaja	No	Mayor a 11
1209	Desaprobado	Hombre	< 25 años	Nacional	Academia	Si trabaja	No	Mayor a 11
1210	Desaprobado	Hombre	< 25 años	Particular	Academia	Si trabaja	Si	Menor a 11
1211	Desaprobado	Hombre	< 25 años	Particular	Pre	Si trabaja	No	Menor a 11
1212	Desaprobado	Hombre	< 25 años	Nacional	Academia	Si trabaja	No	Menor a 11
1213	Aprobado	Hombre	< 25 años	Particular	Academia	No trabaja	No	Mayor a 11
1214	Aprobado	Hombre	< 25 años	Particular	Academia	No trabaja	No	Menor a 11
1215	Desaprobado	Hombre	< 25 años	Particular	Academia	Si trabaja	No	Menor a 11
1216	Desaprobado	Hombre	< 25 años	Nacional	Pre	Si trabaja	Si	Menor a 11
1217	Desaprobado	Hombre	26 a +	Nacional	Academia	Si trabaja	No	Menor a 11

Figura A.2 Codificaciones de variables categóricas- Regresión Logística

			Codificación de parámetro		
		Frecuencia	(1)	(2)	(3)
LP	Academia	565	,000	,000	,000
	Pre	382	1,000	,000	,000
	Grupo de estudio	106	,000	1,000	,000
	Casa	27	,000	,000	1,000
Asistencia	Si	716	1,000		
	No	364	,000		
Edad	< 25 años	784	1,000		
	26 a +	296	,000		
CP	Nacional	751	1,000		
	Particular	329	,000		
CL	Si trabaja	213	1,000		
	No trabaja	867	,000		
Satisfaccion	Si	456	1,000		
	No	624	,000		
Horas de estudio	3	725	1,000		
	3	355	,000		
Nota EP	Mayor a 11	354	1,000		
	Menor a 11	726	,000		
Sexo	Hombre	526	1,000		
	Mujer	554	,000		

Figura A.3.Modelo de árbol de la muestra de entrenamiento

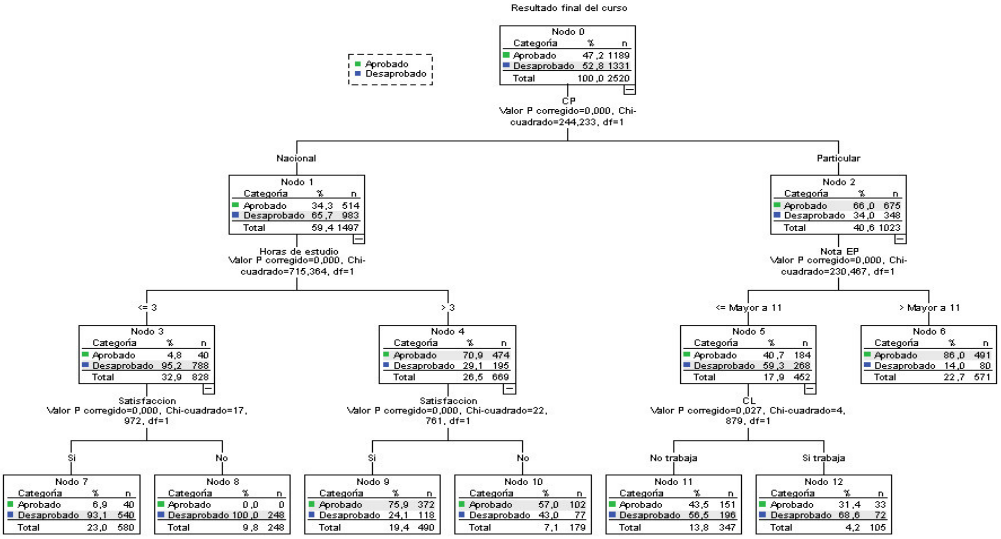


Figura A.4. Modelo de árbol de la muestra de validación