UNIVERSIDAD POPULAR AUTÓNOMA

DEL ESTADO DE PUEBLA

DECANATO DE INGENIERÍAS

Doctorado en Tecnologías de Información y Negocios Electrónicos

Construcción de modelos predictivos de la deserción universitaria utilizando minería de datos, caso de estudio: CETYS Universidad campus Ensenada

Tesis que para obtener el grado de

Dra. Tecnologías de Información y Negocios Electrónicos

Presenta

Lucía Beltrán Rocha

Matrícula 14240014

Directora de Tesis

Dra. Rosa María Croda Cantón

Co-director de tesis

Dr. Damián Emilio Gibaja Romero

Mayo del 2022

Quiero agradecer:

A mis padres por su fe y amor incondicional. Siempre en mi corazón. Viejita, esto te lo dedico a ti.

A mi esposo Leonardo, por su acompañamiento, por sus enseñanzas y por su presencia permanente en esta ruta de mi formación profesional, compleja y llena de retos.

A mis maravillosos hijos: Fausto, Alejandra y Carlos, por su amor y su existencia, son el motor que me impulsa y me impulsará hasta el último día.

Mi especial agradecimiento a mi hermana favorita, Lupita, por su apoyo y fortaleza, por siempre ser y estar y a mi sobrina/ahijada Lucia, por estar presente y atenta a esta experiencia de vida.

A toda mi familia, somos un gran equipo.

A mi Institución CETYS, por la confianza e impulso en mi desarrollo profesional.

A mis amigos y amigas, muchas gracias por su cariño.

Resumen

El propósito del estudio es diseñar y construir un modelo que permita predecir la deserción, a través de herramientas de minería de datos y de algoritmos predictivos. El caso de estudio es en uno de los campus de un Sistema Universitario (IES) multicampus privado en el noroeste del país. Los datos que se analizaron fue la información cuantitativa y cualitativa histórica de los estudiantes que se dieron de baja del campus del 2008 al 2018. Se diseñó y construyó un modelo lógico de una base de datos, a través de un proceso de ETL se almacenaron un total de 355 instancias, cada una representando a un desertor con 102 atributos personales y académicos, que los caracterizaron antes y durante su estancia en la IES. Se aplicaron algoritmos supervisados como regresión logística (RL) y bosque de árboles (RF), para la construcción de modelos predictivos, demostrando que hay una correlación entre las variables que estos modelos identificaron como predictoras.

Se puede concluir que los desertores entraron con un desempeño regular, la mayoría eligieron programas académicos de la escuela de Administración y Negocios y obtuvieron una beca no asociada a la excelencia académica, una gran proporción de estos desertores son de los primeros semestres. Ambos modelos coinciden en su capacidad de predecir aquellos estudiantes que se convertirán en desertores, mejor que la capacidad para detectar a los que se dieron de baja para migrar a otros campus y por consecuencia se quedaron en el Sistema CETYS. Se considera la métrica de Recall o sensibilidad como la más relevante; 95.45% para el modelo de RL y de 94.93% para el modelo de RF, no así la capacidad de predecir a los que permanecerán en el Sistema, con una métrica de especificidad del 40% para el modelo de RL y de 53.3% para el modelo de RF. La métrica de armonía F1, es del 85.13% para el modelo de RL y de 87.33% para el modelo de RF, es una buena métrica para ambos modelos.

Abstract

The study aims to design and build a model that allows for predicting desertion through data mining tools and predictive algorithms. The case study is on one campus of a private multi-campus University System (IES) in the northwest Mexico country. The analyzed data was historical quantitative and qualitative information of the students who dropped out of the campus from 2008 to 2018. A logical model database was designed and built, through an ETL process, 355 instances, each representing a dropout with 102 personal and academic attributes which

characterized them before and during their stay at the IES. Applied supervised algorithms such as logistic regression (RL), and forest of trees (RF) were applied to build predictive models, demonstrating a correlation between the variables that these models identified as predictors.

Can be concluded that the dropouts entered with a regular performance; the majority chose academic programs from the Business and Administration school and obtained a scholarship not associated with academic excellence. A large proportion of these dropouts are from the first semesters. Both models coincide in their ability to predict those students who will become dropouts, better than their ability to detect those who dropped out to migrate to other campuses and consequently stayed in the CETYS System. Recall or sensitivity metric is considered the most relevant; 95.45% for RL model and 94.93% for the RF model, but not the ability to predict those who will remain in the System, with a specificity metric of 40% for the RL model and 53.3% for the RF model. The harmony metric F1, which is 85.13% for the RL model and 87.33% for the RF model, is a good metric for both models.

Palabras clave en español e inglés

Deserción Universitaria, Minería de datos, Aprendizaje de máquina, Modelos predictivos, Minería de datos educativa.

Undergraduate Dropout, Data Mining, Machine Learning, Predictive models, Educational Data Mining.

Índice General

Índice de tablas

Índice de figuras

Índice de gráficas

Introducción

Prevenir la deserción, parcial o total, en las Instituciones de Educación Superior (IES) públicas y privadas es de vital importancia debido a la relevancia de la educación superior en un país en desarrollo como lo es México. Aunque México ha tenido un incremento en la cobertura del nivel educativo de educación superior desde 1950, pasando de un millón a 36 millones de estudiantes (Organización para la Cooperación y el Desarrollo Económicos [OECD], 2019), la tasa de abandono escolar en México se ha incrementado en los últimos 10 años de un 7% a un 8.2% (Instituto Nacional de Estadística y Geografía [INEGI], 2021). Además de enfocarnos en el caso particular de nuestro caso de estudio, es importante resaltar que, en Baja California, la deserción en el nivel universitario pasó del 1.1 en (2005/2006) a 6.1 en (2020/2021).

La OCDE estima en 40% el promedio de adultos entre 25 y 64 años con una licenciatura en el 2020 a nivel mundial. En dicho año, México tenía un 19%, estando por debajo del 20% de Brasil y del 25% que reportan Chile y Colombia. Es importante notar que Canadá fue el país con la mayor cobertura con un 60%, mientras que China es el peor de todos los países de la OCDE pues sólo el 10% de su población adulta tiene licenciatura. Notemos que este indicador mide indirectamente el capital humano, y representa una señal del nivel de las habilidades de un individuo; y se asocia con resultados económicos y sociales positivos para las personas y para los países(*Education at a Glance 2021: OECD Indicators*, 2021). La deserción total se define como "el porcentaje de alumnos que abandonan las actividades escolares durante el ciclo escolar (desertores intracurriculares) y al finalizar éste (desertores intercurriculares) respecto al total de alumnos inscritos en el ciclo escolar" (Tamez et al., 2006).

El propósito del estudio es diseñar y construir un modelo que permita predecir la deserción, a través de herramientas de minería de datos y aplicación de algoritmos predictivos, para ello se busca identificar la relación entre las variables que definen la deserción desde que el estudiante ingresa a la IES hasta que decide truncar sus estudios.

El caso de estudio es uno de los campus de un Sistema Universitario multicampus privado en el noroeste del país, el cual fue fundado en 1961, por un grupo de empresarios conformando el Consejo del Instituto Educativo del Noroeste A.C. quien auspicia a esta Universidad (IENAC)

(*Quiénes Somos - Centro de Enseñanza técnica y superior [CETYS] Universidad*, n.d.). Cuenta con tres campus distribuidos en tres ciudades: Mexicali, Tijuana y Ensenada, siendo este último el campus de estudio, siendo además el más pequeño. La institución actualmente atiende a más de 7,800 estudiantes a nivel sistema, sin contar los estudiantes del área de Educación Continua. Y sus estadísticas muestran que a la fecha hay más de 37,000 egresados. En el caso de campus Ensenada, se cuenta con 397 estudiantes de profesional y alrededor de 6,000 egresados (CETYS, 2018).

Se aplicaron algoritmos supervisados para la identificación de patrones de clasificación y correlación entre los atributos que caracterizan a los estudiantes desertores de la IES. Los datos a analizar fueron la información cuantitativa y cualitativa histórica que los estudiantes desertores generaron durante su estancia en la IES, también se consideró información previa como sus atributos personales y académicos.

Se identificaron algunos datos ausentes que son relevantes para un estudio de deserción que la Institución no colecta a pesar de tener acceso a ella al momento de inscribir a un nuevo estudiante o durante su estancia académica, logrando generar una serie de recomendaciones para efecto de consolidar bases de datos con información que aporte elementos para la toma de decisiones y diseño de estrategias de retención.

Se probó el modelo predictivo en retrospectiva con información histórica y estimando resultados futuros, con la información generada en el campus objeto del estudio, no se logró validar el modelo predictivo en los otros dos campus del Sistema de la IES, ya que la deserción de los últimos dos años ha cambiado debido a la situación de pandemia por el COVID-19, siendo el sector educativo uno de los más afectados por la ausencia de los estudiantes en el aula para desarrollar actividades de aprendizaje para el desarrollo de competencias vitales para el logro de las competencias y el desarrollo académico, así como la situación de salud mental en estos y sus familias, por lo que se descartó la posibilidad de probar los modelos debido a que se deben considerar variables adicionales con las que no se contaban dentro de la muestra de datos colectadas para este estudio.

Se consideró el tema de migración entre los campus del mismo Sistema educativo un fenómeno que debe estudiarse ya que se pudo detectar que hay una cantidad de desertores-migrantes que no son considerados como desertores pero que afectan a ciertos indicadores como

la eficiencia terminal, la poblacional estudiantil, las finanzas del campus entre otras, por lo que se recomendó entender y caracterizar la migración para trabajos futuros.

Capitulo 1 Propósito y Organización

La deserción es un problema que se presenta en todos los niveles educativos de la República Mexicana, siendo los niveles medio superior y superior los que tienen mayor índice (Secretaria de Educación Pública, 2020) con un 15.2% en educación media superior comparada con un 5.8% en educación secundaria y un 0.8% en educación primaria, esta investigación tiene el propósito no solo de identificar las causas que inciden en la deserción universitaria sino proponer modelos predictivos que permitan a las instituciones construir estrategias de intervención temprana para reducirla.

En el presente capítulo se plantean los objetivos generales y particulares, las preguntas de investigación, el alcance y las limitaciones. Se utilizará como caso de estudio el de los desertores de los programas académicos de Licenciatura de uno de los tres campus de una IES privada en el noroeste de México. Este estudio caracteriza la deserción tomando como fuente de datos todos los alumnos desertores desde agosto de 2008 hasta agosto 2018, con ello, se sustenta que fue lo que sucedió para que los estudiantes desertaran.

1.1 Planteamiento del problema

Se considera que la deserción se puede caracterizar, de tal forma que, a través de metodologías, técnicas y herramientas de ciencia de datos, como la minería de datos, se puede construir un modelo que pueda predecir cuándo un estudiante es susceptible de desertar, con ello, las instituciones podrían diseñar e implementar estrategias de retención exitosas y puntuales.

Las instituciones académicas cuidan dentro de sus procesos, algunos indicadores de calidad educativa que les permiten medir el cumplimiento de su misión y objetivos (Sosa, 2016). El Consejo para la acreditación de la Educación Superior, A.C. (COPAES), define una serie de categorías a valorar por los organismos acreditadores, dentro de los cuales se encuentra la categoría de Estudiantes, misma que considera una serie de criterios e indicadores de eficiencia como: rezago, deserción, eficiencia terminal, resultados del EGEL-CENEVAL y la titulación (Marco General de Referencia Para Los Procesos de Acreditación de Programas Académicos de Tipo Superior Ver 3.0, 2016). Es por ello que para cualquier institución educativa es importante

implementar estrategias de retención y éxito académico, siendo la deserción todo aquello que dará como resultado que el estudiante no culmine sus estudios en esa institución educativa.

El caso de estudio que se abordará es el de CETYS Universidad, una Institución educativa privada sin fines de lucro en el noroeste del país, que cuenta con 30 programas académicos de nivel bachillerato, licenciatura y posgrado, estos programas se encuentran bajo la administración de 3 Colegios: Ingeniería, Administración y Negocios, y Sociales y Humanidades.

Según cifras oficiales, el porcentaje de graduación en campus Ensenada para la escuela de Administración y Negocios y la escuela de Ingeniería se muestra en la tabla 1. Se puede observar los porcentajes de deserción están por arriba de la meta planteada del 30% por Rectoría en el plan 2020 (CETYS, 2011).

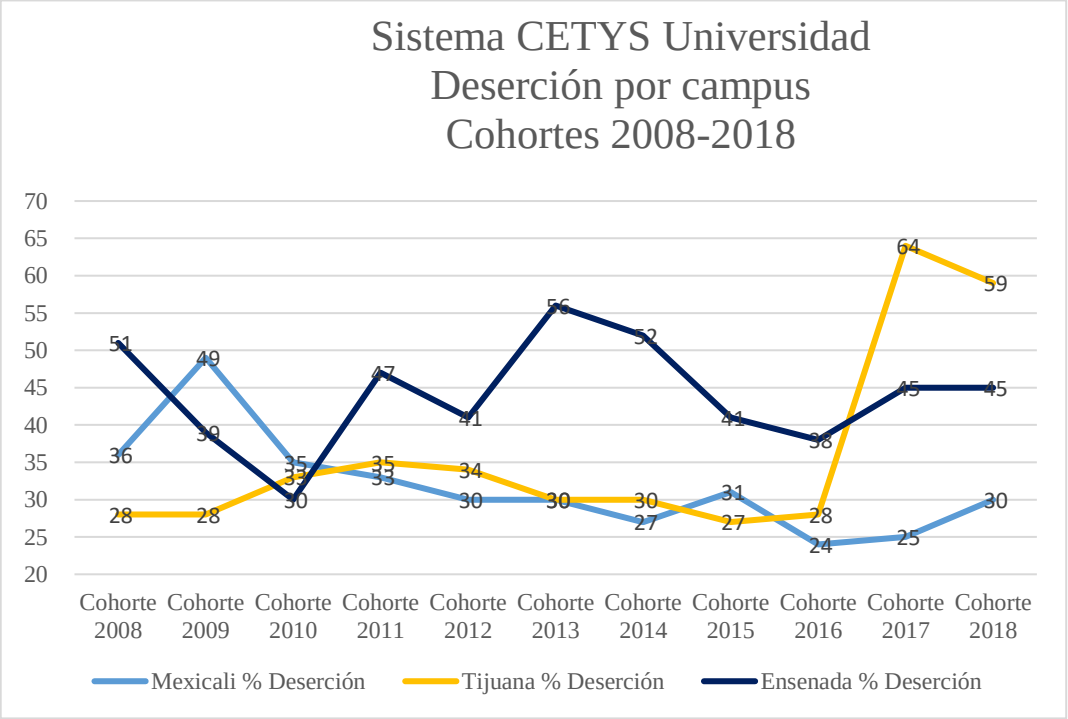
Tabla 1-1 Porcentaje de deserción de campus Ensenada (CETYS, 2018).

Escuela	**Administración y Negocios**			**Ingeniería**			**Campus Ensenada**		
Indicadores	**Total Alumnos**	**Número de Desertores**	**Porcentaje de Deserción**	**Total Alumnos**	**Número de Desertores**	**Porcentaje de Deserción**	**Total Alumnos**	**Número de Desertores**	**Porcentaje de Deserción**
Cohorte 2008	39	18	46%	24	14	58%	63	32	51%
Cohorte 2009	39	15	38%	34	14	41%	73	29	39%
Cohorte 2010	94	28	30%	94	28	30%	188	56	30%
Cohorte 2011	64	29	45%	35	18	51%	99	47	47%
Cohorte 2012	33	15	45%	47	18	38%	80	33	41%
Cohorte 2013	**46**	**29**	**63%**	47	23	49%	93	52	56%
Cohorte 2014	82	33	40%	**43**	**32**	**74%**	125	65	52%
Cohorte 2015	63	29	46%	87	33	38%	150	62	41%
Cohorte 2016	99	42	42%	98	32	33%	197	74	38%
[1]Cohorte 2017	43	11	26%	59	35	59%	102	46	45%
Cohorte 2018	70	35	50%	116	68	59%	186	103	45%
TOTALES	**672**	**283**	**42%**	**684**	**315**	**46%**	**1356**	**599**	**44%**

[1] Las cifras de las cohortes 2017 y 2018, corresponden a la deserción del segundo año.

En un análisis previo de la IES, no fue posible identificar las causas de deserción tan significativa entre los años 2013 y 2014 para Ingeniería y entre los años 2012 y 2013 para Administración y Negocios. Como se observa en la Tabla 1-1, en los años 2013 ambas escuelas presentan porcentajes altos de deserción, y en los años 2017 la escuela de Administración y Negocios muestra mejores cifras que la escuela de Ingeniería, y ésta a su vez presenta una importante mejora en los años 2015 y 2016.

Incluso, si comparamos la deserción con los otros campus del Sistema CETYS, vemos que los campus Mexicali y Tijuana presentan más baja deserción que el campus Ensenada, con excepción de las Cohorte 2017 y 2018 de campus Tijuana y la cohorte 2009 de campus Mexicali, el detalle se puede apreciar en la gráfica 1-1.



Gráfica 1-1 Deserción Sistema CETYS Cohortes 2008-2018 (CETYS, 2018)

Cuando un estudiante se da de baja de la Institución, contesta una encuesta de salida, en ésta manifiesta las causas de su baja. Es común que indique que su baja es por problemas financieros, vocacionales, cambio de campus entre otros, cuando se da este momento, el estudiante ya pasó por un proceso de toma de decisiones debido a circunstancias que la Institución no tuvo control o que sucedieron sin que ésta pudiera hacer algo al respecto, sin embargo, si se contara con las herramientas para observar el comportamiento previo o tener caracterizado al desertor, se podría predecir como un estudiante con riesgo de baja y la IES tendría la capacidad de hacer una intervención temprana.

1.2 Justificación de la investigación

Según el resumen de la estadística de alumnos del 2016-2017 publicada por la Secretaría de Educación Pública (SEP), del total de 3´352,256 alumnos de Educación Superior, en licenciatura

un total de 983,777 alumnos pertenecen al sistema privado, el 29.4% (Secretaria de Educación Pública, 2018). La Secretaría de Educación Pública calcula una serie de indicadores y pronósticos educativos, dentro de los cuales, tres hablan de eficacia: la reprobación, la deserción total y la eficiencia terminal; la deserción total, se define como "el porcentaje de alumnos que abandona las actividades escolares durante el ciclo escolar (desertores intracurriculares) y al finalizar éste (desertores intercurriculares) respecto al total de alumnos inscritos en el ciclo escolar " (Tamez et al., 2006)

De acuerdo a las estadísticas de deserción por entidad federativa, del 2003/2004, en el nivel licenciatura, Baja California ocupaba el 4to lugar a nivel nacional, con un 4.5 % (Tamez et al., 2006). La deserción en las IES, es un tema que se ha abordado desde muchas aristas: el historial académico, el desempeño académico de los estudiantes, la situación económica de las familias, las estructuras y políticas de las Instituciones, las oportunidades de desarrollarse de forma exitosa en el estudio entre otras (Fonseca & García, 2016).

Se utilizará como muestra poblacional, el estudiantado de los programas de licenciatura del campus Ensenada de la Institución educativa privada, CETYS Universidad o Instituto Educativo del Noroeste, la cual se encuentra en el estado de Baja California, México; en la ciudad de Ensenada. Esta Institución se encuentra afiliada a la SEP, a la Federación de Instituciones Mexicanas Particulares de Educación Superior (FIMPES), a la Asociación Nacional de Universidades e Instituciones de Educación Superior (ANUIES) y cuenta con acreditaciones nacionales e internacionales como CACECA, CACEI, WASC, ABET y ACBSP y COMAPROD.

El caso que se analizará, se eligió por la importancia que representa poder determinar porque se han dado los niveles de deserción en campus Ensenada, fundamentalmente porque en los otros campus esta situación no se presenta, en campus Tijuana en promedio tiene un porcentaje de retención del 52% para la escuela de Administración y Negocios y de 59.7% para la de Ingeniería, y en campus Mexicali un promedio de hasta 54.1% para la escuela de Administración y Negocios y de 56.9% para la escuela de Ingeniería, como se puede observar, en campus Mexicali y Tijuana los porcentajes están por debajo de la meta Institucional, pero presentan cifras más elevadas que en campus Ensenada, saber que causas son las que realmente provocan que el estudiante decida no continuar con sus estudios en la Institución, y con ello, poder determinar las estrategias que se deben implementar para lograr una mayor retención y éxito académico. El tema de deserción es una preocupación creciente a nivel local, regional, nacional y en América Central y en América

Latina, porque las consecuencias son relevantes para la productividad económica, el desarrollo y tiene una repercusión directa en lo social, según indica el Banco Mundial en su resumen de tendencias, causas y consecuencias e intervención (Adelman & Székely, 2016).

1.3 Objetivos de la investigación

1.3.1 Objetivo General

El propósito de la investigación es diseñar uno o más modelos predictivos a través de la minería de datos, que permitan identificar los casos de riesgo de deserción, para la efectiva toma de decisiones, en el diseño, implementación y seguimiento de estrategias de retención estudiantil en CETYS Universidad.

1.3.2 Objetivos específicos

1. Recolectar la información de los estudiantes que se dieron de baja de agosto del 2008 a agosto del 2018, de las diferentes fuentes de datos disponibles, diferenciando aquellos de los que permanecieron en Sistema CETYS de los que se dieron de baja de forma definitiva.
2. Caracterizar los desertores, desde su ingreso, su estancia y la salida de la Institución, identificando la correlación o asociación entre estos atributos.
3. Construir un modelo predictivo de la deserción, que permitan identificar los atributos predictores más significativos y con ello poder identificar a un estudiante en riesgo n.
4. Comprobar el modelo predictivo, seleccionando un conjunto de datos históricos de alumnos desertores y no desertores a los cuales se les aplicaría el modelo.
5. Se probará el o los modelos con los datos de estudiantes activos de algún semestre en curso, con la intención de validar al final de este tiempo si el modelo predijo a los desertores de forma correcta o no.

1.4 Preguntas de investigación

1. Qué atributos cualitativos y/o cuantitativos caracterizan a los desertores, que expliquen por qué los estudiantes abandonaron sus estudios?
2. ¿Qué variables influyen en la deserción de los estudiantes de licenciatura de campus Ensenada en el período agosto 2008 a agosto 2018?
3. ¿Existe una correlación multivariada entre las variables que influyen en la deserción?

4. ¿La minería de datos permite construir un modelo que predigan la deserción de los estudiantes de Licenciatura de CETYS Universidad campus Ensenada?
5. ¿El modelo diseñado predice la deserción de los estudiantes de todos los programas académicos de licenciatura de campus Ensenada?
6. ¿Se puede validar el modelo en otros campus de la misma Institución?

1.5 Hipótesis

H1: Hay una correlación de variables multivariadas causales de la deserción.

H2: Se puede predecir la deserción de los alumnos de Licenciatura de CETYS Universidad, campus Ensenada, con uno o más modelos.

H3: El modelo predictivo puede predecir la deserción en otros campus de CETYS Universidad.

1.6 Alcances y limitaciones del proyecto

Describir la deserción de los estudiantes de licenciatura de CETYS Universidad, campus Ensenada, a través de la correlación entre las variables que la caracterizan, construyendo uno o más modelos predictivos. Estos modelos permitirán predecir la deserción, para ello, se requiere analizar la información cuantitativa y cualitativa con la que se cuenta, misma que el estudiante va generando desde que es un prospecto, por los atributos de su perfil de ingreso; durante su estancia en la IES a través de su historial académico, financiero y extracurricular, y por las condiciones y atributos del mismo estudiante cuando tramita una baja de forma definitiva.

Se busca explicar la deserción de los estudiantes de Licenciatura de CETYS Universidad, tomando como muestra la población de campus Ensenada, en el período de agosto de 2008 a agosto de 2018. La deserción, para efectos de ésta investigación, está considerándose como el momento en el que el estudiante deja de pertenecer al sistema educativo de la IES. En esta investigación no se consideran las variables cualitativas ni cuantitativas, fuera del contexto de la IES, como la situación de vida del estudiante en el hogar, en el trabajo o en algún espacio y actividad deportiva, política o social que no fuere el de la IES, a menos que ésta sea una razón explicita y manifiesta de su deserción al momento de darse de baja de la Institución.

Capitulo 2 Marco Teórico

2.1 Introducción

La presente investigación contemplará los siguientes componentes teóricos para su desarrollo: La deserción, la minería de datos y la deserción explicada a través de técnicas de minería de datos. Es importante identificar de dónde surgen las técnicas de análisis para la toma de decisiones, el contexto histórico y la necesidad, cada vez más inminente, de la aplicación de técnicas y herramientas computacionales, estadísticas y de analítica de datos, ante la oportunidad de contar con grandes volúmenes de información para la identificación de patrones de comportamiento de las variables significativas que pueden o no explicar algún fenómeno dentro de las organizaciones, que finalmente se traducen en oportunidades, para diseñar estrategias que favorezcan o disminuyan aquellos factores que afectan la productividad y la eficiencia.

La deserción se ha intentado explicar desde mucho tiempo atrás, de esto, han surgido modelos teóricos que permiten contextualizar, organizar y clasificar las variables que deben ser tomadas en cuenta para caracterizarla, explicarla y en el resultado más deseable, predecirla. Con la aplicación cada vez más frecuente de minería de datos, se han diseñado una cantidad importante de algoritmos que pueden ser utilizados para el análisis descriptivo y predictivo de un fenómeno, sin embargo, es importante darles relevancia a todas las etapas de la metodología de minería de datos, para garantizar la calidad e integridad de los datos, con la finalidad de que los resultados obtenidos sean confiables y pertinentes para la toma de decisiones.

2.2 Tendencias teóricas de la minería de datos

Desde la década de los 60 se hablaba del aprendizaje automático, ya se empezaba a trabajar en el reconocimiento de patrones y aprendizaje de máquina, como lo abordó Fu, K.S (1968). en su libro *Sequential methods in pattern recognition and machine learning*. Estos métodos tienden a optimizar el rendimiento predictivo directamente, mediante el manejo automático de la no linealidad, de datos ruidosos, de una gran cantidad de predictores potenciales, entre otros. El énfasis de estas herramientas radica más en la predicción que en la explicación e interpretación de los efectos de las covariables (Tollenar & Van der Heijden, 2013).

En 1996 Fayad et al., hablan del campo emergente de extracción del conocimiento de bases de datos y de minería de datos y del problema de la interpretación de los grandes volúmenes de

datos que se estaban generando en la era de la información digital. En principio, la minería de datos es una fase dentro de un proceso global denominado descubrimiento de conocimiento con base en datos (Knowledge Discovery in Database –KDD por sus siglas en inglés), aunque generalmente se asocia el concepto de minería de datos a todo el proceso, en lugar de la fase de extracción de conocimiento. En otras palabras, la minería de datos se refiere a la extracción o el conocimiento "minería" de grandes cantidades de datos (Fayyad et al., 1996).

Frecuentemente se tiende a creer que es lo mismo análisis de negocio, minería de datos e inteligencia de negocios. Sin embargo, son distintos, ya que la analítica de negocio es una estrategia que consiste en la utilización de los datos procesados y analizados para aprovechar la información y está íntimamente relacionada con el manejo de grandes volúmenes de datos (Big Data) (Goldman, 2015), mientras que la minería de datos, es el motor del análisis del negocio (Lee, 2013), las diferentes técnicas y algoritmos de la minería de datos se implementan actualmente en muchas industrias, con la finalidad de identificar los datos significativos del negocio, que permitan analizarlos, evaluarlos, integrarlos y generar indicadores que funcionen como el termómetro del negocio en sus diferentes áreas y niveles de toma de decisiones, finalmente la inteligencia de negocios, es la unión del intelecto humano y las nuevas tecnologías de minería de datos, en la gestión y toma de decisiones en los diferentes problemas de la organización (Mishra et al., 2016).

Actualmente, las grandes empresas utilizan herramientas de inteligencia de negocios para analizar grandes cantidades de información y tomar decisiones adecuadas en el negocio. De acuerdo con Rodríguez et. al. (2016), la inteligencia de negocios es uno de los procesos gerenciales que le da a la organización la capacidad de tener un aprendizaje continuo. Para lograr una adecuada toma de decisiones, las organizaciones utilizan visualizaciones analíticas haciendo uso de herramientas de inteligencia de negocio, como la minería de datos (Data Mining, DM). El DM es un mecanismo de explotación de datos, que consiste en la búsqueda de información valiosa en grandes volúmenes de datos que proporciona la información histórica, con la cual, aplicando algunos algoritmos de DM se puede obtener la información necesaria para la toma de decisiones (Andrea et al., 2016).

2.3 Aplicación de la minería de datos en la educación

La Inteligencia de Negocios (BI, por sus siglas en inglés) es un término acuñado por Hans Peter Luhn en 1958 y, en 1989 Howard Dresner lo amplió incluyendo en su definición el uso de sistemas de apoyo basado en datos reales. BI es cada vez más importante en la gestión estratégica y en el apoyo de estrategias comerciales, la cual permite a los administradores y usuarios finales convertir grandes cantidades de datos no transparentes en información útil que proporciona a las empresas grandes capacidades. Por tanto, es una herramienta analítica que transforma los datos internos y externos en un conocimiento apropiado que respalda el proceso de toma de decisiones, y al combinar datos operacionales con las herramientas analíticas proporcionan a los planificadores y gerentes corporativos información competitiva (Alnoukari, M. & Hanano, 2017). Por esta razón, los investigadores consideran la inteligencia de negocios como un diferenciador competitivo (Brinkmann, 2015).

Con la llegada del Internet y la Web 2.0, el volumen de datos ha crecido exponencialmente, de diferentes fuentes y de diferentes formas: estructuradas y no estructuradas, contenidas en bases de datos, documentos de texto, recursos multimedia, entre otros (Abdelrahman & Farah, 2018). Las empresas, al no poder sustentar sus decisiones operativas, tácticas y estratégicas por simple intuición, deben aprovechar la información que se está generando, dentro y fuera y ante los nuevos retos propuestos por las naciones en el ámbito de la sustentabilidad, la eficiencia energética, la reducción del daño al medio ambiente, entre otros, implica la implementación de nuevas estrategias y técnicas para manejar eficientemente los recursos y por ende tomar mejores decisiones (Ge et al., 2017).

El área ha ganado recientemente mucha atención de todos los campos de aplicaciones como industria, economía, medicina, CRM, comercio, debido a la existencia de grandes colecciones de datos en diferentes formatos, y la creciente necesidad de análisis y comprensión de ellos (Khademolqorani y Hamadani, 2013); las organizaciones educativas también incorporan el uso de inteligencia de negocios, para tomar decisiones efectivas en el ámbito de la administración, de las condiciones en las que se encuentran y hacia dónde quieren llegar (Reyes Dixson y Nuñez Maturel, 2015); también hay una necesidad de implementar el análisis predictivo para diseñar estrategias educativas (Boticario et al., 2015).

2.4 Técnicas de Minería de Datos.

El uso de minería de datos para facilitar el soporte de decisiones permite nuevos enfoques para la resolución de problemas, mediante el descubrimiento de patrones y relaciones ocultos en los datos y; por lo tanto, permite un enfoque inductivo para los sistemas de soporte de decisiones (Decision Support Systems, DSS).

Los criterios que se mencionaron en el reporte del panel de la conferencia KDD-2006, para identificar si un problema es susceptible de ser resuelto a través de minería de datos fueron los siguientes: si el problema no se ha resuelto, si la minería es un elemento fundamental para su solución, si se cuenta con volúmenes representativos de los datos, si tiene un propósito específico, si interesa a los investigadores y si tiene un beneficio público significativo el resolverlo, esto clarifica cuando se justifica el uso de las herramientas de minería de datos en la búsqueda de una solución (Piatetsky-Shapiro et al., 2006).

Las tareas claves de la minería de datos son la clasificación, la regresión y el agrupamiento, su esencia es generar de forma inductiva a partir de los datos (se le conoce como entrenamiento) un modelo (se le conoce como conocimiento) que puede aplicarse a nuevos datos (predictivo) (Gironés et al., 2017). Los algoritmos de minería de datos se dividen en supervisados y no supervisados, los primeros son aquellos que requieren un grupo de datos previamente clasificados y en donde conocemos el valor de los atributos de antemano y los segundos son aquellos donde los datos no tienen ninguna etiqueta o clasificación, no se conoce ningún valor objetivo categórico o numérico *a priori* (Gironés et al., 2017).

Existen numerosos algoritmos de DM, incluidos los algoritmos predictivos que dan como resultados clasificadores que se pueden usar para predicción y clasificación, y algoritmos descriptivos que sirven para otros fines como encontrar asociaciones, clústeres, entre otros.

2.4.1 Aprendizaje supervisado

Los algoritmos de aprendizaje más utilizados son los predictores lineales ya que son muy fáciles de interpretar e intuitivos, como la regresión lineal y la logística, con estos se puede estimar el valor de una variable usando sus atributos distintivos; también están los de clasificación, que permiten organizar, ordenar y/o clasificar los datos de acuerdo a sus atributos y comportamientos, y poder predecir cómo se conforman esas categorías o se agrupan de acuerdo a ciertas decisiones, como el algoritmo de árboles de decisión, redes neuronales, redes bayesianas entre otros (Shalev-

Shwartz y Ben-David, 2013), los algoritmos de aprendizaje supervisado utilizan variables de entrada preestablecidas conocidas como variables independientes, las cuales pueden tener un valor cuantitativo, cualitativo o categórico, con la intención de predecir los valores de las salidas o variables dependientes, cuya naturaleza puede ser cuantitativa o categórica, el propósito fundamental es identificar un patrón para predecir la respuesta esperada (Hastie et al., 2013).

Bāliņa et.al (2016) indican que la tecnología de minería de datos trabaja con una gran cantidad de métodos y algoritmos. Estos autores los clasifican en:

- Algoritmo de clasificación, método para extraer datos y ordenar las entidades de acuerdo a sus atributos y comportamientos.
- Algoritmo de regresión, utilizado para la estimación del valor variable usando atributos distintivos.
- Gutiérrez & Molina (2016) incluye como algoritmos de clasificación los siguientes métodos;
- Algoritmo de AdaBoost, para anticipar un comportamiento de acuerdo a un grupo de datos.
- Naive Bayes, usado para predecir la salida de una entidad basado en observaciones conocidas.
- Algoritmo Classification and Regression Tree (CART), utilizado para predecir la probabilidad de un evento utilizando la regresión y la clasificación.
- Máquinas de vectores de soporte: Con datos de entrada, buscan predecir cuál de las categorías los contienen
- Árboles de decisión, utilizados para organizar los datos en decisiones que compiten organizadas en ramas de influencia.

2.4.2 Aprendizaje no supervisado

Los algoritmos de aprendizaje no supervisado buscan descubrir asociaciones relevantes en las observaciones del fenómeno. Por ello, estos algoritmos obedecen a descubrir subgrupos de variables u observaciones para descubrir patrones o respuestas no esperadas (Casella et al., 2013). Estos algoritmos se pueden clasificar en: clusterización, reglas de asociación y reducción de la dimensionalidad (Velayutham, 2020).

La clave en el análisis de los datos con estos modelos no supervisados es la elección de los componentes o atributos principales de las observaciones elegidas para estudiar el fenómeno. Mientras menos componentes principales se analicen, es mejor la capacidad de explicarlo; por ello es importante determinar cuáles serán los componentes que se introducirán al modelo no supervisado (Casella et al., 2013). Para lograr lo anterior se visualizan todas las observaciones (n) con una serie de atributos (p), como parte de un análisis de datos explicativo, pero cuando (p) es muy grande la información se fracciona demasiado que llega a no ser informativa. Para evitar esto, se utiliza la herramienta del principal componente de análisis (principal component analysis, PCA por sus siglas en inglés), la cual busca un pequeño número de dimensiones que sean lo más interesantes posible, donde el concepto de interesante se mide por la cantidad que las observaciones varían a lo largo de cada dimensión (Casella et al., 2013), también se le conoce como reducción de la dimensionalidad (Rosasco, 2017). Otra forma de seleccionar las variables es el enfoque de fuerza bruta, que consiste en considerar todos los subconjuntos de variables, conformando pares de variables, trillizas, y así sucesivamente, de cada conjunto de entrenamiento se seleccionarán las que tengan mejor rendimiento (Rosasco, 2017).

El rendimiento de un método de aprendizaje está relacionado con algunos valores como la capacidad de predicción del modelo, el ajuste del mismo, es decir, si está subajustado es que no captura la tendencia de los datos o sobre ajustado que muestra un sesgo, pero alta varianza, o si el modelo clasifica correctamente los datos de prueba elegido (Velayutham, 2020).

Gutiérrez & Molina (2016) define los siguientes algoritmos como no supervisados:

- Algoritmo K-means, basado en análisis de grupos (clústeres).
- Algoritmo Apriori, para controlar datos de tipo transacciones.
- Algoritmo Esperanza Maximización (EM), usado para predecir una salida futura o evento aleatorio analizando los datos y definiendo un parámetro.
- Algoritmo de PageRank, utilizado como base para los motores de búsqueda, señalando y estimando la relevancia dentro de un gran grupo de datos.
- Algoritmo k-Nearest Neighbor: Identifica la ubicación de los datos de acuerdo a patrones, relacionados con una entidad mayor.

Además Bāliņa, Žuka, & Krasts (2016)menciona los siguientes algoritmos de aprendizaje no supervisado:

- predecir la secuencia de eventos en el tiempo.

- Algoritmo de segmentación, el cual identifica los diferentes tipos de clústeres en función de algunos atributos específicos.
- Algoritmo de detección de anomalías, como uno de los métodos de segmentación.
- Algoritmo de asociación, utilizado para identificar patrones entre las diferentes entidades.

2.5 Metodología de la minería de datos

Como se puede observar en la figura 2-1 el esquema propuesto por Ge et al., (2017) muestra la visión general de la metodología de minería de datos, la cual implica varios procedimientos como la preparación de datos desde bases de datos históricas y en tiempo real, pre procesamiento de datos, selección de modelos de entrenamiento, validación, rendimiento y evaluación; minería de datos y análisis basados en la implementación de algoritmos, análisis de patrones, análisis predictivos, análisis de tendencias, visualización de datos; entre otros, con la finalidad de descubrir conocimiento, mejorar los procesos, la toma de decisiones y el conocimiento automático.

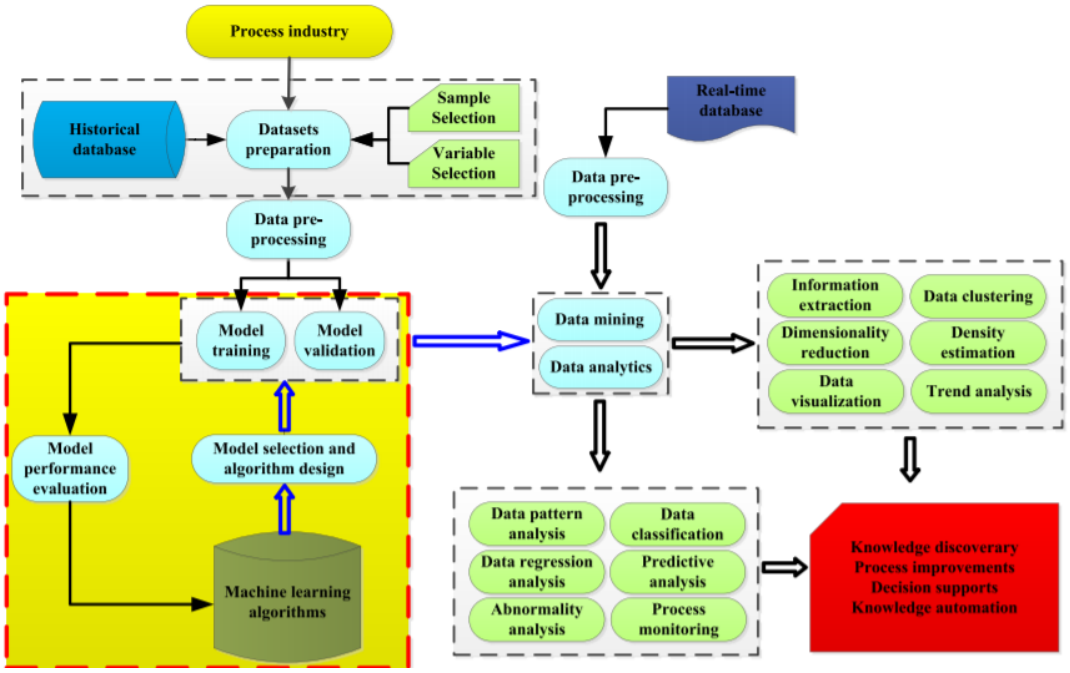


Figura 2-1 Visión general de la metodología de minería de datos y análisis (tomada de Ge et al., 2017)

El proceso de minería de datos y análisis requiere el uso de una serie de herramientas durante cada una de las etapas de desarrollo. El proceso de extracción, transformación y carga de los datos

es el primer paso, en esta etapa del proceso se extraen los datos de las fuentes de origen que pueden ser internas o externas, después se transforman según las necesidades específicas de integración, para que en una etapa siguiente se pueda realizar un análisis de los mismos. Las fuentes de datos pueden ser diversas y pueden contener datos estructurados, no estructurados o semi-estructurados (Curto Diaz, 2016).

Algunas de estas herramientas son:

- Tableau prep: es una herramienta visual que permite combinar, dar y limpiar los datos a través de la construcción de flujos, los datos pueden ser editados y modificados directamente, implementa la agrupación de datos similares, puede utilizar como fuente de datos hojas Excel, bases de datos relacionales, entre otras. Los flujos pueden programarse para reutilizarse eventualmente (*Tableau Prep Builder y Tableau Prep Conductor: Una Solución de Preparación de Datos de Autoservicio*, n.d.).
- SAS *ETL Enterprise* (SAS), combina datos de diferentes fuentes, se puede construir un *datawarehouse* (almacén de datos), el proceso es extraer los datos de diversos orígenes, se transforman en datos que puedan ser analizados y almacenados (*What Is ETL? | SAS*, n.d.).
- *Data integration solutions from IBM*, sirve para extraer grandes volúmenes de datos de diversas fuentes, se transforman y se pueden cargar en un almacén de datos (*Data Integration | IBM*, n.d.).

Después del proceso de extracción, transformación y carga de datos, se puede conformar la base de datos en donde se concentrarán las diferentes tablas de datos. Algunas bases de datos que pueden utilizarse son las siguientes.

- *MySql Database*, el cual es un servicio de base de datos utilizado para implementar aplicaciones nativas de la nube, es una herramienta de acceso abierto (*MySQL*, n.d.).
- *Apache Hive*, es un software de almacenamiento de datos permite la escritura y lectura de grandes volúmenes de datos, residiendo en almacenamiento distribuidos mediante SQL (*Apache Hive TM*, n.d.).

Las herramientas de análisis estadístico y minería de datos, son aquellas que permiten trabajar con los datos, aplicando diferentes algoritmos, Alban & Mauricio (2019) mencionan en su estudio documental sobre publicaciones acerca de la deserción escolar aplicando técnicas de minería de

datos de enero del 2006 a diciembre del 2017 identifican las siguientes como las más populares, el orden no representa la relevancia en el estudio:

- WEKA, es un software de aprendizaje automático de código abierto al que se puede acceder a través de una interfaz gráfica de usuario, aplicaciones de terminal estándar o una API de Java (*Weka 3 - Data Mining with Open Source Machine Learning Software in Java*, n.d.).
- *SPSS Statistics*, es una solución de análisis estadístico avanzado, algoritmos de aprendizaje automático, análisis de texto, integración de big data e implementación dentro de otras aplicaciones (*SPSS Software | IBM*, n.d.).
- *SPSS Modeler*, es una solución visual que implementa la analítica predictiva, la gestión e implementación de modelos y *machine learning* (*SPSS Modeler - Overview | IBM*, n.d.).
- Matlab, software que facilita las tareas de la *data science*, cuenta con herramientas para acceder y pre-procesar datos, crear modelos predictivos y de *machine learning*, y desplegar modelos en sistemas de IT empresariales (*SPSS Modeler - Overview | IBM*, n.d.).
- *Rapid Miner*, es una plataforma de ciencia de datos que permite limpiar y transformar datos, selección y validación de algoritmos de minería de datos con los que se pueden construir modelos operacionales visuales (*Why RapidMiner | RapidMiner*, n.d.).
- R, es un lenguaje y entorno para gráficos y computación estadística., está disponible como software libre bajo los términos de la Licencia Pública General GNU de la Free Software Foundation en forma de código fuente. Se compila y se ejecuta en una amplia variedad de plataformas; tales como: UNIX, FreeBSD, Linux, Windows y MacOS (*R: What Is R?*, n.d.).
    - Excel, aplicación de Microsoft utilizada para el manejo de hojas de cálculo, permite realizar tareas contables y financieras, así como el análisis de datos, cuenta con diferentes funciones para el manejo de la información y la creación de gráficos

2.6 La deserción

La deserción universitaria, es un tema que se ha abordado desde diferentes aspectos: el historial académico, el desempeño académico de los estudiantes, la situación económica de las familias, las estructuras y políticas de las Instituciones, las oportunidades de desarrollarse de

forma exitosa en el estudio, entre otras (Vera Noriega et al., 2012). La deserción se busca describir desde tiempo atrás, Spady (1970), hace un análisis de diferentes estudios realizados a mediados de los años cincuenta y sesenta en donde se identifican diferentes causas de la deserción, las variables más significativas que caracterizan la deserción en estos estudios, son varias como su desempeño académico pasado, el potencial, la familia, la orientación vocacional, las bajas calificaciones, el poco interés por terminar una carrera y la decisión de estudio tomada por terceros; también se identificaron algunas variables asociadas al género y a la motivación por iniciar una carrera, ya que algunos estudios indican que la mujer en estos tiempos, no era presionada o no se esperaba que terminara un programa universitario, y éstas bajas en su mayoría se daban de forma voluntaria, no por causas académicas, no así los hombres, de quienes se esperaba que si obtuvieran un grado académico universitario.

Tinto (1989), otro referente de los estudios de deserción, indica que los estudiantes pueden salir de una Institución educativa y regresar a esa u otra entidad educativa después, cuando las condiciones de su salida cambien, es por ello, que definir deserción es el primer paso para determinar qué es lo que se desea estudiar.

Existen varios estudios para predecir la deserción, con la finalidad de reducirla y por ende incrementar la eficiencia terminal, estos estudios parten del análisis de diferentes variables, desde aquellas que caracterizan al individuo, como la autoeficacia, siendo esta una característica personal que varía de persona a persona, (Navarro Charris et al., 2017) buscando encontrar una correlación entre estos atributos y la permanencia o deserción del estudiante, realizaron un estudio que se efectuó en una IES de la ciudad de Barranquilla en Colombia, se tomó una muestra de 322 estudiantes del período 2015-1, de los cuales 127 fueron desertores y 195 fueron estudiantes activos, en cuanto al género 162 mujeres y 160 hombres, representando el 50% y 29% respectivamente; el instrumento aplicado fue el de la Escala de Autoeficacia General (EAG), demostrando que no existe una correlación significativa entre eficacia y permanencia, así como tampoco existe una correlación significativa entre eficacia y deserción, sin embargo, arrojó como resultados que existen más desertores laborando que los que no laboran, ya que hay más estudiantes activos que dependen de sus padres.

Ramírez y Grandón (2018), realizaron un análisis basado en los datos de 5.288 estudiantes de una IES estatal chilena, correspondiente a cuatro cohortes consecutivas de estudiantes de 44 programas universitarios de las áreas de humanidades, artes, educación, ingeniería, y salud,

utilizando árboles de decisión con parámetros optimizados logrando una precisión del 87.27%, los atributos analizados en el estudio son algunas variables demográficas del estudiante (edad y género), antecedentes de su ingreso a la IES (puntaje del examen de admisión y promedio final de sus estudios de enseñanza media), atributos económicos (nivel de ingreso familiar y tipo de escuela de enseñanza media), y datos de su desempeño académico (avance, promedio académico y desviación estándar calificaciones). Se usó el software *RapidMiner* Studio 7.5, dando como resultado las variables determinadas por el análisis, el promedio de las calificaciones, los años de avance en la carrera y el puntaje en la prueba de selección a la IES.

En un estudio comparativo realizado en México, en el Instituto Tecnológico Superior de Misantla-México, y Comunicaciones, utilizando *SQL Server Data Tools de Microsoft Visual Studio 2012*, se aplicaron cuatro algoritmos: regresión logística, clusterización, arboles de decisión y redes neuronales, a los datos obtenidos del sistema de información escolar de la Institución. Se logró obtener los factores que más influyen en la deserción escolar, que son principalmente el lugar de procedencia y algunas asignaturas en específico y el algoritmo con mejores resultados fue el de regresión logística con un 100% en precisión, sensibilidad, exactitud y la medida-F (la media armónica de precisión y sensibilidad) (Hernández et al., 2016).

En cuanto a la caracterización de la deserción, Eckert y Suénaga(2015), analizaron la información académica de los estudiantes que desertaron, de la carrera de Ingeniería en Informática de la Universidad Gastón Dachary en Argentina, con el propósito de identificar los factores que influyeron en este fenómeno. El período seleccionado para el estudio corresponde a los estudiantes ingresados desde el año 2000 al 2009, totalizando 855 casos analizados. Se aplicaron los algoritmos de clasificación Algoritmo Clasificador C4.5 - Árbol de Decisión, Algoritmo Clasificador Naïve Bayes Aumentado a Árbol (TAN) y Algoritmo Clasificador Reglas OneR; se detectaron atributos que muestran una fuerte relación con la deserción siendo el más importante la cantidad de materias reprobadas en el primer año de estudios; otros son la cantidad de materias cursadas, cuando el estudiante no regulariza en caso de reprobación, la edad de ingreso, la procedencia; estas variables explicativas mostraban una cantidad de verdaderos positivos, de entre un 76% y un 80% de los casos.

Se realizó una revisión de los modelos teóricos de la retención y la deserción más citados por varios autores al momento de analizar los referentes teóricos de éste fenómeno (Donoso & Schiefelbein, 2007),(Berger et al., 2012), (Manyanga et al., 2017), (Morrison y Silverman, 2012),

(Aljohani, 2016), con la finalidad de identificar los atributos que estos modelos destacan como explicativos de la deserción. Estos modelos de la deserción y la retención surgen a partir de los años 60´s, ya que la IES pasa de ser un nivel académico de poco acceso, a un nivel que ya resultaba aspiracional y que contaba con un buen número de alumnos inscritos, para identificar los atributos a los que hacen referencia estos modelos como relevantes en la caracterización del desertor.

2.6.1 Modelo teórico del comportamiento de la deserción.

Uno de los modelos asociado a los aspectos psicológicos, es el modelo de la persistencia del estudiante desarrollado por Tinto (1975), el cual contempla no solo las habilidades o estatus social, también enfatiza en incluir características del individuo y de la institución; al hablar de las características del estudiante. Adopta elementos del modelo teórico de Durkheim del suicidio, el cual indica que cuando la persona no se siente parte de la sociedad o de un colectivo, incrementa la probabilidad de suicidarse, en esa analogía. Tinto refiere que el abandono escolar del estudiante es porque no puede integrarse al sistema social de la IES, sin olvidar el ámbito académico, ya que pudiera sentirse parte del colectivo social, pero tener un bajo desempeño, y viceversa.

Este autor menciona algunas variables que deben considerarse al momento de entrar a la IES tanto las generales como raza, género, capacidad, estatus social, experiencias académicas previas, residencia, pero enfatiza en aquellas que hablan de motivación como expectativas de la carrera elegida y motivación para lograr el objetivo académico, ya que son indicadores del compromiso del estudiante con este objetivo y con la Institución, lo cual puede ser determinante en la decisión de abandono del estudiante.

Sin embargo, también contempla factores externos que pudieran influir en la deserción, como los financieros y la percepción de los beneficios, es decir lo que invierte el estudiante contra lo que recibe. Las variables que representan los antecedentes familiares según Tinto, son el estatus socioeconómico, nivel de estudios de los padres, el ingreso familiar, la relación familiar, las expectativas de los padres con respecto a los logros académicos del hijo. En cuanto a los atributos individuales son aquellos que caracterizan al individuo como la raza, el género, las experiencias académicas previas y la residencia (urbana o rural). El compromiso y las metas influyen en la resiliencia y la persistencia, es por ello que este modelo teórico del comportamiento de la deserción, indica que las intenciones y el nivel de compromiso, son un factor dinámico que puede cambiar en el tiempo. La interacción con la Institución se puede definir como las diferentes

interacciones del estudiante con la Institución: sus pares, los maestros, la administración, entre otros, así como su desempeño académico. Las características propias de la Institución, si es privada o pública, el nivel de calidad, el tamaño de la población, sus métodos de selectividad, son variables que pueden influir en la deserción, sea esta voluntaria o por el desempeño académico del estudiante.

2.6.2 Modelo de la persistencia del estudiante.

Este modelo está caracterizado fundamentalmente por los rasgos de la personalidad del estudiante, ya que asume que la deserción es un debilitamiento de las intenciones iniciales y la retención como un fortalecimiento de éstas; también menciona las perspectivas teóricas de la deserción desde ciertos modelos psicológicos porque enfatiza sobre el apoyo familiar y el nivel aspiracional del estudiante, ya que genera expectativas de éxito y valores que lo harán persistente y por ende evitará la deserción, (Ethington, 1990), se inspira mucho en el Modelo de Tinto, al incluir una serie de variables exógenas como los antecedentes familiares y el rendimiento académico previo.

Incorpora también algunas variables endógenas como la autoconcepción del estudiante con respecto al éxito académico y el estímulo familiar. Sustenta muchos elementos del Modelo de (Eccles & Wigfield, 2002) la teoría del valor de la expectativa, que aborda variables de autopercepción del niño y adolescente sobre sus expectativas del éxito.

Algunos modelos que empiezan a incorporar otras variables externas a las propias del individuo, es decir, variables sociológicas son:

2.6.3 Modelo sociológico explicativo del proceso de deserción.

El modelo de Spady (1970) , define una relación significativa entre el ambiente familiar y el potencial académico y la congruencia normativa, como el apoyo de pares y la integración social, para que el estudiante decida desertar o no. Asumiendo que son diversas las fuentes a las están expuestos los estudiantes dentro de la IES, por ejemplo: cursos, profesores, administradores, procesos académicos y administrativos, entre otras, la relación establecida en el modelo, entre el compromiso institucional y la congruencia normativa, adquiere relevancia porque esta es cíclica, en cada etapa de la vida universitaria, y puede influir en la actitud, el interés y las motivaciones del estudiante. Spady contrario al modelo de Tinto, considera que las calificaciones no son un

atributo significativo, porque estas pudieran adquirirse sin reflejar realmente la capacidad del alumno o si este puede irse o no de la Institución.

En cuanto a los modelos que están orientados hacia un enfoque económico, son aquellos que obedecen al costo-beneficio, ya que, si el estudiante percibe que el beneficio no está respaldado por el costo, o si el estudiante percibe su incapacidad para solventar ese costo.

2.6.4 Modelo teórico de la retención del estudiante

Cabrera et al. (1993), integraron algunas variables de los modelos de Tinto (1975) y, Bean y Metzner (1985) en un estudio longitudinal de la generación de otoño de 1988, durante el primer año de estudios en una institución académica en Australia, a un total de 2,459 estudiantes, creando un nuevo modelo.

También, incluyeron otras variables económicas, enfatizando sobre la necesidad de los estudiantes, de recibir apoyo financiero, la relación costo-beneficio, ya que estas variables ejercen una influencia directa sobre la decisión de permanencia o deserción.

En este modelo, se analizaron diferentes variables ambientales como el apoyo de los amigos y la familia, los apoyos financieros que recibe el estudiante como: becas, préstamos, apoyo económico familiar, empleo, entre otros, siendo estos los más significativos del modelo. También se observan algunas variables endógenas, que hablan del grado de integración social y académica, así como el compromiso institucional con los objetivos del estudiante.

Algunos autores asocian la permanencia con las características de la institución educativa, teniendo una gran relevancia la calidad desde todos los ángulos como: cursos, profesores, experiencias del estudiante en el aula; también consideran los beneficios, como: deporte, cultura, apoyo académico, recursos, infraestructura, entre otros.

Estos modelos son considerados como organizacionales, como el caso del modelo del comportamiento de la deserción de Tinto (Figura 3), y otros modelos como:

2.6.5 Modelo causal

(Pascarella y Terenzini, 1980), sostiene que el desarrollo y cambio de los estudiantes están relacionados con cinco conjuntos de variables. Dos conjuntos de variables relacionadas con antecedentes y características personales, otro conjunto relacionadas con el entorno institucional, uno más que considera las variables asociadas a la frecuencia y contenido de las interacciones

con los miembros de la institución y sus pares; y un quinto grupo de variables relacionadas con el esfuerzo del propio estudiante.

Pascarella se basa mucho en el Modelo de Tinto (1975), pero enfatiza en las variables académicas previas al ingreso, el programa académico elegido y a las características de la IES, como variables que guardan una correlación relevante. También hace una distinción entre el desertor voluntario, siendo éste aquel que se va de la Institución por razones no académicas y el desertor persistente, el cual se va de la Institución por razones académicas.

2.6.6 Modelo del síndrome de abandono estudiantil universitario.

En este modelo Bean (1985), enfatiza en los factores académicos, sociales, psicológicos y ambientales al momento de entrar a la IES, como variables que definen el desempeño académico que el estudiante observará durante su carrera, como es el caso de las calificaciones que obtendrá durante el desarrollo de su aprendizaje ya que éstas incluyen parte de la evaluación del desempeño pasado del estudiante. En lo que respecta a lo social, la adaptación institucional representa, la impresión subjetiva de un estudiante de las normas y valores de sus compañeros, profesores y tutores; y el compromiso institucional muestra cómo se identifica el estudiante con la institución.

Los modelos organizacionales enfatizan las variables de calidad de la docencia y de las experiencias del estudiante en el aula. El modelo de Tinto es el más significativo para explicar la deserción, en el cual se identifica que la influencia sobre la deserción proviene de los atributos previos al ingreso y se define a partir de las experiencias institucionales del estudiante.

Existen otros estudios para predecir la deserción, con la finalidad de reducirla y por ende incrementar la eficiencia terminal, estos estudios parten del análisis de diferentes variables, desde aquellas que caracterizan al individuo, como la autoeficacia, siendo esta una característica personal que varía de persona a persona, o el involucramiento como la teoría de Astin (1999), el cual asegura que la cantidad de energía física y psicológica que el estudiante dedica a su experiencia académica es equivalente al nivel de involucramiento del estudiante.

2.6.7 Teoría de la participación

Astin (1999) plantea que el hecho de que el estudiante esté expuesto a una serie de cursos no significa que se va a generar el deseo de aprendizaje y desarrollo, su teoría de la participación coloca al estudiante al centro de su aprendizaje de forma activa, para ello su teoría contempla

cinco postulados: 1. La participación es la inversión de energía física y psicológica en algún objeto que puede ir desde lo general como la experiencia del estudiante o tan específica como la preparación para un examen. 2. La participación es un proceso continuo y no depende del objeto, se puede dar en diferentes grados. 3.La participación tiene características cualitativas y cuantitativas. 4. La cantidad de aprendizaje del estudiante y el desarrollo personal es directamente proporcional a la calidad y cantidad de participación del estudiante; y 5. La efectividad de cualquier práctica o política educativa ésta relacionada con la capacidad de esa política o práctica de incrementar la participación de los estudiantes.

Indica que la teoría de la participación está más enfocada en los mecanismos o procesos qué facilitan el desarrollo del estudiante, como:

a) El tiempo del estudiante, lo cual implica revisar los horarios de clases, el tiempo que dedica a actividades extracurriculares, los traslados de su residencia a la escuela, entre otros
b) La residencia, cuando los estudiantes residen en la IES les permite participar de diversas formas como la interacción con los profesores, con los grupos de la IES como las fraternidades, clubes, entre otros,
c) Los programas honoríficos, en ellos los estudiantes que participan tienen mayor autoestima interpersonal, autoestima intelectual e intereses artísticos o académicos, aunque podría no favorecer la relación con otros compañeros fuera de estos programas
d) Participación académica, hábitos de estudio, el tiempo que ocupa en el estudio, el grado de interés por los cursos, recompensas institucionales
e) Interacción con los académicos, mientras más interacción se dé, tiene una mejor percepción de la institución
f) Participación deportiva, similar a la participación académica, también los aísla de los otros estudiantes, pero tienen una mejor percepción de la institución
g) Participación en grupos estudiantiles representativos, los estudiantes que participan activamente en el gobierno estudiantil interactúan frecuentemente con sus compañeros, y esta interacción parece favorecer la experiencia universitaria.

Astin, es el autor que integra de forma detallada, cada interacción del estudiante con el ambiente estudiantil y sus diferentes actores de forma tan significativa, más allá del aula y de las calificaciones obtenidas en los diferentes cursos.

2.7 La deserción y la minería de datos

Existen varios estudios que recopilan diferentes trabajos e investigaciones que utilizan la minería de datos dentro del contexto educativo, como un medio que provee a las Instituciones de educación superior de recursos para apoyar la toma de decisiones, no solo para implementar acciones que permitan incidir en la retención o acompañamiento en la vida académica del estudiante, sino para mejorar la calidad educativa de las Instituciones. La minería en este ámbito educativo se reconoce como Minería de datos educacional (EDM, Educational Data Mining, por sus siglas en inglés) (Urbina y De la Calleja, 2017), (Proano y Villamar, 2018), (Diaz Pedroza et al., 2019) y es importante porque los sistemas que mantienen la información de los estudiantes dentro de las instituciones permiten almacenar y recuperar la información que contienen, no necesariamente con la inteligencia necesaria para discriminar la información relevante de la que no lo es para facilitar la toma de decisiones (Devasia et al., 2016).

Como se ha mencionado, la deserción es un problema que enfrentan todas las IES, esta situación tiene un impacto social, económico y político que trasciende no solo a las Instituciones educativas, públicas o privadas, sino también al Sistema Educativo en México, que tiene que crear políticas y/o estrategias que permitan que la sociedad logre un nivel educativo que genere un impacto positivo en la sociedad, a través de capital humano más calificado. Por eso es muy importante poder determinar los factores que inciden en la deserción, ya que las estrategias de retención serán más efectivas (Patiño, L. y Cardona, 2012).

Hay algunos estudios que abordan este problema utilizando la analítica de datos por medio de minería de datos (Eckert y Suénaga, 2015), identificando modelos que permitan predecir la deserción y la graduación de los estudiantes universitarios, algunos utilizando árboles de decisión, clusterización, redes bayesianas, (Miranda & Guzmán, 2017), regresión logística entre otros (Hernández et al., 2016). Estas investigaciones coinciden en muchas de sus conclusiones como lo son algunas variables comunes significativas como el examen de admisión y el promedio con el que ingresan. Sin embargo, muy pocas estudian el proceso y desempeño académico durante su estancia en la IES, algunos estudios incluyen algunas variables cuantitativas como el promedio académico, pero no mencionan índices de reprobación o algunos momentos significativos en los que se da la deserción, como el número de semestre, por ejemplo. Diaz Pedroza et al., (2019) en su investigación documental sobre técnicas, herramientas, algoritmos y atributos para la minería

de datos utilizados en la deserción estudiantil, mencionan que los atributos más utilizados por los autores se clasifican en variables académicas, incluyendo las calificaciones durante su estancia como su promedio al ingresar a la IES así como el examen de admisión, las variables familiares, como el nivel de estudios de los padres y su ocupación y las variables demográficas como la edad, estado civil, numero de hermanos entre otros,

Se encontraron muchas investigaciones descriptivas, de las cuales varias de ellas se apoyan en encuestas de salida o en los resultados obtenidos de cuestionarios aplicados a grupos de interés. Algunas de estas investigaciones fueron realizadas en IES privadas como la Corporación Universitaria Lasallista (Londoño, 2013), en donde se mencionan factores institucionales obtenidos de una serie de encuestas, o como el realizado en la Universidad católica de Temuco (Peralta et al., 2017) donde se aborda deserción y graduación como un análisis comparativo de las variables que explican ambos fenómenos.

Al aplicar herramientas de análisis que permitan describir y predecir el fenómeno de la deserción, facilita identificar las variables que inciden en este. Alban & Mauricio (2019) presentan una revisión de la literatura sobre la predicción de la deserción universitaria a través de técnicas de minería de datos. Para ello se tomaron en cuenta tres etapas: la planeación, en donde se identifica el protocolo de revisión y la necesidad de la investigación; la segunda etapa fue la implementación, donde se aplica la inclusión y exclusión de criterios; y la tercera etapa donde se evalúan los resultados obtenidos y el tipo de análisis estadístico seleccionado.Los resultados encontrados fueron, que la mayoría de los autores se concentran en análisis multifactoriales, clasificados en cinco dimensiones: personal, académico, económico, social e institucional, y el más común es el personal, que considera factores como la edad, la etnicidad y el género. Las técnicas de análisis más utilizadas fueron las estadísticas y las de inteligencia artificial, siendo las estadísticas las más frecuentes, sin embargo, los modelos diseñados con técnicas de inteligencia artificial presentan mejores indicadores de exactitud.

Alban & Mauricio (2019) analizaron 1,681 artículos, de los cuales 67 cumplieron con los criterios de selección: modelos que proveen una solución a la deserción universitaria, que contemplen factores que influyen en la deserción, que incluyan predicciones basadas en minería de datos, que presenten métricas que evalúan la calidad de los modelos predictivos y aquellos que si responden a las preguntas de investigación. El volumen de artículos sobre el tema implica que es un tema de interés para la comunidad científica y que aún no ha sido resuelto.

En cuanto al uso de algoritmos supervisados en el contexto educativo, Urbina y De la Calleja (2017) indican que los algoritmos de aprendizaje de máquina más comunes son las redes neuronales, los árboles de decisión, los métodos basados en instancias y el aprendizaje bayesiano; sin embargo, cada uno de estos muestran ventajas y desventajas. Por ejemplo, los árboles de decisión son modelos que se pueden interpretar fácilmente y son robustos cuando se presentan atributos atípicos o redundantes, su desventaja es que se puede obtener como resultado árboles muy grandes, con muchas reglas de decisión dando como resultado predicciones con bajo desempeño.

Sobre la deserción universitaria se revisaron algunos estudios en donde se aplicaron diferentes algoritmos de clasificación como Multilayer Perceptron, árboles de decisión J48, Random Forest y Random Tree (Rodriguez Maya et al., 2017), mostrando mejor rendimiento el de Random Forest (RF). El algoritmo RF, crea un número grande de árboles de clasificación, construyendo cada uno con un algoritmo determinista, dividiendo cada nodo con un subconjunto de predictores aleatorias y utilizando una muestra de las observaciones. Evalua la exactitud de la precisión con un tercio de estas observaciones, sin poda ni recorte, permitiendo con una mayor precisión y exhaustividad (recall, la capacidad de predecir todos los casos verdaderos, sean estos positivos o negativos), la distribución de las observaciones en el set de datos es relevante debido a la varianza y al sesgo. Con este algoritmo, si existiera una representación de clases desequilibrada, el modelo tenderá a predecir mejor la clase que más observaciones tiene ya que se entrena para ello. Los árboles de decisión son fáciles de interpretar ya que se construyen con una serie de reglas que permite identificar las condiciones del modelo, hay estudios que muestran una importante mejora en la medida de exhaustividad (recall) (Limsathitwong et al., 2018).

Capitulo 3 Metodología de Investigación

3.1 Introducción

La metodología implica la definición de las etapas del estudio, así como los elementos que deben contemplarse en cada una de estas, indicando los pasos más relevantes para la identificación, cuantificación, evaluación y análisis de la información requerida para cumplir con los objetivos específicos y generales de la Investigación, así como las técnicas y herramientas que deberán aplicarse para lograr estos resultados.

Esta investigación es explicativa ya que busca exponer el fenómeno de la deserción por medio de diferentes variables, por lo tanto, es de carácter multivariado no experimental, debido a que se tomará la información que es generada por los desertores, fundamentalmente cuantitativa, aunque se incluirán datos categóricos de algunas variables cualitativas que caracterizan al estudiante. La investigación incluirá un análisis longitudinal de tendencia, ya que se analizará el fenómeno durante el período de agosto de 2008 a agosto de 2018 considerando a cada grupo y programa académico (cohortes) para analizar el comportamiento de cada generación.

La identificación de las fuentes de información inicia con el proceso de admisión del prospecto hasta la salida del estudiante por deserción o por cambio de campus, la extracción y transformación de datos se da a partir de la identificación de los atributos que caracterizan al desertor en dos momentos: cuando entra a la IES y cuando sale, en estos momentos se identifican algunos de los atributos personales, académico e institucionales, mismos que sirven para entender y describir el fenómeno de la deserción para el caso de estudio del campus Ensenada del Sistema CETYS Universidad.

La minería de datos permite descubrir patrones y modelos que pueden explicar y predecir un fenómeno como el de la deserción, el cual es un problema no resuelto y del que se sigue investigando. En este capítulo revisamos algunos algoritmos de la minería de datos que se utilizaron para descubrir y discriminar los atributos más significativos del fenómeno y los modelos que se construyeron para probar la hipótesis planteada. Es relevante que los atributos puedan ser identificados desde el valor que representan, ya que estos valores permiten que en términos de probabilidad o de variabilidad sean los predictores del fenómeno.

Para construir los modelos predictivos se utilizaron algoritmos de aprendizaje supervisados lineales y de clasificación, siendo los de regresión logística y el de árboles de decisión los más adecuados por la composición del set de datos recolectado y preparado para construir los sets de entrenamiento, la distribución de clases (desertor y no desertor) no están balanceadas es por ello que se eligió el algoritmo de bosque de árboles aleatorios. La evaluación del desempeño de estos modelos fue buena, las variables predictoras son algunos atributos que el estudiante tiene al momento de ingresar a la IES asociados al desempeño académico previo y otros que definen su permanencia asociada al tema financiero.

3.2 Diseño de la Investigación

CETYS Universidad es una Institución educativa privada sin fines de lucro en el noroeste del país, que cuenta con 30 programas académicos de nivel bachillerato, licenciatura y posgrado, estos programas se encuentran bajo la administración de tres colegios: Ingeniería, Administración y Negocios, y Sociales y Humanidades. Fue fundada en 1961, por un grupo de empresarios conformando el Consejo del Instituto Educativo del Noroeste A.C. (IENAC), quien auspicia a esta IES (CETYS Universidad, 2017). Cuenta con tres campus distribuidos en tres ciudades: Mexicali, Tijuana y Ensenada, siendo este último el campus internacional de la institución y el más pequeño.

La Institución actualmente atiende a más de 7,800 estudiantes a nivel sistema, sin contar los estudiantes del área de educación continua. Las estadísticas muestran que a la fecha hay más de 37,000 egresados. En el caso de campus Ensenada, se cuenta con 395 estudiantes de profesional y alrededor de 6,000 egresados (CETYS Universidad, 2017). Según cifras oficiales, el porcentaje de graduación en campus Ensenada para la escuela de Administración y Negocios y la escuela de Ingeniería del 2010 al 2015, fluctúa entre el 35% y 53%, la meta planteada por Rectoría en el plan 2020 es del 70% (CETYS, 2011).

Como se puede observar en la Figura 3-1, la metodología se encuentra dividida en las siguientes etapas:

I. Identificación de la población foco

Se revisará toda la información documental sobre normativa, conceptos y políticas sobre retención, deserción, eficiencia terminal, los programas académicos y los ejes de competencias definidas en el currículo, las cohortes y sus características; con la finalidad de

identificar a los desertores y sus atributos. Esa información servirá para el diseño del modelo lógico de la base de datos.

II. Identificación de las fuentes de información**.**

Se buscarán todas las fuentes de datos en donde se va concentrando la información del desertor desde que inicia su contacto con la Institución, durante su estancia en ella y su salida. Estas fuentes de datos pueden ser resultado de un proceso automatizado o manual. Todos los procesos que existen en la operación de la Institución generan información estructurada o no estructurada.

III. Recolección de la información.

En esta etapa se seleccionará y recolectará la información de las diferentes fuentes de datos, esta información debe corresponder a los desertores de Licenciatura desde agosto de 2010 a agosto de 2018; con la intención de extraerla, depurarla, transformarla e insertarla a la base de datos de desertores, la cual tendrá el concentrado de datos para su análisis. Se busca que la información contemple información de la cohorte, las condiciones de ingreso, su estancia y su salida de la Institución,

IV. Aplicación de algoritmos supervisados y herramientas de minería de datos

De acuerdo a la revisión de la literatura sobre aprendizaje de máquina y herramientas de minería de datos, se utilizarán los algoritmos supervisados más adecuados para la información recolectada. Se buscará que el modelo pueda responder a los diferentes enfoques de análisis: por sus atributos: género, programa académico, etc., por cohorte, por su etapa en la Institución, entre otros.

V. Prueba experimental de los modelos

Se harán diferentes pruebas aplicando los modelos desde los diferentes enfoques de análisis mencionados y se evaluará su efectividad.

VI. Conclusiones

Se documentarán las predicciones, su evaluación y las conclusiones de la investigación.

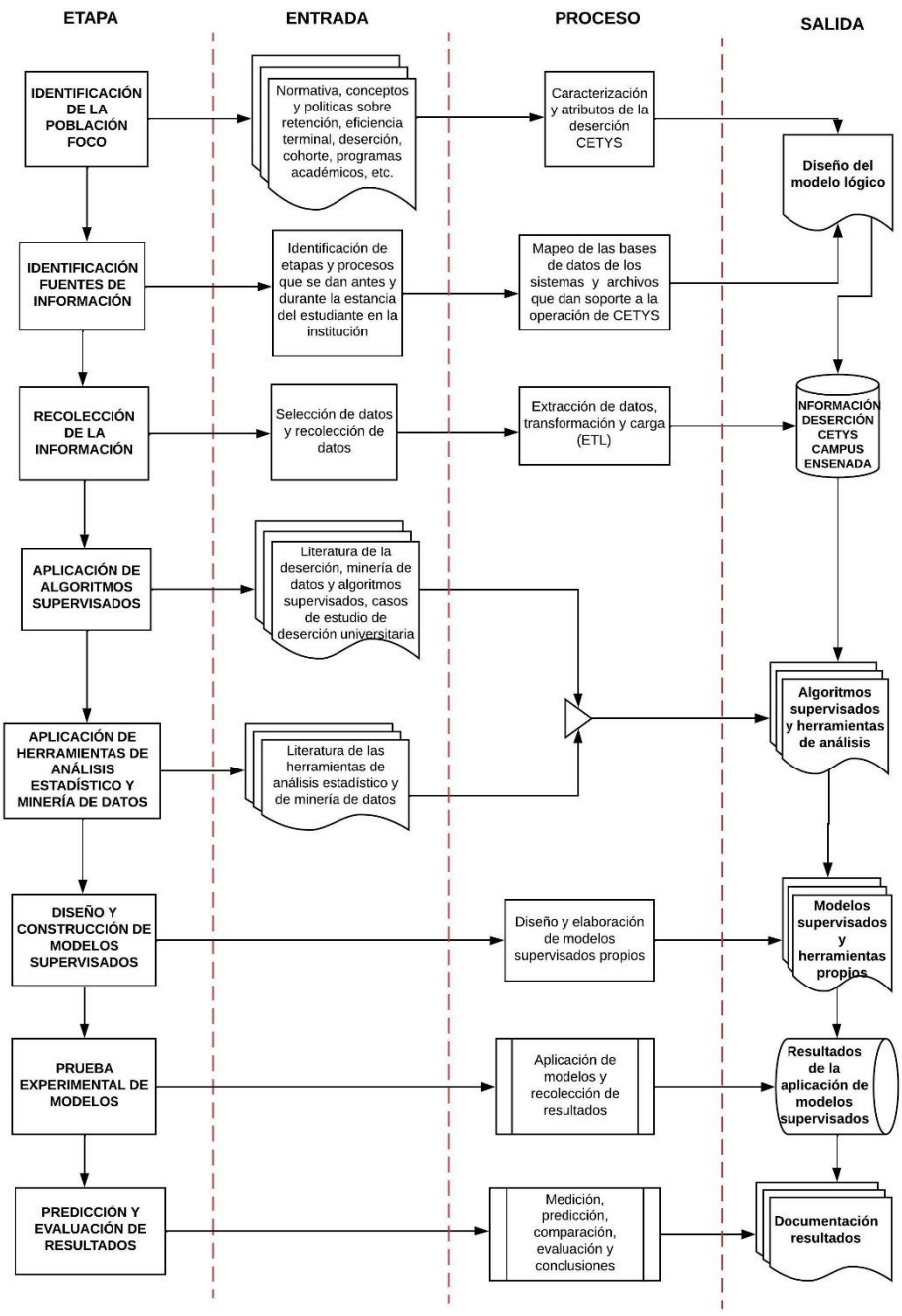


Figura 3-1 Metodología, Fuente: Elaboración propia

Descripción de los datos

3.3.1 Atributos

Los atributos son aquellas variables que caracterizan al estudiante y al desertor, pueden ser cuantitativos o cualitativos; y clasificarse como psicológicos, sociológicos o económicos;

además, algunos ya están definidos desde antes de que el estudiante ingresa a la IES y otros cuando se inserta al colectivo universitario; además, estos pudieran cambiar durante su permanencia en la institución debido a las experiencias vividas.

En la Tabla 3-1, se presentan las variables significativas que explican la deserción y la retención, de acuerdo a los modelos teóricos revisados. Se pueden distinguir tres contextos relevantes: el personal, el académico y el institucional, los cuales pueden presentarse en diferentes momentos: al ingresar a la IES, ya que son atributos que tienen un valor previo al ingreso a la IES, o durante su estancia en ésta, como una variable que se genera y toma valor cuando el estudiante ya forma parte del colectivo universitario o como un valor dinámico que puede cambiar en el tiempo, sobre todo porque la expectativa se convierte en una experiencia que puede incidir de forma favorable o no, en la permanencia del estudiante.

Tabla 3-1 Variables significativas de la deserción según los autores revisados. Fuente: elaboración propia

Variable o atributo	**Modelo teórico**	**Clasificación**	**Momento en que se puede medir**
Raza	(Tinto, 1975), (Ethington, 1990) (Pascarella et al., 1986)	Contexto Personal	Al ingresar
Género	(Tinto, 1975) (Ethington, 1990) (Pascarella et al., 1986)	Contexto Personal	Al ingresar
Experiencias académicas previas	(Tinto, 1975) (Ethington, 1990) (Pascarella et al., 1986) (Bean, 1985)	Contexto Académico	Al ingresar
Logro académico previos	(Ethington, 1990) (Pascarella et al., 1986) (Pascarella & Terenzini, 1980) (Bean, 1985)	Contexto Académico	Al ingresar
Actividades extracurriculares previas	(Pascarella & Terenzini, 1980) (Bean, 1985)	Contexto académico	Al ingresar
Residencia (urbana o no)	(Tinto, 1975)	Contexto Personal	Al ingresar

	(Pascarella et al., 1986)		
Orden de importancia para asistir a esa Institución	(Pascarella & Terenzini, 1980) (Cabrera et al., 1993)	Contexto Personal	Al ingresar
Nivel de confianza de haber elegido la mejor Institución	(Pascarella & Terenzini, 1980) (Cabrera et al., 1993)	Contexto Personal	Al ingresar Durante su estancia
Resultados del examen de admisión	(Pascarella et al., 1986)	Contexto académico	Al ingresar
Nivel de Motivación y compromiso con el propósito académico y la Institución	(Tinto, 1975) (Ethington, 1990) (Pascarella et al., 1986) (Cabrera et al., 1993)	Contexto Personal	Al ingresar Durante su estancia
Integración social y académica	(Tinto, 1975) (Spady, 1970) (Pascarella et al., 1986) (Cabrera et al., 1993) (Bean, 1985) (Astin, 1999)	Contexto Personal	Durante su estancia
Estatus socioeconómico	(Tinto, 1975) (Ethington, 1990) (Spady, 1970) (Pascarella et al., 1986)	Contexto Personal	Al ingresar
Nivel de estudios de los padres	(Tinto, 1975) (Ethington, 1990) (Spady, 1970) (Pascarella et al., 1986)	Contexto Personal	Al ingresar
Ingreso familiar	(Tinto, 1975) (Ethington, 1990) (Spady, 1970) (Pascarella et al., 1986)	Contexto Personal	Al ingresar
Relación familiar	(Tinto, 1975) (Ethington, 1990) (Spady, 1970)	Contexto Personal	Al ingresar
Expectativas de los padres con respecto a los	(Tinto, 1975) (Ethington, 1990) (Spady, 1970)	Contexto Personal	Al ingresar

logros académicos del hijo.	(Cabrera et al., 1993)		
Programa académico	(Pascarella & Terenzini, 1980)	Contexto académico	Al ingresar
Evaluación de la institución	(Tinto, 1975) (Spady, 1970) (Pascarella et al., 1986) (Cabrera et al., 1993) (Bean, 1985)	Contexto Institucional	Durante su estancia
Evaluación del profesorado	(Tinto, 1975) (Spady, 1970) (Pascarella et al., 1986) (Cabrera et al., 1993) (Bean, 1985)	Contexto Institucional	Durante su estancia
Evaluación del programa académico	(Tinto, 1975) (Spady, 1970) (Pascarella et al., 1986) (Cabrera et al., 1993) (Bean, 1985)	Contexto Institucional	Durante su estancia
Evaluación del ambiente estudiantil	(Tinto, 1975) (Spady, 1970) (Pascarella et al., 1986) (Cabrera et al., 1993) (Bean, 1985)	Contexto Institucional	Durante su estancia
Calificaciones	(Tinto, 1975) (Spady, 1970) (Pascarella et al., 1986) (Cabrera et al., 1993)	Contexto académico	Durante su estancia
Logros académicos	(Tinto, 1975) (Spady, 1970) (Pascarella et al., 1986) (Cabrera et al., 1993)	Contexto académico	Durante su estancia
Autoconcepción del estudiante sobre el éxito académico	(Ethington, 1990) (Pascarella et al., 1986)	Contexto Personal	Al ingresar Durante su estancia
Percepción de dificultad	(Spady, 1970)	Contexto Personal	Al ingresar

Expectativas políticas, económicas, sociales	(Ethington, 1990) (Spady, 1970) (Cabrera et al., 1993)	Contexto Personal	Al ingresar
Deseo de reconocimiento	(Ethington, 1990) (Spady, 1970)	Contexto Personal	Al ingresar
Becas	(Cabrera et al., 1993)	Contexto Institucional	Al ingresar Durante su estancia
Financiamiento	(Cabrera et al., 1993)	Contexto Institucional	Al ingresar Durante su estancia
Prestamos	(Cabrera et al., 1993)	Contexto Institucional	Al ingresar Durante su estancia
Trabajo como fuente de ingreso durante la estancia en la IES	(Cabrera et al., 1993)	Contexto Institucional	Al ingresar Durante su estancia
Administración del tiempo	(Astin, 1999)	Contexto Institucional	Durante su estancia
Participación en grupos representativos (sociedades, consejos, deportivos, etc)	(Astin, 1999)	Contexto Institucional	Durante su estancia
Lugar donde vive	(Astin, 1999)	Contexto Personal	Durante su estancia
Participación en programas honoríficos	(Astin, 1999)	Contexto Institucional	Durante su estancia
Hábitos de estudio	(Astin, 1999)	Contexto Personal	Durante su estancia
Relación con los profesores, tutores, asesores, entre otros	(Astin, 1999)	Contexto Institucional	Durante su estancia

Para visualizar los contextos en los que se da la deserción y cómo es que los atributos van cambiando su valor, o van surgiendo nuevos a medida que el estudiante inicia y desarrolla su vida académica, se diseñó el modelo entidad-relación que se presenta en la Figura 3-2; en el cual se puede observar que el estudiante se va caracterizando con sus propias expectativas y las de su familia, los atributos de la Institución y su experiencia durante su estancia en la IES.

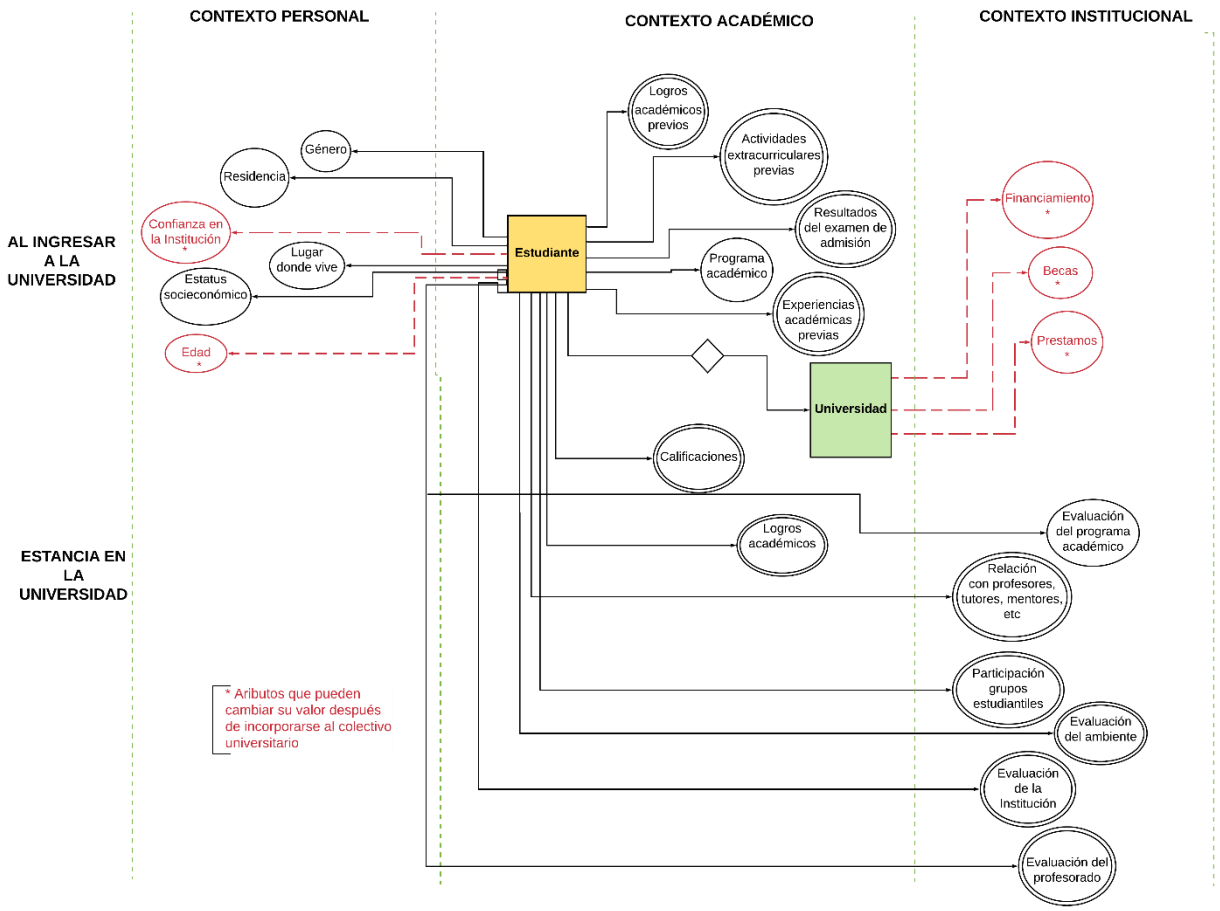


Figura 3-2 Modelo entidad–relación de los atributos del estudiante al ingresar y durante su estancia en la IES (Elaboración propia).

3.3.2 Caracterización de los atributos.

Se utilizaron como referentes, los atributos que se identificaron como significativos según los modelos teóricos analizados y para el análisis de las variables, se seleccionarán aquellas que estén disponibles en las bases de datos de la Institución, y algunos otros que podrían ser relevantes para este estudio, para determinar si son explicativos del fenómeno e incluso podrían conformar el modelo predictivo buscado.

3.3.3 Desertores de Licenciatura de CETYS Universidad campus Ensenada.

La deserción en CETYS Universidad es considerada como el momento donde el alumno se da de baja definitiva sin haber terminado su programa de estudios, si un estudiante se da de baja de un programa académico y se integra a otro, no es considerado un desertor. Se puede apreciar

en la Tabla 3-2, los estados por los que pasa un estudiante durante su estancia en la institución (Vargas, 2015)

Tabla 3-2 Clasificación de los estados en los que se puede ubicar un estudiante para fines de retención, deserción y eficiencia terminal. (Vargas, 2015)

Estado (i): Ei	Descripción	Observación para fines de retención, deserción y eficiencia terminal
E1	El estudiante es de nuevo ingreso (NI) y pertenece a la Generación X y se mantiene siempre en el mismo programa académico. Será observado por un periodo de 6 años a partir de su ingreso para fines de retención, deserción y eficiencia terminal.	Sera parte de la población a observar por un periodo de 6 años a partir del año de ingreso.
E2	El estudiante acompaña a la Generación X e ingresó (con o en el siguiente semestre después de la Generación X) mediante equivalencia de estudios ya sea proviniendo de otra institución de educación superior o de los campus del Sistema CETYS.	No se toma en consideración para fines de retención y eficiencia terminal.
E3	El estudiante causa baja definitiva de la Generación X, ya no forma parte de la población escolarizada de la institución.	Se toma en consideración para fines de retención y eficiencia terminal.
E4	El estudiante causa baja definitiva, pero forma parte de la población estudiantil que acompaña a la Generación X. Ya no forma parte de la población escolarizada de la institución	No se toma en consideración para fines de retención y eficiencia terminal.
E5	El estudiante egresa como miembro de la Generación X.	Se toma en consideración para fines de retención y eficiencia terminal.
E6	El estudiante egresa acompañando a la Generación X.	No se toma en consideración para fines de retención y eficiencia terminal.
E7	El estudiante pertenece a la Generación X, pero representa una baja temporal o ha interrumpido sus estudios de manera temporal (se incluyen aquí los cambios de semestre).	Se toma en consideración para fines de retención y eficiencia terminal,
E8	El estudiante acompaña a la Generación X, pero representa una baja temporal o ha interrumpido sus estudios de manera temporal (se incluyen aquí los cambios de semestre)	No se toma en consideración para fines de retención y eficiencia terminal.
E9	El estudiante egresa como miembro de la Generación X pero después del periodo de seguimiento de 6 años.	No se toma en consideración para fines de retención y eficiencia terminal.
E10	El estudiante fue incorporado a la Generación X como consecuencia de un cambio de carrera, campus o colegio.	Se toma en consideración para fines de retención y eficiencia terminal.

3.3.4 Cohorte

La generación o cohorte, según la política 1, de la normatividad para medir la retención y la eficiencia terminal en CETYS Universidad es*: "Para el caso de los programas tradicionales de licenciatura y de educación media superior, se identifica al conglomerado de estudiantes inscritos en el segundo periodo semestral de cada año natural, como la población estudiantil que integra a la generación de ese año para fines de seguimiento en materia de retención, deserción y eficiencia terminal"* (Vargas, 2015).

3.3.5 Programas académicos

Los programas académicos de la Institución están sustentados en su modelo educativo basado en competencias que se presenta en la Figura 3-3, es así como cada mapa curricular de los programas académicos de licenciatura de la Institución, está dividido en tres ejes de formación; generales, básicas y profesional.

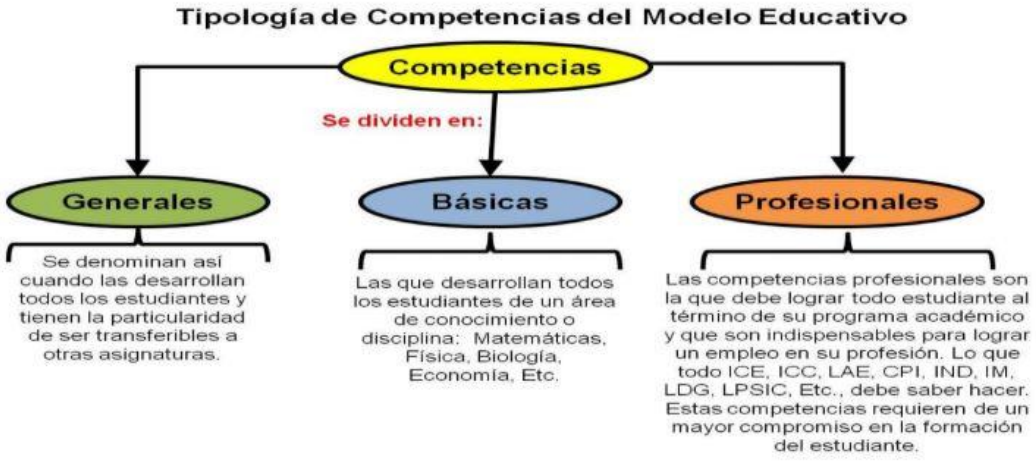


Figura 3-3Tipología para formular las competencias de los programas académicos en el modelo educativo basado en competencias de CETYS Universidad.

Los programas de licenciatura de CETYS Universidad se renuevan cada 6 años a través de un proceso de revisión de programas académicos y planes de estudio, para efectos del tiempo de deserción que se está analizando tienen diferentes características, que son presentadas en la Tabla 3-3.

Tabla 3-3 Características de los programas académicos y planes de estudio. Fuente: Elaboración propia

Programa académico	Planes de estudio	Año de inicio	Número de cursos	Créditos	Duración (semestres)	Colegio
Negocios Internacionales (LNI)	NI2004	2004	42	328	8	Administración y Negocios
Negocios Internacionales (LIN)	NI2007	2007	46	360	8	Administración y Negocios
Negocios Internacionales (LNI)	NI2015	2015	45	353	8	Administración y Negocios
Administración de Mercadotecnia (LAM)	AM2000	2000	52	335	9	Administración y Negocios
Administración de Mercadotecnia (LAM)	L02004	2004	42	328	8	Administración y Negocios
Administración de Mercadotecnia (LAM)	LM2007	2008	46	360	8	Administración y Negocios
Administración de Mercadotecnia (LAM)	LM2015	2015	45	353	8	Administración y Negocios
Administración de Empresas (LAE)	LA2000	2001	54	403	9	Administración y Negocios
Administración de Empresas (LAE)	LA2004	2004	42	328	9	Administración y Negocios
Administración de Empresas (LAE)	LA2007	2007	47	360	8	Administración y Negocios
Administración de Empresas (LAE)	LA2015	2015	44	337	8	Administración y Negocios
Diseño Gráfico (LDG)	LDG05	2002	42	326	8	Administración y Negocios
Diseño Gráfico (LDG)	LDG07	2009	46	360	8	Administración y Negocios
Diseño Gráfico (LDG)	DG2015	2016	45	353	8	Administración y Negocios
Ingeniería Industrial (IIND)	II2004	2005	42	328	8	Ingeniería
Ingeniería Industrial (IIND)	II2007	2011	46	360	8	Ingeniería
Ingeniería Industrial (IIND)	II2015	2015	45	360	8	Ingeniería
Ingeniería de Software (IS)	IS07	2005	42	328	8	Ingeniería
Ingeniería de Software (IS)	IS2007	2008	46	360	8	Ingeniería
Ingeniería de Software (IS)	IS2015	2015	45	360	8	Ingeniería
Ingeniería Mecánica (IMEC)	IMEC05	2014	42	328	8	Ingeniería
Ingeniería en Diseño gráfico (IDGD)	IDGD05	2009	42	328	8	Ingeniería
Ingeniería en Diseño gráfico (IDGD)	ID2015	2015	45	360	8	Ingeniería

Ingeniería en Energías Renovables (IER)	IER2015	2015	45	360	8	Ingeniería
Ingeniería Cibernética (ICE)	IE2004	2006	42	328	8	Ingeniería
Ingeniería Cibernética (ICE)	IE2007	2007	46	360	8	Ingeniería
Ingeniería Cibernética (ICE)	CE2015	2015	45	360	8	Ingeniería

3.3.6 Ingreso a la Institución

La promoción de los programas que ofrece CETYS Universidad está a cargo del departamento de Promoción de la Institución, mismo que cuenta con una coordinación en cada campus, y se lleva a cabo a través de los siguientes mecanismos:

a) El directorio de escuelas lo registra y actualiza el área de promoción en cada campus, en una plataforma CRM (Customer Relationship Management, por sus siglas en inglés) con la finalidad de contar con información del universo de prospectos.
b) Se cuenta con el Programa de Inscripción Personalizada (PIP) que pertenece al área de Promoción y se encarga de la atención-inscripción personalizada de alumnos), en el cual se da seguimiento a información a aquellos candidatos interesados a ingresar a la Institución. Igualmente, la Dirección de Promoción Académica en conjunto con las Escuelas, realiza diversos eventos.

Para la admisión en CETYS Universidad se deben cumplir dos criterios generales: cumplir con el nivel académico anterior, y la obtención de la puntuación mínima en el examen de admisión establecida por la vicerrectoría académica. Para todos los programas de licenciatura, sin importar su modalidad y forma de entrega (presencial, híbrida, o en línea) se usará la prueba estandarizada Prueba de Aptitud Académica, PAA, del College Board (o la prueba estandarizada Scholastic Aptitude Test, SAT, del College Board para el caso de los aspirantes extranjeros provenientes de los Estados Unidos).

De igual manera, se considera el rendimiento académico basado en el promedio final de preparatoria y su calificación en el examen de admisión para el Programa de Apoyo Financiero de estímulo a la calidad académica (beca excelencia/beca talento), dicho programa de apoyo solo aplica a los programas de licenciatura. Los criterios para identificar el estatus de entrada del estudiante que ingresa a la institución son determinados por el puntaje obtenido en el examen, información presentada en la Tabla 3-4.

Tabla 3-4 Criterios de admisión según resultados del examen de admisión (CETYS Universidad, 2016).

Puntuación en PAA	Puntuación en SAT	Tipo de admisión	Condiciones
Igual o mayor a 1,100 puntos	Igual o mayor a 1,100 puntos	Automática	Ninguna (*)
De 1,000 a 1,099 puntos.	De 900 a 999 puntos.	Condicional 1	Registrarse en el programa de apoyo académico para estudiantes de licenciatura (*).
De 950 a 999 puntos.	De 850 a 899 puntos.	Condicional 2	Registrarse en el programa especial de apoyo académico para alumnos de nuevo ingreso a Profesional. De igual forma la carga académica queda sujeta a revisión.
Igual o menor a 949 puntos	Igual o menor a 849 puntos	No admisión	
(*) Todos los estudiantes extranjeros deberán evidenciar su fluencia en el idioma español mediante el resultado oficial de una prueba estandarizada con antigüedad no mayor a un año, o en su defecto presentar en CETYS Universidad la Prueba de Aprovechamiento de Español y obtener un resultado favorable para su programa académico.			

Los resultados de los exámenes de admisión obtenidos por los prospectos, se capturan en un archivo de Excel, y se exportan al Sistema de Admisiones y Escolar. A cada prospecto aprobado se le asigna una matrícula, de esta manera, la información de cada nuevo estudiante estará disponible al momento de que éste se inscriba en el Departamento de Servicios Escolares. Este departamento recibe la documentación por parte de promoción, firmada por el alumno y el padre de familia o tutor: la solicitud de admisión, las políticas institucionales, el aviso de privacidad y la carta de aceptación emitida por admisiones; además se hace la recepción de los documentos requisitos de inscripción, según lo indica el reglamento de alumnos. En este departamento se realiza el mantenimiento de la información de los alumnos; así como el seguimiento de su trayectoria académica desde su inicio, desarrollo y/o conclusión de los estudios cursados, mediante el sistema de escolar WEB.

La coordinación de informática hace entrega de su usuario y contraseña para el ingreso al portal de la institución, la red inalámbrica (WiFi), y la plataforma Blackboard, mediante estos accesos los alumnos podrán revisar calificaciones e historiales, realizar inscripción en línea e ingreso a la biblioteca digital. Por otro lado, se cuenta con la modalidad de ingreso a través de revalidación y equivalencia para aspirantes que están interesados en continuar sus estudios en la Institución y provengan de otra institución de educación superior. El Reglamento de Estudiantes

de Programas de Licenciatura (CETYS Universidad, 2014) en sus artículos 43 y 44, menciona los procesos de revalidación y equivalencia.

3.3.7 Apoyos financieros.

CETYS Universidad se caracteriza por brindar apoyos financieros a los estudiantes, se reconoce como una Institución en donde 8 de cada 10 estudiantes cuentan con este tipo de apoyo, si revisamos la Tabla 3-5 se pueden observar los criterios de otorgamiento de becas, en los cuáles los resultados del examen de admisión, el promedio que obtuvieron en el nivel académico anterior y su nivel socioeconómico son factores determinantes. Todas las becas son evaluadas cada semestre, ya que existen sanciones al dejar de cumplir los criterios, o bien, algunas becas puedes ser incrementadas durante la vida académica del estudiante (*Reglamento de Becas y Descuentos Del Sistema CETYS Universidad*, 2010).

Tabla 3-5 Becas y descuentos del Sistema CETYS

Categoría	Nombre de la beca	Descripción	Criterios de otorgamiento
Becas Académicas	Beca Talento Puede ser desde el 30 al 90% en base al puntaje del College Board.	Alumnos cuya trayectoria académica, potencial intelectual y conocimientos adquiridos en preparatoria sean los mejores.	• Promedio de 9.00 en adelante y no tener materias reprobadas una constancia oficial. • Examen de admisión College Board aprobado. • Estudio socioeconómico por parte de CETYS Universidad
	Beca Promedio 15% de beca	Reconoce el esfuerzo académico de los estudiantes	• Hacer constar su calificación global de 90.00 o más en historial a dos decimales. • No tener materias reprobadas • Datos socioeconómicos
	Beca Pro-Ingeniería 50% de beca	Apoya la promoción a las carreras de ingeniería	• Promedio general de 8.50 en Bachillerato y no tener materias reprobadas, comprobable con una constancia oficial • Puntaje integrado de 1,200 puntos en el examen de admisión

			• Datos socioeconómicos • Carta de una cuartilla dirigida al Comité de Becas donde explique el motivo
	Beca PAFENI 25% de beca	Estudiantes destacados a lo largo de su desempeño, que desean estudiar en campus Ensenada	• Promedio general de 8.00 en Bachillerato sin materias reprobadas, comprobable con una constancia oficial. • Alumno de nuevo ingreso • Iniciar sus estudios en el periodo académico inmediato a su admisión a la escuela.
Beca Laboral	Prestación laboral 100%	Exención total o parcial del pago de la colegiatura de un empleado activo de planta de tiempo completo o de medio tiempo, de su cónyuge o de sus hijos, en los programas Preparatoria, Profesional, Maestría	• Cumplir con todos los requisitos de inscripción
Becas patrocinadas	Montos específicos	Becas con fondos externos a CETYS y que la institución administra de acuerdo a los criterios acordados entre el patrocinador y la Institución.	• Cumplir con todos los requisitos de inscripción • Criterios pueden variar de acuerdo a lo que el patrocinador define
Beca Deportiva	Beca Deportiva Hasta el 100%	Deportistas de alto rendimiento de alguna disciplina contemplada en el campus	• Examen de admisión aprobado • Currículo deportivo • Examen de rendimiento deportivo
Descuentos de Apoyo Institucional	Descuento Egresado Hasta el 10%	Reconocer y apoyar a aquellos personas que son egresados de cualquier programa escolarizado del Sistema CETYS Universidad	• Ser considerado como alumno egresado de cualquier programa escolarizado de CETYS • Ser aceptado en el proceso de admisión.
	Descuento Hermano	Apoyo a la economía de aquellas familias que confían en la Institución	• Inscritos con carga completa • Datos socioeconómicos

			• El descuento aplica al hermano que tenga el plan de estudios más avanzado
	Descuento Hijo de Egresado Hasta el 10%	Se otorga a hijos de egresados	• El descuento es en el importe de la colegiatura y de la cuota de inscripción • El porcentaje no es acumulable con Descuento egresado.

3.3.8 Nivel Socioeconómico

Los estudios socioeconómicos son realizados por terceros, los cuales entregan los resultados al área de promoción para que los integren en los expedientes de los estudiantes, mismos que son enviados al Comité de Becas para su revisión y determinación del tipo y monto de beca a la que éste puede acceder (*Reglamento de Becas y Descuentos Del Sistema CETYS Universidad*, 2010).

A Ingreso Mensual Familiar		
Desde	Hasta	Puntos
85,000	mas	5
35000	84999	8
11600	34999	12
6800	11599	20
2700	6799	25

B Opción Ubicación	Opción Valor	
Zona 1 Chapultepec, Moderna, Amp. Moderna, San Marino, Loma Dorada, Cibolas del Mar, Coronita, Las Rosas, Juan Diego Residencial, Quinta Sta. Lucia, Lomas del Mar, Real de San Marino	Casas de más de 1,000,000 pesos	5
Zona 2 Valle Dorado, Col. Obrera, Carlos Pacheco, Costa Azul, Buenaventura, Nueva Ensenada, Bosque de los Olivos, Las Fincas, Arboledas, Bahia, Bahia Sur, Acapulco, Villa las Rosas sauzal, Colonia Maestros,Playa Ensenada,Puerta del Mar (siena y san borja residencial)	Casas hasta 1,000,000 pesos	10
Zona 3 Villas del Real, Ejido Chapultepec, Aviación, empleados, Sauzal, Jalisco, Marquez de León, Casos Geo, Casas de Urbi,Villas del Prado, Villa Bonita, Fovistte, Cumbre de la presa, Col. Azteca, Villas del sol, Vista Hermosa, Unidad Militar, Valle Verde, Popular 1,2 y 3, (hidalgo, granjas el gallo, villas del rey, aeropuerto)	Casas hasta 600,000 pesos	15
Zona 4 La presa, Ruiz Cortinez, Popular 89, Bella Vista, Emiliano Zapata,Aguajito, La Joyita, Francisco Zarco, Vista al mar, maneadero, Indeco 1,2 y 3,(loma linda)	Casas hasta 250,000 pesos	20

C Trabajo	
grandes o medianos empresarios (en el ramo industrial, comercial y de servicios); como gerentes, directores o destacados profesionistas.	5
empresarios de compañías pequeñas o medianas, gerentes o ejecutivos secundarios en empresas grandes o profesionistas independientes	10
pequeños comerciantes, empleados de gobierno, vendedores, maestros de escuela, técnico y obreros calificados,Encargado de ventas,	15
taxistas (choferes propietarios del auto), comerciantes fijos o ambulantes (plomería, carpintería), choferes de casas, mensajeros, cobradores, obreros,Enfermera, militares de medio rango, Encargado de Almacen	20
obreros, empleados de mantenimiento, empleados de mostrador, choferes públicos, maquiladores, comerciantes,Secretaria, Albañil, Mecanico	25
subempleos o empleos eventuales, jubilados	30

Figura 3-4 Nivel socioeconómico: criterios y puntajes (CETYS Universidad)

En este momento, la Institución no se cuenta con los resultados de los estudios socioeconómicos de los aspirantes; cuestión por la cual se ha construido una tabla con los criterios utilizados para identificar el nivel socioeconómico y se analizaron los expedientes para definirlo. La tabla construida se muestra en la Figura 3-4, los aspectos evaluados son: ingreso mensual familiar, zona residencial, tipo de trabajo del estudiante, padre o tutor, se puede observar que a mayor puntaje mayor necesidad económica.

3.3.9 Estancia en la Institución

Durante el trayecto por la vida de Educación Superior, el estudiante tiene un acompañamiento en donde participan diferentes instancias para asegurar el desempeño académico, su permanencia y eficiencia terminal, como los coordinadores cada programa académico, los directores de escuela; además, a partir del año 2015 la Institución cuenta con un Centro de Desarrollo Estudiantil (CEDE) el cual brinda servicios de apoyo educativos y psicológicos que contribuyan al desarrollo profesional, académico y personal de los alumnos de nivel profesional, así como contribuir a las metas institucionales de retención y eficiencia terminal.

El CEDE cumple sus objetivos a través de diversos programas preventivos, de intervención y de seguimiento, por ejemplo:

- Programa de Apoyo Universitario.
- Programa con alumnos deportistas.
- Programa de asesorías entre compañeros.

En cuanto al Programa de Apoyo Universitario, se desarrollan talleres de fortalecimiento académico dirigidos a estudiantes admitidos condicionados y todo alumno interesado de nivel profesional. Los carteles promocionales y las asistencias se registran en una hoja de Excel compartida a través de Google Drive con los coordinadores de programa y mentores, para su seguimiento, también se lleva una bitácora de registro de los alumnos que acuden a CEDE para una sesión de orientación individual, ya sea de seguimiento a alumno deportista, académica o asesorías. El seguimiento que se da a los alumnos busca recuperar la información académica del estudiante, así como información personal que permita realizar las valoraciones adecuadas, ofrecer los servicios de apoyo al estudiante y brindar información al cuerpo académico para la toma de decisiones con respecto a cada caso. En el CEDE el estudiante inicia el proceso de baja y con base en la información obtenida por el estudiante se define el tipo de baja, la clasificación de las bajas se muestra en la Tabla 3-6; es importante hacer notar que la deserción a otros campus del Sistema CETYS no se considera baja, ya que el estudiante permanece en el mismo sistema.

El portal *mi Campus* es una herramienta que consolida páginas web, sistemas administrativos, servicios académicos, que en conjunto satisfacen de manera integral los requerimientos de información de alumnos y padres de familia del CETYS. Este sistema integra información académica, financiera, inscripciones, portafolio de proyectos por materia, es el medio

por el cual los profesores pueden registrar las evaluaciones parciales y finales de acuerdo al calendario escolar.

A través del Sistema de Inscripciones en línea, el alumno selecciona el semestre o trimestre a cursar, selecciona las materias que desea cursar, -las cuales previamente fueron seleccionadas por el Coordinadores académico y/o Director de escuela-, y registradas por el área de Operación Académica (OPA) en una plantilla académica. Completado el proceso correspondiente, el alumno quedará registrado y contará con su horario de clases del semestre.

Tabla 3-6 Clasificación de los motivo de baja.

Motivo de baja	Descripción
BD	Baja definitiva
BT	Baja temporal
CM	Cambio campus
CS	Cambio de semestre
CT	Cuestiones de trabajo
FA	Falta de adaptación
IE	Incapacidad económica
MS	motivos de salud
CCA	Cambio de carrera
OC	Otras causas
PH	Problemas de horario
CC	cambio de ciudad
OC	Otras causas
CS	Cambio de semestre
DV	Desorientación vocacional
CC	cambio de ciudad
CE	Cambio de escuela
PF	Problemas familiares
CT	Cuestiones de trabajo
IE	Incapacidad económica
CE	Cambio de escuela
PH	Problemas de horario
CP	Cambio programa

En CETYS Universidad se disponen los cursos que se ofertarán a los estudiantes en la plataforma Blackboard, ya sean en modalidad presencial o en línea (e-campus). El departamento de Sistemas y Redes de cada campus se encarga de registrar los cursos de cada semestre y a los estudiantes inscritos en cada curso.

Para los programas de licenciatura, durante el ciclo escolar se cuenta con dos períodos de evaluación parcial. Al finalizar cada uno de ellos el CEDE genera en el portal mi Campus el

reporte de desenvolvimiento académico, en él se concentra el promedio obtenido, además de materias aprobadas y no aprobadas. Cuando un estudiante muestra una reprobación mayor al 40%, se procede a notificar a las direcciones y coordinaciones de escuela, y se cita al estudiante para identificar y solventar oportunamente sus necesidades académicas. El área académica y administrativa de la Institución, cuenta con un Sistema de información (SICU), cuya función es la de generar reportes y estadísticos referentes al alumnado y profesorado de todos los programas.

Como parte del proceso de mejora continua y detección de necesidades, la Institución realiza una encuesta de satisfacción al semestre; para lo cual, selecciona una muestra obtenida con la fórmula de poblaciones finitas y por selección, que represente a la población estudiantil de profesional. El área responsable es la Dirección de Promoción y Desarrollo Institucional, después es procesada por el área de Investigación y Evaluación Institucional, quien finalmente obtiene las conclusiones y reportes de esta encuesta (Encuesta Satisfacción 2017-2 Campus Ensenada, 2017).

Esta encuesta empezó a aplicarse desde 2002 y evalúa las siguientes áreas:

1. Servicios
2. Instalaciones
3. Coordinador / Director de Carrera
4. Actividades extracurriculares
5. Aspectos cualitativos
    a. Orgullo y pertenencia
6. Satisfacción
7. Ambiente estudiantil

Criterios para la obtención de conclusiones:

1. Para efectos de medición se toman en cuenta los datos que representen oportunidades de mejora. (pésimo, malo y regular), "regular" se toma como parámetro negativo, ya que implica que se puede mejorar.
2. Para identificar las áreas de mejora se suman las calificaciones "pésimo", "malo" y "regular". Esto es el FM=Factor de Medición Si esta suma es igual o superior a 25%, se anota en la tabla de hallazgos.
3. De acuerdo a los estándares de calidad de la Institución, no se puede obtener una calificación menor al 75%.

4. En las preguntas abiertas, como: "¿Por qué?", se toman en cuenta generalmente las 3 de mayores porcentajes.

El nivel de consulta de los resultados es: Programa Académico, Escuela y Nivel Académico, por ejemplo: Programa: Ingeniería de Software, Escuela: Ingeniería y Nivel Académico: Licenciatura. Este reporte es entregado a cada Director de Escuela y de Programa según corresponda. Cada año el instrumento es revisado para hacer los ajustes y adecuaciones pertinentes.

3.3.10 La vida del estudiante en la Institución

La Dirección de Formación Integral Universitaria tiene como objetivo consolidar la formación integral de los estudiantes mediante la renovación e innovación de estrategias cocurriculares. Se vincula con distintas áreas de la Institución, con el propósito de articular los servicios educativos y alinearlos a los resultados de aprendizaje.

La vinculación tiene que ver con:

- La academia: de manera particular con la Escuela de Humanidades que ofrece el Eje de Formación General a todos los estudiantes.
- Asuntos Estudiantiles: cuyas distintas coordinaciones ofrecen acompañamiento a los distintos grupos estudiantiles y promueven actividades que se denominan de “vida estudiantil”.
- Otras áreas de apoyo a los estudiantes, con quienes se generan programas que inciden en la formación integral. Ejemplo importante de estos programas es el “Sí a la Vida” que aglutina iniciativas que promueven estilos de vida saludables, así como la detección de conductas de riesgo.
- Difusión Cultural: promoción del aprecio por la cultura y de habilidades artísticas.
- Liderazgo Estudiantil: desarrollo de habilidades de liderazgo, trabajo en equipo y responsabilidad social.
- Bienestar Integral: desarrollo de la capacidad funcional y promoción del deporte recreativo.
- Deporte representativo: desarrollo de habilidades deportivas de alto rendimiento.
- Eje de Formación General: nueve asignaturas que promueven: el desarrollo de “Habilidades blandas”, la Cultura de la Información, comprender ventajas y riesgos

de la globalización de nuestra cultura, el aprecio por el arte y la cultura contemporáneos, la reflexión sobre preguntas trascendentales que tienen que ver con las relaciones del ser humano:

  - Como parte de una sociedad humana
  - Como parte de un entorno natural
  - Como ser moral

- Talleres de media y fin de carrera: momentos especiales de reflexión en el que los alumnos revisan su trayecto universitario y ajustan sus metas a futuro.

La normatividad, las políticas y reglamentos de ingreso y permanencia en la Institución, así como los sistemas de información con los que opera CETYS Universidad, permiten que la información del estudiante que ingresa a la Institución vaya dejando una huella digital de su experiencia académica y personal. Si bien es cierto que la institución cuenta con tecnologías de información, también carece de sistemas de información integrados que permitan registrar y concentrar la información del estudiante para facilitar su seguimiento y la toma de decisiones para asegurar su permanencia y favorecer la eficiencia terminal.

3.4 Recolección de los datos

Se debe recolectar la información relevante del estudiante desde su ingreso a la Institución hasta su salida, para ello se construye un modelo lógico que permita hacer el análisis de los datos con minería de datos y con ellos describir y explicar la deserción, para finalmente construir un modelo predictivo a través de la información histórica de la deserción. Para ello, se ejecuta el proceso de extracción, transformación y carga de datos, ETL por sus siglas en inglés (*Extract, Transform and Load*), lo cual permite concentrar los datos que se generan de diferentes fuentes y plataformas, haciendo una limpieza y transformación de éstos dentro un contexto semántico para proveer de conocimiento que permita analizar la información (Alnoukari, M. and Hanano, 2017).

3.4.1 Identificación de las fuentes de información

Las fuentes de datos son los siguientes sistemas de información:

- Sistema Escolar WEB: Se localizan los datos de inscripción como matrícula, programa de estudios, periodo, ficha de inscripción (materias, créditos, semestre, pagos, becas,

entre otros), historial académico, grupos inscritos (materia, profesor, horario, salón asignado, entre otros).

- CEDE: Hojas de cálculo con control de asistencia a cursos de alumnos condicionados, control de visitas del estudiante, asesorías entre pares, información sobre la baja del estudiante.
- Sistema de Finanzas: Se cuenta con los datos financieros del estudiante como: datos del padre o tutor, asignación de beca por tipo y por período, pagos, tipo de pagos, períodos de pago, fichas de inscripción, entre otros.
- CRM: Información recolectada por el área de promoción y admisiones, como datos del examen de admisión, fecha de admisión, datos personales como: dirección, datos de los padres, escuela de egreso, promedio de egreso de bachillerato, entre otros.
- Portal mi campus: Información curricular que contiene calificaciones parciales, finales por materia, por semestre, beca aplicada por semestre, desempeño académico del estudiante, extraordinarios y calificación obtenida, además se actualiza la información cocurricular como participación en grupos de impacto y liderazgo social, grupos deportivos, viajes de estudio, experiencias de intercambio internacionales y nacionales, prácticas profesionales, servicio social, entre otros.
- Sistema de información de CETYS Universidad (SICU): Información general de cada periodo escolar como evaluaciones docentes, estudios de satisfacción, entre otros.

3.4.2 Proceso de Extracción, transformación y carga de datos (ETL)

La extracción de los datos se inició desde el Sistema de Escolar, seleccionando a todos los estudiantes que no completaron su programa académico y que tuvieran un estatus de baja, tomando como base esta información se recuperaron los datos del estudiante durante su estancia en la Institución.

Como se puede observar en la Figura 3-5, se realizó una extracción de datos actuales e históricos, formales e informales, que pudieran contener información relevante para cumplir el propósito del estudio, se identificaron a los estudiantes desertores del período del 2008 al 2018; es decir, aquellos que se dieron de baja en algún período de su carrera sin concluir sus estudios y se identificó cuál fue el motivo de la baja.

Enseguida, se reunió la información que se genera alrededor del estudiante desde su primer contacto con la Institución; los datos fueron obtenidos de diversos sistemas digitales de CETYS tales como: el sistema de promoción (CRM), el sistema WEB de escolar, el sistema de finanzas, el portal de información del estudiante y del Centro de Desarrollo del Estudiante (CEDE).

La recolección de datos se hizo con el objetivo de construir una base de datos, cuyo modelo lógico representará las distintas dimensiones que caracterizan a los desertores. En el proceso de transformación se prepararon los datos utilizando la herramienta Tableau prep de Tableau Inc, utilizando técnicas de discretización, convirtiendo algunas variables numéricas en categóricas como la escala numérica que el estudiante utiliza para evaluar su nivel de satisfacción en la encuesta semestral, el género, el nivel socioeconómico, la clave de la escuela de procedencia, la clave de la ciudad de procedencia, la condición del estudiante al ingresar a la IES (condicionado, aprobado), el tipo de escuela de la que procede el estudiante (privada o pública), el tipo de estudiante (foráneo o local). Se detectaron y corrigieron las instancias que contenían datos corrompidos, se hicieron tareas de agregación, unión, agrupamiento, reducción, incluso se crearon atributos a partir de otros para cada instancia como aquellos que indicaban que tenía curriculum deportivo, si tenía logros académicos, si existía alguna relación previa por lo tanto tenían confianza en la institución (si era egresado de la preparatoria CETYS el y/o sus padres). Es importante mencionar que se hizo una imputación de datos ausentes que se encontraron en las áreas del examen de admisión: área verbal, matemática y redacción, mismo que se describirá más adelante; el proceso completo puede verse en la Figura 3-6.

Se observaron algunos datos ausentes, ya que alrededor del 14% de los registros no contaban con los puntajes de las áreas que conforman el resultado del examen de admisión, las inconsistencias fueron más frecuentes cuando se trataba de la información personal del estudiante, sin embargo, por el volumen de las instancias registradas se hizo una revisión manual de la información, aunque las fuentes fueron diversas, se pudo probar la fidelidad de la información, ya que se comprobaron a través de la documentación original presentada por el estudiante en cada expediente físico con el que contaba la Institución. Sin embargo, por la relevancia que representan las áreas de evaluación del examen de admisión, se realizó una imputación de datos por medio de regresión lineal por la naturaleza numérica de sus datos, se definió como set de entrenamiento aquellos registros que contaban con toda la información del puntaje en cada área, representando el 67% del total de registros.

Las ecuaciones obtenidas fueron las siguientes:

Razonamiento matemático = -75.37 + 0.57026(puntaje global de admisión)

Razonamiento verbal = 51.757 + 0.44727(puntaje global de admisión)

Área de redacción = 103.713 + 0.22696 (puntaje global de admisión) + 0.34750 (verbal)

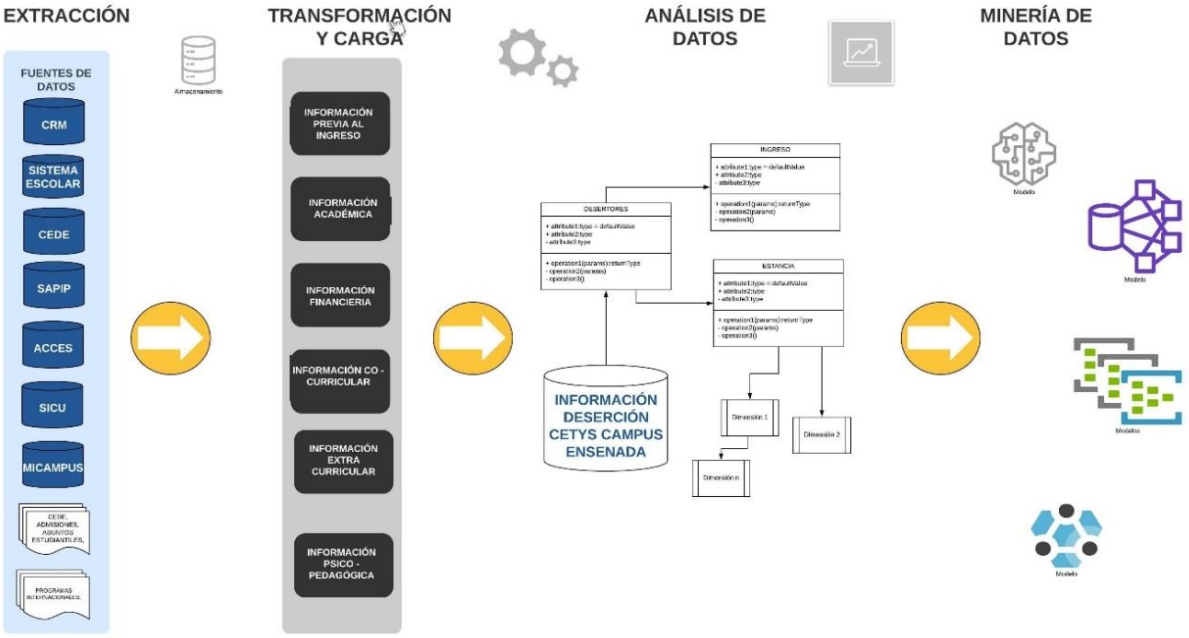


Figura 3-5 Modelo ETL. Fuente: Elaboración propia.

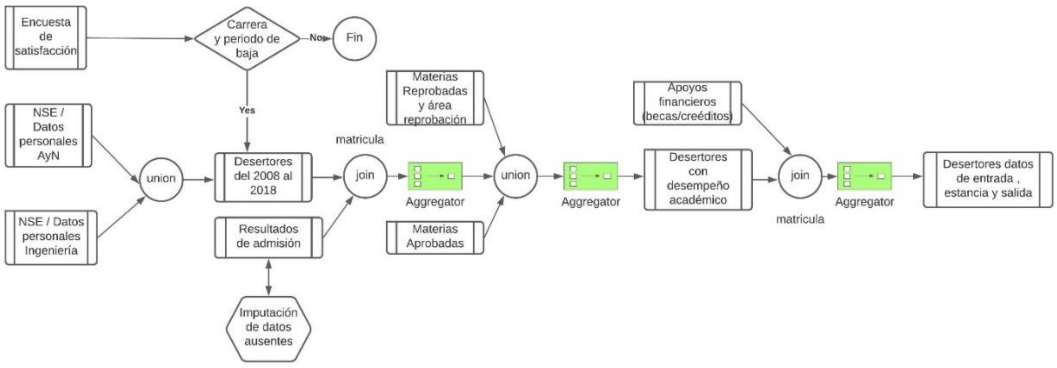


Figura 3-6 Transformación de datos. Elaboración propia

Modelo multidimensional

El modelo lógico se diseñó visualizando las dimensiones que explican la deserción de los estudiantes de CETYS campus Ensenada, en el período de agosto de 2010 a agosto de 2018, a partir de las fuentes de datos disponibles. Como se mostró en la Figura 3-2, en el modelo entidad–relación de los atributos del estudiante al ingresar y durante su estancia en la IES, se identificó la información disponible en las fuentes de datos de la Institución, para el diseño del modelo multidimensional que se presenta en la Figura 3-7.

El modelo multidimensional tiene definida la tabla de hechos de los desertores y las dimensiones de los datos de ingreso y de su estancia en la IES. La tabla de hechos contiene la definición de los estudiantes desertores, la matrícula que se ha definido como la llave primaria de la relación, la escuela, el programa académico, el plan de estudios en el que estaba inscrito cuando dejó la Institución, a que cohorte perteneció, la fecha en la que se inscribió y en la que se dio de baja. Las dimensiones que caracterizan a cada uno de los desertores son:

- Los datos del estudiante como su nombre completo, si es foráneo, si vivió un intercambio, el promedio con el que ingresó de su nivel académico previo, la escuela de procedencia, su resultado final en su examen de ingreso.
- El estatus de salida al momento de darse de baja como el motivo de la baja, la fecha de baja, su último programa y plan de estudios inscrito, datos obtenidos de la Tabla 3-3, el último período de inscripción.
- Los resultados de admisión como la fecha de presentación del examen, el puntaje total, el puntaje por área de conocimiento: razonamiento matemático, razonamiento verbal y la habilidad cognoscitiva, además del estado de admisión a la institución, información adquirida de la Tabla 3-4.
- En cuanto al desempeño académico, se conoce el número de materias aprobadas y reprobadas, el área de reprobación según la tipología de competencias mostrados en la Tabla 3-3, los exámenes extraordinarios presentados, las materias totales cursadas en un período determinado, las faltas totales acumuladas, el programa y plan de estudios correspondientes a ese período, además si hubo cambio de programa académico.

- Los apoyos financieros que el estudiante recibió desde su ingreso y durante su estancia en la IES, como el tipo de beca y el monto de apoyo, información de la Tabla 3-5.
- La percepción cualitativa del estudiante a través de los resultados de las encuestas de satisfacción que se aplican cada semestre, en donde se puede visualizar la opinión de los estudiantes sobre las instalaciones, el apoyo de la institución, los servicios, el orgullo de pertenencia, su evaluación del ambiente estudiantil, las actividades extracurriculares, la satisfacción en general.

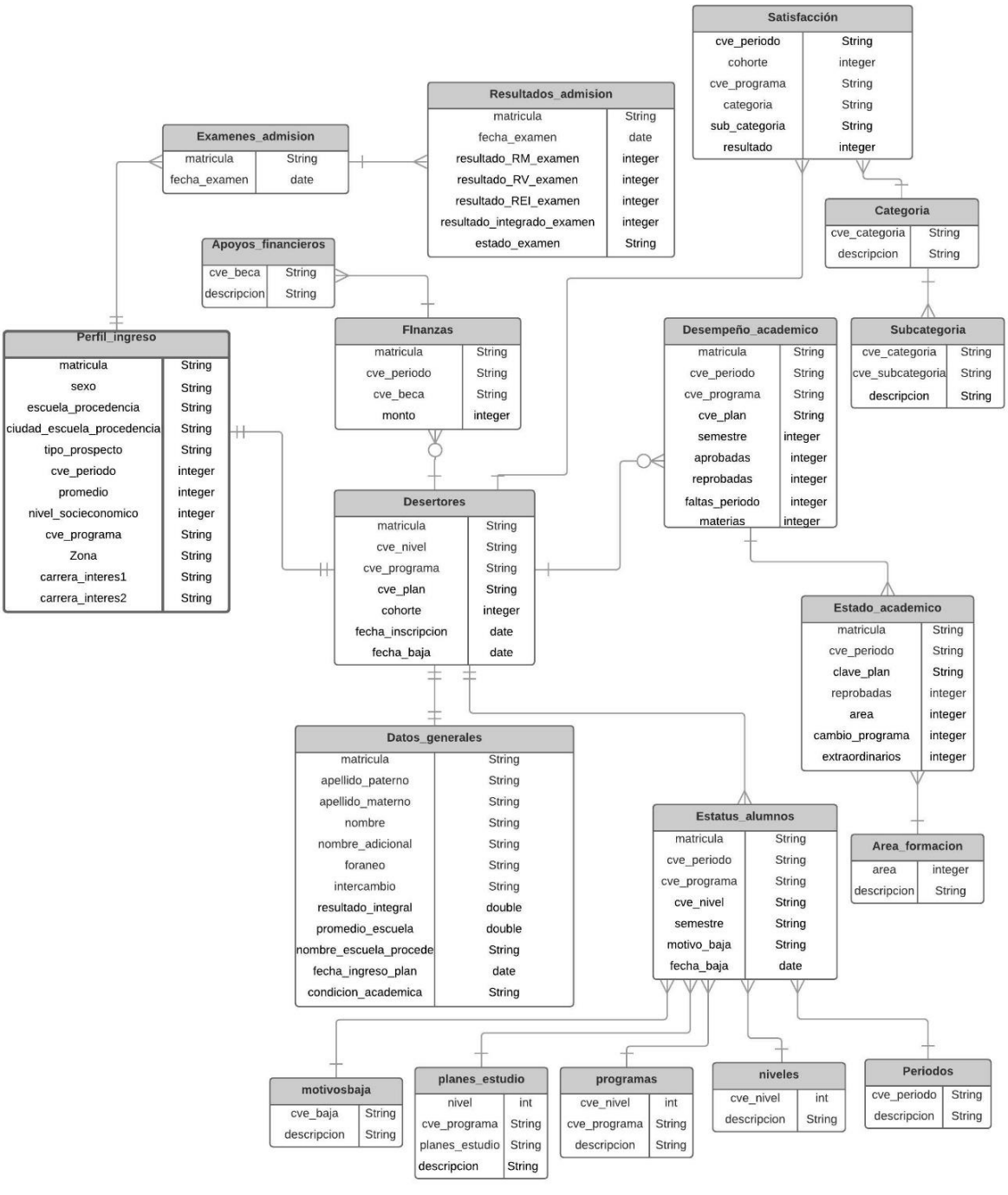


Figura 3-7 . Modelo Multidimensional Deserción. Fuente: Elaboración propia.

Diccionario de datos

Los datos que se extrajeron de las diferentes fuentes de información consolidan la información para el estudio de ésta investigación, como se puede observar en la Figura 3-5, después de la transformación, limpieza y carga de datos se construyó la base de datos con las siguientes tablas: En el Anexo 2 se pueden consultar la definición de cada uno de los datos que contienen las relaciones del modelo multidimensional.

1. **Desertores:** Tabla de hechos, alumnos que desertaron.
2. **Finanzas:** Tabla de apoyos financieros.
3. **Apoyos financieros:** Claves de apoyos financieros.
4. **Estatus alumnos:** Tabla que muestra el estatus del estudiante inscrito o dado de baja.
5. **Motivobaja:** Descripciones de los motivos de baja.
6. **Programas:** Descripciones de los programas académicos.
7. **Planes de estudio:** Descripciones de los planes de estudio.
8. **Niveles:** Descripciones de las escuelas de educación superior.
9. **Periodos:** Descripciones de los períodos académicos.
10. **Perfil_ingreso:** Estatus de los alumnos prospectos a ingresar a la Institución
11. **Resultados_admision:** Tabla de resultados de cada una de las áreas del conocimiento evaluadas en el examen de admisión.
12. **Desempeño_academico:** Tabla con la información de reprobación por área de competencia, cambios de programas y presentación de extraordinarios por período.
13. **Área de formación:** Tabla de áreas de formación según la tipología de competencias.
14. **Satisfacción:** Tabla de resultados de la evaluación semestral de satisfacción del estudiante.
15. **Categorías:** Tabla de categorías o áreas de la evaluación de satisfacción.
16. **Subcategorías:** Tabla de subcategorías de la evaluación de satisfacción.

Base de datos

CETYS Campus Ensenada inició operaciones en 1975, a la fecha cuenta con 3,200 egresados de licenciatura. El reporte oficial del Sistema de Información SICU (CETYS, 2018a), muestra que en el periodo de 2008 a 2018 ingresaron 897 estudiantes, de los cuales desertaron 337, como se observa en la Tabla 3-7, la muestra obtenida para este estudio es de 355 desertores, ya que en

la base de datos se encuentran desertores de cohortes anteriores al 2010 como se indica en las primeras siete columnas.

Tabla 3-7 Registros de ingreso (I) y deserción (D) por cohorte, Sistema de Información CETYS (SICU)

COHORTE		2000	2002	2003	2004	2005	2006	2007	2008		2009		2010		2011		2012		2013		2014		2015		2016		2017		2018	
PROGRAMA	GÉNERO								I	D	I	D	I	D	I	D	I	D	I	D	I	D	I	D	I	D	I	D	I	D
LAM	HOMBRES								4	2	0	0	5	2	3	2	4	3	4	2	3	1	6	5	2	2	0	0	4	0
LAM	MUJERES								2	1	6	3	11	4	7	4	2	1	10	7	7	3	8	2	5	0	4	3	5	1
LDG	HOMBRES								7	2	0	0	5	0	2	0	2	0	1	0	0	0	1	0	3	1	3	1	3	0
LDG	MUJERES								5	0	11	3	6	3	3	1	7	4	6	5	2	1	8	5	4	3	7	3	3	0
LNI	HOMBRES								12	9	7	3	14	4	29	7	8	3	8	5	17	7	7	1	17	6	12	0	3	1
LNI	MUJERES								9	5	15	7	14	4	20	8	10	4	10	5	13	5	9	4	14	7	17	4	12	1
ICE	HOMBRES								6	2	6	3	7	2	7	3	2	0	2	1	0	0	2	0	2	0	1	0	0	0
ICE	MUJERES								1	0	1	0	0	0	2	1	1	0	0	0	0	0	2	1	2	0	0	0	0	0
IDGD	HOMBRES								0	0	7	4	5	4	3	2	5	1	2	0	3	1	3	1	3	0	3	0	3	0
IDGD	MUJERES								0	0	8	2	3	2	1	0	3	2	3	1	2	0	7	3	4	1	3	0	4	0
IIND	HOMBRES								2	1	2	0	5	3	2	1	11	7	9	3	3	0	6	3	7	3	11	7	9	3
IIND	MUJERES								1	0	7	4	3	0	5	2	3	0	1	0	4	3	5	2	2	0	3	0	0	0
IMEC	HOMBRES								7	5	0	0	10	7	9	7	8	3	8	3	11	10	11	3	11	4	9	5	16	1
IMEC	MUJERES								1	0	0	0	2	0	0	0	3	2	4	2	2	1	5	1	3	0	3	1	1	0
ISW	HOMBRES								3	2	2	1	6	0	4	1	6	2	7	3	3	1	7	2	3	1	9	2	8	1
ISW	MUJERES								3	1	1	0	1	0	2	1	2	1	2	1	2	0	2	0	2	1	6	2	4	0
Desertores del estudio		1	2	2	4	3	8	15		39		37		48		48		46		35		43		27		9		2		
TOTAL REGISTROS		1	2	2	4	3	8	15	63	30	73	30	97	35	99	40	77	33	77	38	72	33	89	33	84	29	91	28	75	8
	DESERTORES	372																												
	INGRESARON	897																												

3.5 Construcción del modelo predictivo

3.5.1 Descripción de la base de datos

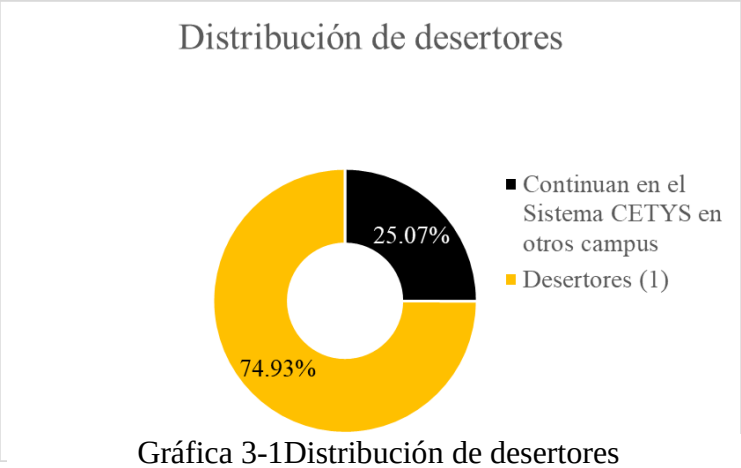
La muestra de 353 desertores organizada por cohorte, representa la población de deserción del 2008 al 2018 analizada en esta investigación. Los desertores de este período pertenecen a cohortes diferentes, como se muestra en la Tabla 3-7, se puede identificar que algunos de los estudiantes que desertaron en este periodo de tiempo pertenecen a cohortes del 2000 al 2007, sin embargo, estos desertores representan un 9.7% del total, por lo que se concluyó que el análisis desde el enfoque de la cohorte no es significativo para la construcción de un modelo desde esa agrupación, sin embargo la valoración cualitativa en el contexto institucional durante su estancia en la IES como se observa en el modelo entidad-relación de la Figura 3-2, habla de una relación de satisfacción importante con la institución de estos estudiantes, porque el 31.3% cursó más de la mitad de los semestres de su programa académico como se puede observar en la distribución presentada en la Tabla 3-8 Último semestre cursado.

Tabla 3-8 Último semestre cursado

Último semestre cursado									Total de desertores	% del total
	1	2	3	4	5	6	7	8		
Total	131	64	49	37	35	18	6	16	355	**100.00%**
% del Total	36.9%	18%	13.8%	10.4%	9.8%	5%	1.6%	4.5%		

Para describir el fenómeno de la deserción, se hizo un muestreo estratificado por escuela y programa académico, asumiendo que los perfiles vocaciones y las intenciones de ingreso a la institución pudieran estar asociadas a la carrera seleccionada, como lo menciona Pascarella & Terenzini (1980) en su modelo causal, el desertor no solo es aquel que abandona la IES voluntariamente o circunstancialmente sino aquel desertor persistente, que pudo haber sido dado de baja por no cumplir con los requerimientos académicos de la institución, muchas veces por las exigencias del programa académico elegido o las propias competencias del estudiante.

De acuerdo al modelo entidad–relación (MER) de los atributos del estudiante al ingresar y durante su estancia en la IES, mostrado en la Figura 3-2, se realiza un análisis descriptivo en los tres contextos: personal, académico e institucional, y se lleva a cabo una clasificación de los atributos contenidos en el MER. La descripción estadística se lleva a cabo con los registros cuyo valor es de uno en la variable dependiente *MotivoBaja*, ya que estos son los estudiantes desertores y representan el 74.93% de la muestra. En la gráfica 3-1 se muestra la distribución de los estudiantes desertores.



Gráfica 3-1Distribución de desertores

3.5.2 Atributos personales al ingresar a la IES

Cuando los prospectos se convierten en estudiantes de la Institución pasan por el proceso de admisión descrito en este capítulo en la sección de atributos de la población foco, en este proceso se registran los datos personales, académicos y las variables que definen el tipo de beca o crédito al que el prospecto puede acceder según el programa de apoyos financieros con el que cuenta la Institución, esta información fue descrita en la Tabla 3-5 de becas y descuentos del sistema CETYS. Los atributos personales, académicos e institucionales que se incluyen en este estudio se describe en las Tablas 3-9, 3-10 y 3-11 respectivamente.

Tabla 3-9 Atributos personales al ingresar a la IES

Momento	Contexto	Variable	Descripción	Tipo	Fuente
Al ingresar a la Universidad	Personal	edad	Edad del estudiante al ingresar a la IES	numérico	Promoción
		genero	Genero del estudiante (1Femenino, 2 Masculino)	nominal	Promoción
		NSE	Nivel socioeconómico (1 Bajo, 2 Medio, 3 Alto)	ordinal	Promoción
		VExperienciacInstitucion	Experiencia previa con la institución, hijo de egresado y/o egresado (1 Si, 0 No)	bandera	Promoción
		cve_prospecto	Estudiante local, foráneo nacional e internacional (1 Local, 2 Foráneo, 3 Internacional)	nominal	Promoción
		ciudad_escuela_procedencia	Ciudad de origen (ID Ciudad)	categórico	Promoción

Tabla 3-10 Atributos académicos al ingresar a la IES

Momento	Contexto	Variable	Descripción	Tipo	Fuente
Al ingresar a la IES	Académico	cve_escuela	Escuela donde se encuentra el programa académico elegido (1 Ingeniería, 2 Administración)	nominal	Escolar

		cve_programa	Programa académico elegido (1.ICE, 2. IDGD, 3.IER, 4. IIND, 5.IM, 6. IMEC, 8.IS de Ingeniería, 9.LAE, 10.LAM, 11.LDG, 12.LNI Administración y Negocios)	nominal	Escolar
		admision	Resultados del examen de admisión	numérico	Promoción
		verbal	Resultados del examen de admisión área verbal	numérico	Promoción
		matematico	Resultados del examen de admisión área matemática	numérico	Promoción
		redaccion	Resultados del examen de admisión área redacción	numérico	Promoción
		promedio_ingreso	Promedio general de bachillerato	numérico	Promoción
		cve_escuelaingreso	Escuela de ingreso (ID numérico)	categórico	Promoción
		tipo_escuela	Sector a la que pertenece la escuela de ingreso (2 privada o 1 pública)	nominal	Promoción
		VExcelenciaacademica	Estudiante con excelencia académica mayor que 9.5 (1 Si, 0 No)	bandera	Promoción
		VPromedioacademico	Estudiante con un promedio académico mayor que 9 (1 Si, 0 No)	bandera	Promoción
		Vdeportista	Experiencia cocurricular deportiva (1 Si, 0 No)	bandera	Promoción

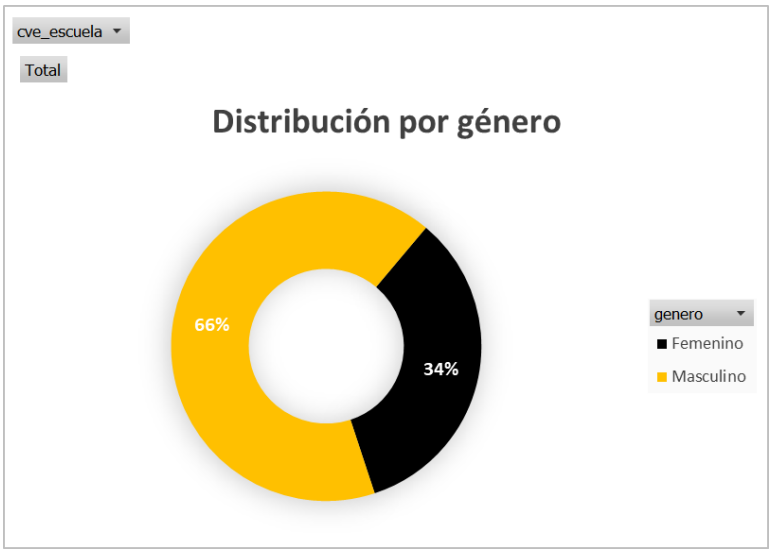
Tabla 3-11 Atributos institucionales al ingresar a la IES

Momento	**Contexto**	**Variable**	**Descripción**	**Tipo**	**Fuente**
Al ingresar a la IES	Institucional	becaingreso	Monto de la beca de ingreso	numérico	Promoción
		Pafeni	Beca para apoyar a aquellos estudiantes destacados en el Campus Ensenada.	numérico	Finanzas
		pro_ing, ping	Beca apoyar la promoción a las carreras de ingeniería.	numérico	Finanzas

		Piafi, piaf	Alumnos que ingresan a cualquier programa de Licenciatura, destacados por sus aptitudes intelectuales y su trayectoria académica en Bachillerato	numérico	Finanzas
		egre	Pertenece a un programa que tiene como objetivo reconocer y apoyar a aquellos personas que son egresados de cualquier programa escolarizado del Sistema CETYS Universidad y que desean continuar sus estudios en cualquiera de los campus	numérico	Finanzas
		cedu	Contrato de apoyo especial de empresa	numérico	Finanzas
		onal	Apoyo especial por campaña promocional	numérico	Finanzas
		talento	La Beca Talento es otorgada a aquellos alumnos cuya trayectoria académica, potencial intelectual y conocimientos adquiridos en preparatoria sean los mejores	numérico	Finanzas
		dg	Apoyo especial de Dirección General por un año	numérico	Finanzas
		depo	Apoyo a los buenos alumnos que sean también excelentes deportistas	numérico	Finanzas
		conc	Beca de concurso	numérico	Finanzas
		excelen	La beca de Excelencia es un programa de reconocimiento académico a las principales preparatorias que a lo largo de los últimos años han proveído de alumnos de nuevo ingreso con alto nivel académico al CETYS Universidad,	numérico	Finanzas
		rect	Apoyo especial de Rectoría por un tiempo determinado	numérico	Finanzas
		herm	Es un apoyo a la economía de aquellas familias que han honrado a la institución con su preferencia inscribiendo a más de un hijo en preparatoria o profesional en	numérico	Finanzas

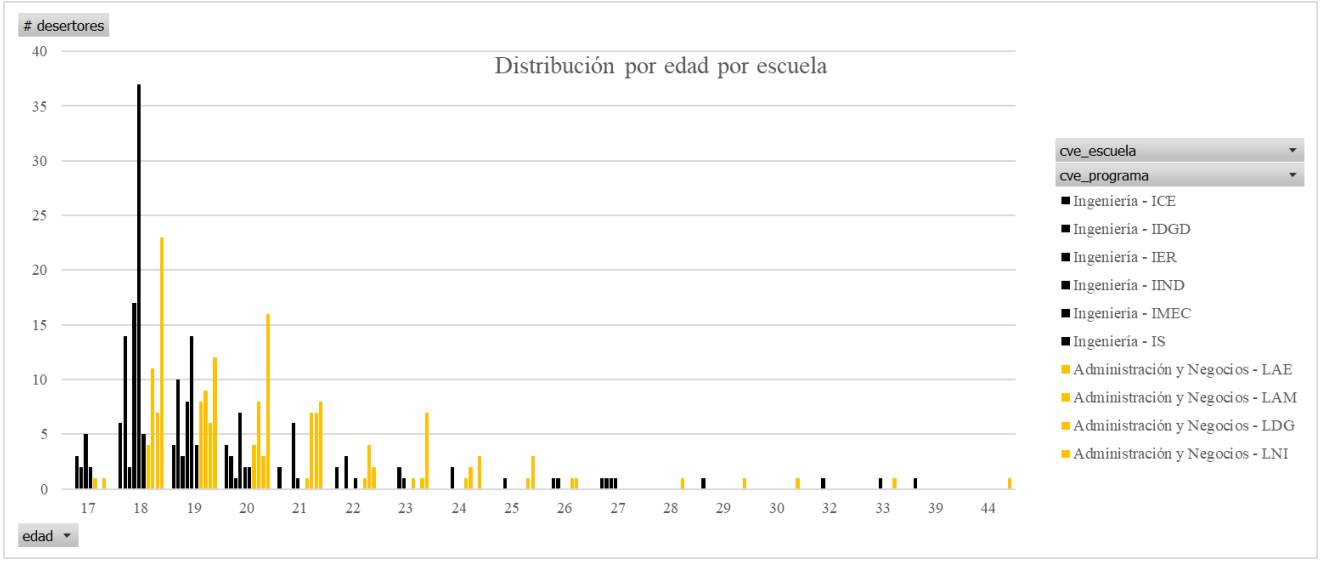
			el mismo período escolar y que tienen necesidad de este apoyo.		
		labo	beca de prestación laboral consiste en la exención total o parcial del pago de la colegiatura de un empleado activo de planta de tiempo completo o de medio tiempo, de su cónyuge o de sus hijos	numérico	Finanzas
		Hieg,	Descuento se otorga a hijos de egresados que desean ingresar a un programa escolarizado en CETYS.	numérico	Finanzas
		sgv	Seguro de vida	numérico	Finanzas

Como se observa en la gráfica 3-2, de todos los desertores en este estudio, el 66% son hombres y el 34% son mujeres; al analizar estos datos por escuela; en el área de Administración y Negocios las mujeres que desertan son el 14.7% del programa de Negocios Internacionales (LNI) y el 13.3% de Administración de Mercadotecnia (LAM); en cuanto a los hombres el 33.09% de los desertores son del programa de Negocios Internacionales (LNI). La distribución de género entre los desertores de la escuela de Ingeniería indica que el 23.85% de los hombres pertenecían al programa de Ingeniería Mecatrónica (IMEC), en cuanto a las mujeres el 7.69% pertenecieron al programa de Ingeniería en Diseño Gráfico Digital (IDGD) y otro 6.15% al programa de Ingeniería Industrial (IIND). Por lo tanto, los programas académicos con mayor deserción es el de LNI para la escuela de Administración y Negocios y el de IMEC para la escuela de Ingeniería.



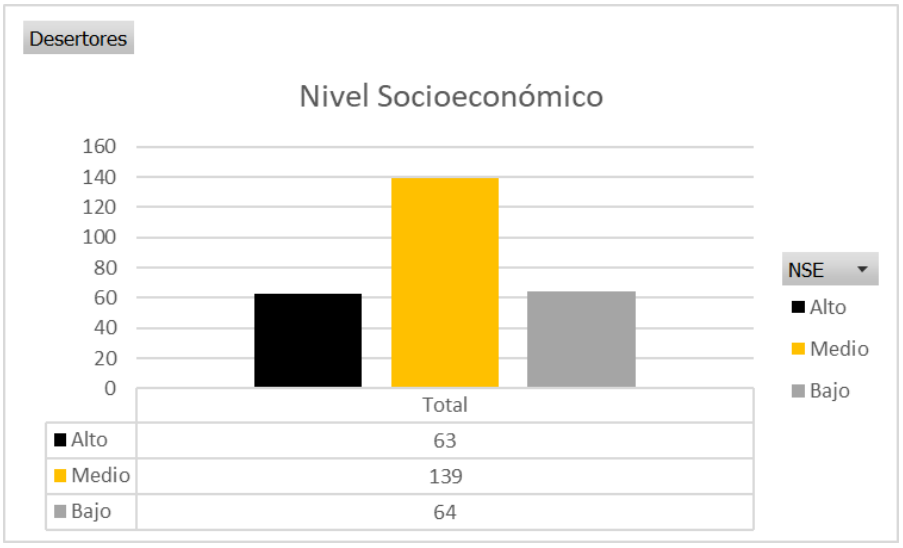
Gráfica 3-2 Distribución de género, desertores del campus Ensenada del 2008 al 2018.

La edad más frecuente de los desertores es de 18 años, aunque el 17% de los casos eran mayores de 21 años cuando ingresaron a la carrera, sin embargo, se puede notar en la gráfica 3-3 que aunque hay algunos casos aislados, la escuela de Administración y Negocios es la que muestra más estudiantes desertores mayores de 20 años particularmente en el programa de LNI.



Gráfica 3-3 Distribución de desertores por edad. Fuente: Elaboración propia.

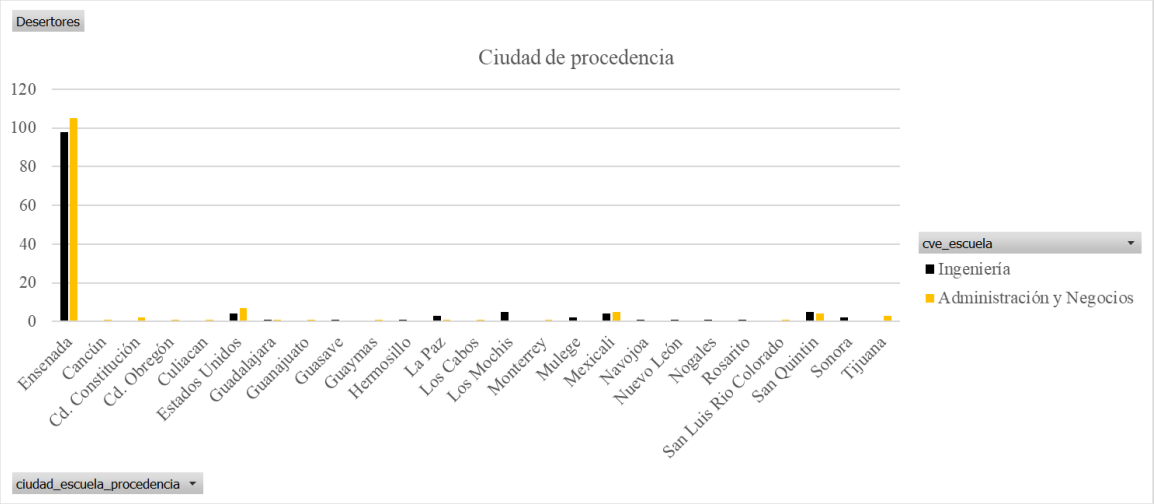
Los desertores pertenecen a diferentes estratos socioeconómicos, de acuerdo a los criterios mencionados en la Figura 3-4, el puntaje para determinar este Nivel Socio Económico (NSE) se asigna con el siguiente criterio: si el estudiante tiene una mayor necesidad económica va a obtener mayor número de puntos, se puede inferir que, si un estudiante tiene un NSE con un mayor puntaje, tiene una mayor necesidad de conservar o incrementar el apoyo financiero con el que entra a la Institución. En la Figura 3-4 se muestra que el puntaje más bajo es de 5 puntos y el más alto de 71 puntos. Para identificar el NSE de los desertores de este estudio se discretizó este atributo en cuatro intervalos $\alpha = (X\max[\![- X\_\min]\!])/k$, en donde k son tres categorías (alta, media y baja), se puede observar en la gráfica 3-4 que la mayoría de los estudiantes se encuentran entre el nivel socioeconómico alto y medio, siendo más numeroso el de los desertores con un NSE medio.



Gráfica 3-4 Nivel socioeconómico del total de desertores. Fuente: Elaboración propia.

Como se observa en la gráfica 3-5 la mayoría de los desertores son locales, aunque hay algunos que proceden de ciudades de fuera del estado. Los desertores tanto para la escuela de Ingeniería como para la de Administración y Negocios están concentrados principalmente en la ciudad de Ensenada, Baja California, lo cual puede estar relacionado con el que la mayoría de

estos desertores provienen de escuelas locales, sin embargo también hay estudiantes foráneos que desertaron.

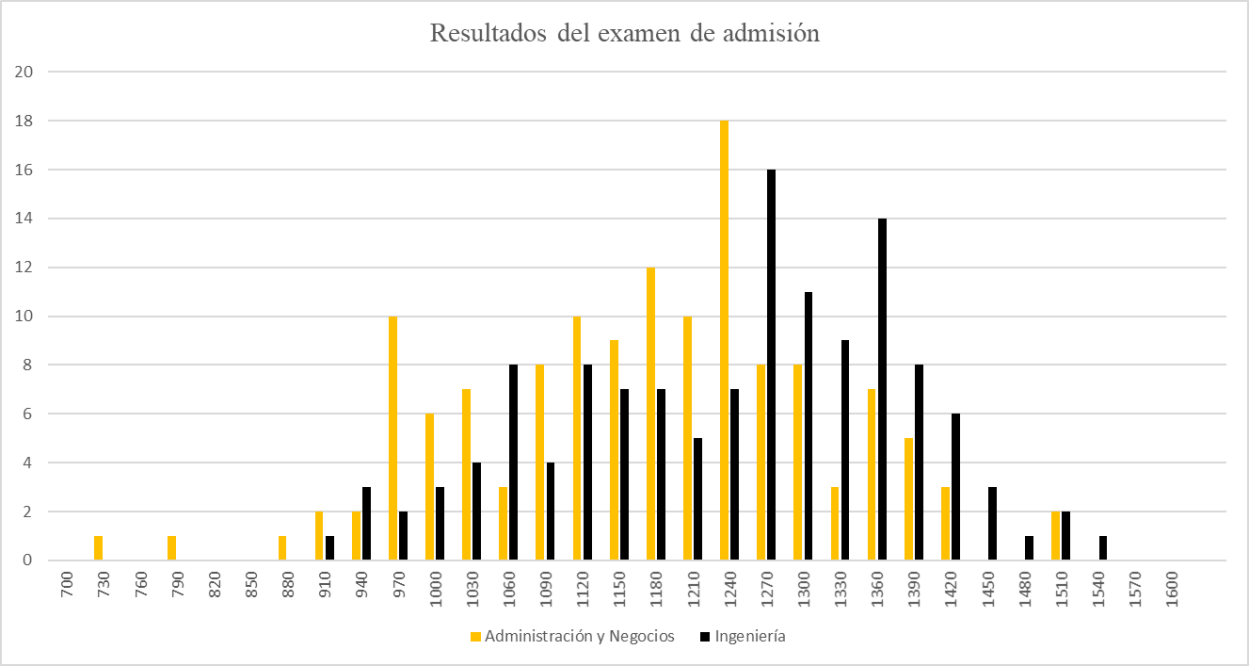


Gráfica 3-5 Ciudad de procedencia (Elaboración propia)

En cuanto al contexto académico, es importante identificar algunas variables que determinan algunos apoyos financieros otorgados al estudiante, así como los compromisos curriculares y extracurriculares que asume al entrar a la Institución, ya que muchos de estos estudiantes traen un currículo deportivo que les da acceso a una beca deportiva y otros cuentan con un historial académico que los hace acreedores a apoyos de excelencia académica; todos los apoyos financieros otorgados tienen criterios de reasignación, tales como el promedio que el estudiante debe mantener y la integridad académica que debe observar en el cumplimiento del reglamento estudiantil para no ser sancionados y correr el riesgo de perder una parte o la totalidad del apoyo.

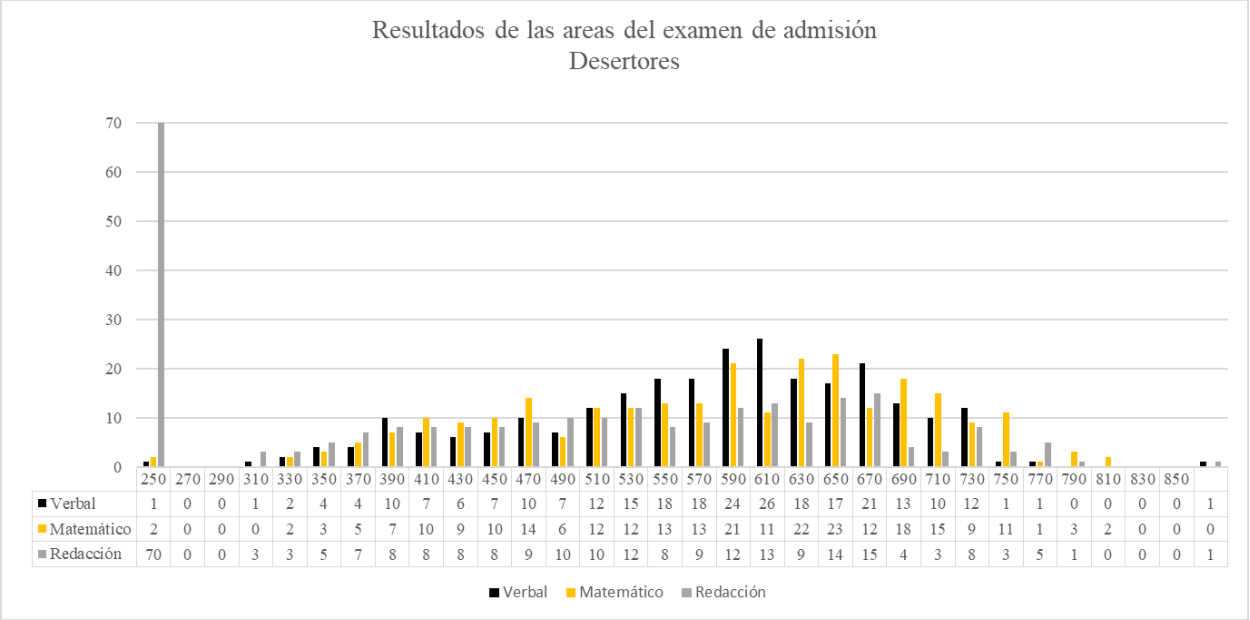
Otro atributo significativo son los resultados del examen de admisión, los estudiantes nacionales aplican la Prueba de Aptitud Académica (PAA), del College Board y los aspirantes extranjeros provenientes de los Estados Unidos aplican la prueba estandarizada Scholastic Aptitude Test, SAT, del College Board. Este examen, forma parte de los factores que inciden en el porcentaje de apoyo financiero recibido, así como la condición con la que ingresa el estudiante a la IES, ya que, junto con el puntaje obtenido en este, indica si es un estudiante condicionado o no, de acuerdo al reglamento referido en la sección de admisiones de este capítulo. Se puede observar en la gráfica 3-6 que la mayoría de los desertores de

Ingeniería obtuvieron puntajes por encima de los 1200 puntos requeridos, para acceder a algún apoyo financiero; por el contrario, se puede ver que en la escuela de Administración y Negocios hay una gran concentración de desertores por debajo de este puntaje.



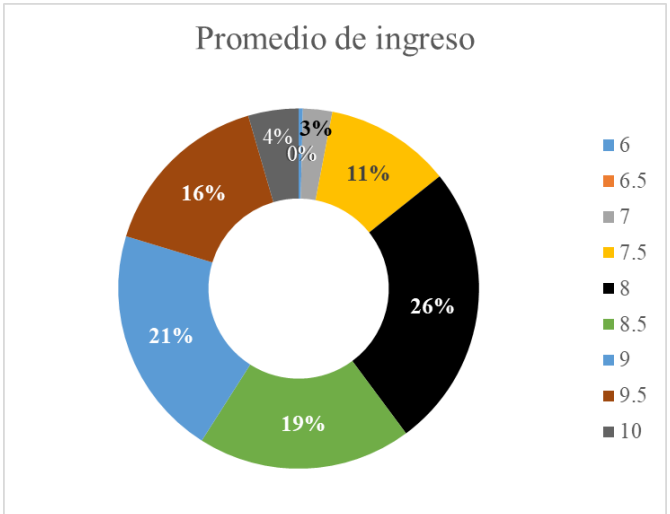
Gráfica 3-6 Resultados del Prueba de Aptitud Académica (PAA), del College Board

El área del conocimiento evaluada en el examen de admisión con resultados más bajos es el de redacción como se indica en la gráfica 3-7, en el área que mide tanto el razonamiento matemático como el aprovechamiento del estudiante en las áreas de aritmética, álgebra, geometría y análisis de datos y probabilidad, así como el verbal que mide el nivel de desarrollo de la habilidad verbal del estudiante, esto es, su capacidad para utilizar el lenguaje verbal para la comprensión e interpretación de la lectura (Board, 2014) los resultados están más concentrados entre los 500 y 650 puntos, con una desviación estándar de 100 puntos.

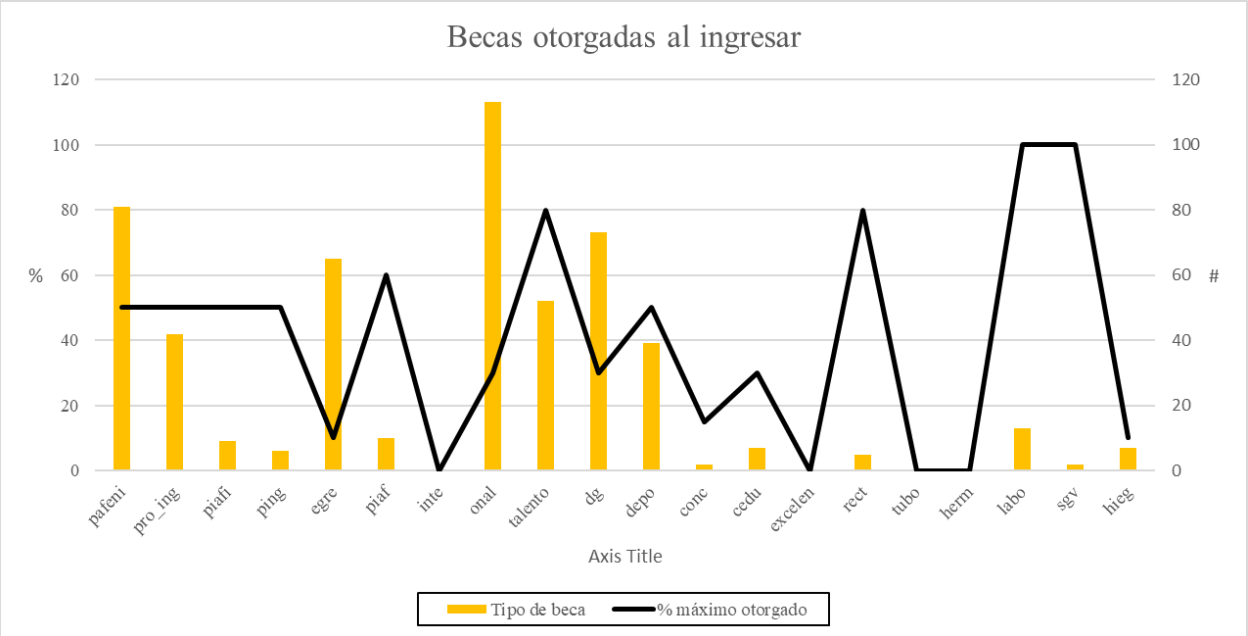


Gráfica 3-7 Resultados de las áreas de conocimiento de la Prueba de Aptitud Académica (PAA), del College Board

El 60% de los desertores ingresó con un promedio mayor a 8.5 (gráfica 3-8), siendo este un factor para acceder a un tipo de apoyo financiero según el reglamento; y eso coloca al 40% del resto de los desertores en una condición de no apoyo o incluso de ser requerido en el programa de nivelación de los alumnos condicionados, que como se mencionó, deben asistir a una serie de asesorías, cursos y talleres adicionales a su carga académica durante un año.



. Gráfica 3-8 Promedio general del nivel académico anterior antes de entrar a la Universidad



Gráfica 3-9 Tipos de apoyos financieros asignados a los desertores al entrar a la universidad

El contexto institucional al momento de ingresar a la IES, representa los atributos que caracterizan a los estudiantes cuyo valor es definido por la institución, entre los cuales se encuentran el estatus de ingreso, es decir, si este es un alumno condicionado, aprobado o sin examen, si obtuvo una o más becas, que tipo de becas fueron asignadas y con ello las condiciones que este debe mantener durante su estancia en la IES, se aprecia en la gráfica 3-9 que la mayoría de las becas otorgadas son aquellas definidas por la campaña promociona ONAL que les otorga un porcentaje de descuento del 20% o menos, la beca PANEFI es una beca otorgada a todos los aspirantes que se hayan destacado en su nivel académico anterior, cabe hacer mención que este apoyo solo lo otorga campus Ensenada, la beca de dirección general (DG) es otorgada por un año y está condicionada a mantener un promedio general mayor de 8.5 durante los semestres correspondientes, también se puede observar en esta gráfica que hay una cantidad importante de estudiantes egresados de la preparatoria del mismo Sistema CETYS, ya que la beca egresado (EGRE) solo se otorga a los estudiantes que egresan de un programa escolarizado de preparatoria, licenciatura o posgrado, el monto de esta beca es del 10% y no se pierde en ningún momento de la vida académica del estudiante.

3.5.3 Durante su estancia en la IES

En las Tablas 3-12 y 3-13 se muestran los atributos de los contextos académicos e institucionales respectivamente, que definen la experiencia de los alumnos desertores durante su

estancia en CETYS campus Ensenada, en estos contextos se pudieron rescatar algunas variables cualitativas que representan la experiencia del estudiante durante su estancia en la Institución, en este capítulo se explicó cómo está diseñado y se aplica la evaluación del nivel de satisfacción del estudiante, instrumento muy relevante para la Institución, ya que algunos de estos resultados que representan áreas de mejora, se incluyen en la planeación y estrategias a desarrollar para los siguientes períodos escolares, esta evaluación es semestral y aunque no se cuenta con los resultados individuales si se conoce de que programa académico son, incluso a que semestre pertenecen los estudiantes que realizaron la encuesta, es por ello que se usaron estos datos para inferir que la evaluación refleja la percepción del desertor si pertenece a ese grupo.

Tabla 3-12 Atributos académicos al desertar de la IES

Momentoo Momentoto	**Contexto**	**Variable**	**Descripción**	**Tipo**	**Fuente**
Durante su estancia en la IES	Académico	Promedio final	Promedio global al momento de su salida	numérico	Escolar
		TotalFaltas	Total de faltas acumuladas en el período anterior a su baja definitiva	numérico	Escolar
		PromedioPeriodoSeleccionado	Promedio del último período cursado	numérico	Escolar
		Materiascursadas	Total de materias cursadas	numérico	Escolar
		TotalReprobadas	Número total de materias reprobadas	numérico	Escolar
		Reprobadasarea1	Número total de reprobadas eje de formación general	numérico	Escolar
		Reprobadasarea2	Número total de reprobadas eje de formación de colegio básicas: Ingeniería o Administración y Negocios	numérico	Escolar
		Reprobadasarea3	Número total de reprobadas eje de formación profesional	numérico	Escolar
		TotalAprobadas	Número total de materias aprobadas	numérico	Escolar
		Extraordinarios	Número total de extraordinarios presentados	numérico	Escolar
		ultimosemestre	Semestre en el que decide su baja	numérico	Escolar

		Avance	Porcentaje de avance del programa académico	numérico	Escolar

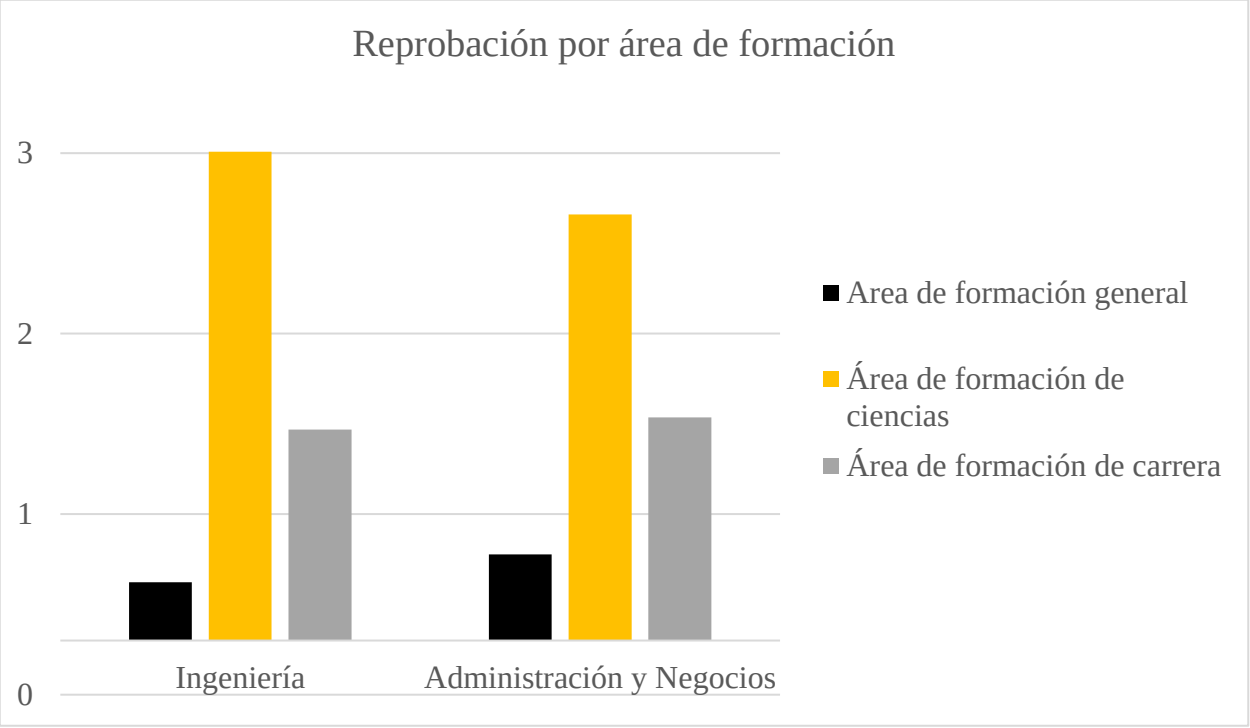
Tabla 3-13 Atributos institucionales al desertar de la IES

Momento	Contexto	Variable	Descripción	Tipo	Fuente
	Institucional	Becapromedio	Monto total del apoyo financiero asignado a la fecha de su baja	numérico	Finanzas
		VPrestamoEquipo VDepartamentoescolar VPagoencaja VCreditoeducativo VBiblioteca VCafeteria VPapeleria VEnfermeria VServiciosocial VProgramasinternacionales VTrmitesdeprcticasprofesionales VConsultacalificacioneseninternet VCampaadereinas	Evaluación de las áreas de servicio de la institución (0=Muy mal, 1=No aplica, 2=No tengo bases para opinar, 3=Regular, 4=Bien, 5=Excelente	Categórico	Evaluación de satisfacción
		VLaboratoriodecmputo VMaestrosenasesorasacadmicas VMaestrosfueradeclases VEnseanzaaprendizajes VLaboratoriosdecmputo VLaboratoriosdeingeniera VSalasaudiovisuales VSalndeclases VCoordinadordecarrera	Evaluación del equipamiento e infraestructura disponibles y servicio educativo las áreas de servicio de la institución (0=Muy mal, 1=No aplica, 2=No tengo bases para opinar, 3=Regular, 4=Bien, 5=Excelente	Categórico	Evaluación de satisfacción

		VEquiposrepresentativos VEventosacadmicos VExposiciones VSemanacultural VSociedaddealumnos VTalleresculturales VTorneosinternos VViajesdeestudio VConferencias VAmbienteestudiantil VDeportes VDifusincultural VAsuntosestudiantiles	Evaluación de las actividades cocurriculaes y extracurriculares (0=Muy mal, 1=No aplica, 2=No tengo bases para opinar, 3=Regular, 4=Bien, 5=Excelente	Categórico	Evaluación de satisfacción
		VOrgullo Calificacion VSatisfaccion VComparativootras escuelas	Evaluación de la Institución, satisfacción y orgullo de pertenencia (0=Muy mal, 1=No aplica, 2=No tengo bases para opinar, 3=Regular, 4=Bien, 5=Excelente	Categórico	Evaluación de satisfacción
		VBlackboard VRedinalmbrica VConsultaelectronicaenbiblioteca VAreasdeportivas VAreasverdesyplazas VAuladevideoconferencias VBaos VBa¤os VCafeteria VCafeDVolada VEstacionamiento	Evaluación de la infraestructura de la institución (0=Muy mal, 1=No aplica, 2=No tengo bases para opinar, 3=Regular, 4=Bien, 5=Excelente	Categórico	Evaluación de satisfacción
		MotivoBaja	Motivo de la baja	nominal	CEDE

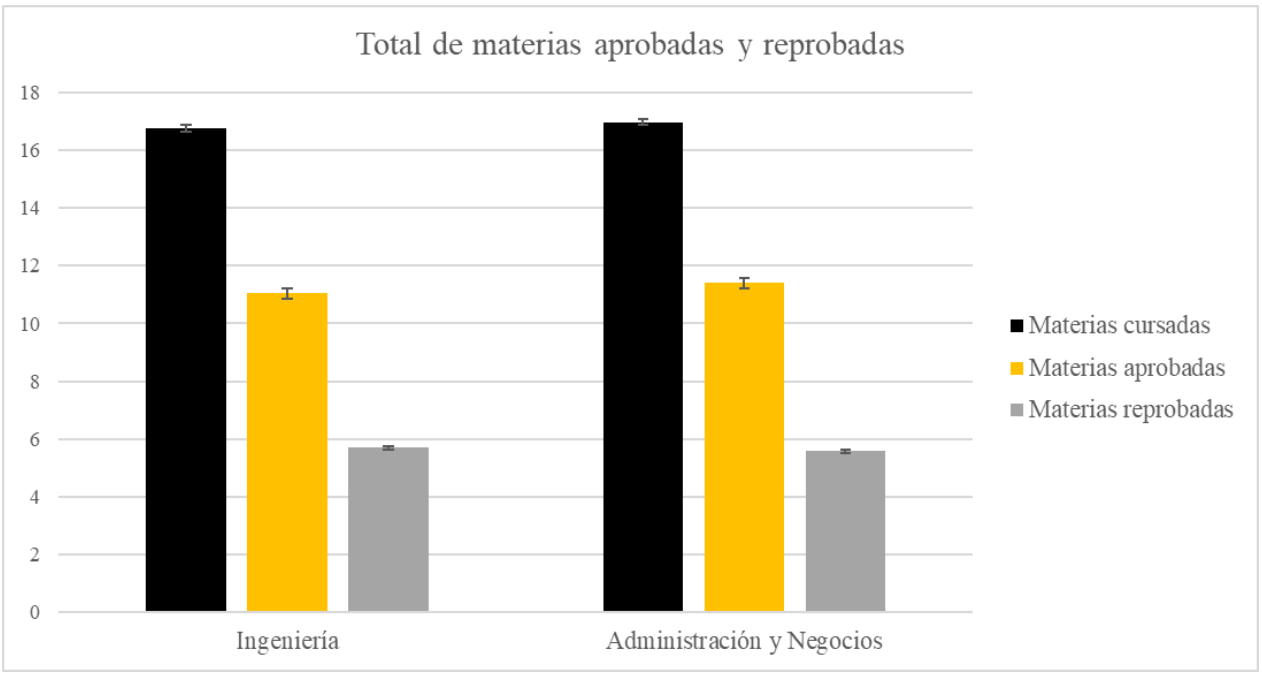
El promedio de reprobación y aprobación de los estudiantes es muy semejante entre las dos escuelas, como se puede apreciar en la gráfica 3-11, un dato interesante es que el número de materias aprobadas es mayor que el de las reprobadas, aunque la reprobación de estos desertores es aproximadamente del 33%. En la gráfica 3-10 se visualizan las áreas de reprobación con mayor

concentración, como se describió en la Tabla 3-12 el área de reprobación 1 representa a las materias que desarrollan las competencias generales o de formación integral, el área de reprobación 2 son aquellas que trabajan en las competencias de formación básica de colegio, las de ciencias de la ingeniería y las de ciencias administrativas según sea el caso, y finalmente el área de reprobación 3 son aquellas materias que desarrollan las competencias propias de la especialidad de carrera, siendo el área de formación de colegio la más representativa y enseguida en orden ascendente la de especialidad, incluso en proporción se observa el mismo comportamiento tanto para la escuela de Ingeniería como la escuela de Administración y Negocios.

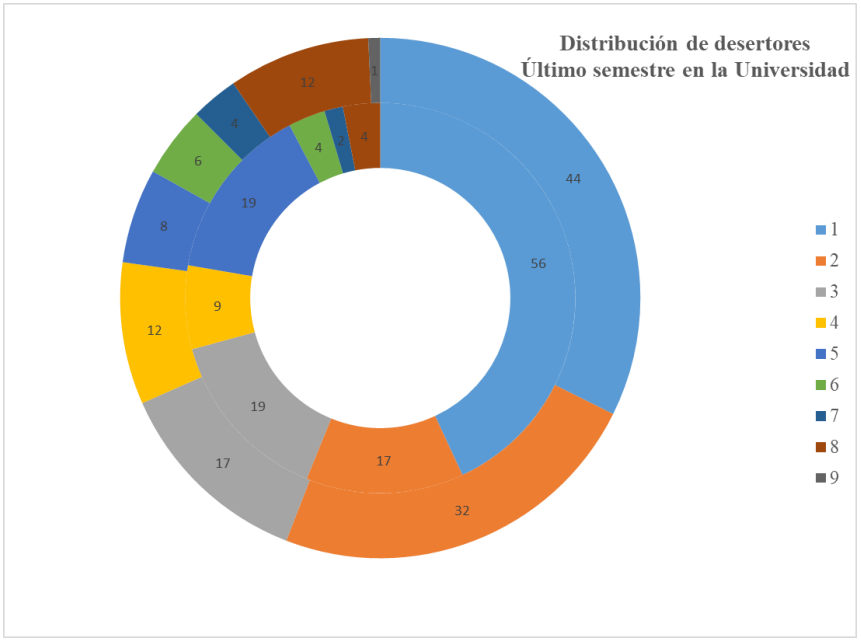


Gráfica 3-10 Reprobación por área de formación, por escuela. Fuente: Elaboración propia.

Los momentos descritos en los contextos históricos de los desertores de este caso de estudio, permite caracterizar el fenómeno, poniendo mayor énfasis en las variables que se identificaron como significativas para la construcción de modelos predictivos, mismos que se mostrarán en el capítulo de análisis y discusión. Al momento de salir de la IES los estudiantes se encontraban cursando el primer año en su mayoría como se puede observar en la gráfica 3-12, lo cual indica que estos estudiantes definen su deserción en los primeros dos años de estudios, siendo mucho más significativo el primero.



Gráfica 3-11 Promedio de materias cursadas, aprobadas y reprobadas. Fuente: Elaboración propia.



Gráfica 3-12 Último semestre en la Universidad

3.6 Algoritmos de aprendizaje automático

Como se mencionó, las tareas claves de la minería de datos son la clasificación, la regresión y el agrupamiento, su esencia es generar de forma inductiva a partir de los datos (se le conoce como entrenamiento) un modelo (se le conoce como conocimiento) que puede aplicarse a nuevos datos (predictivo) (Gironés-Roig et al., 2017). Los algoritmos de minería de datos se dividen en supervisados y no supervisados, en esta investigación se aplicaron los primeros ya que los datos que se recabaron están clasificados en desertores y no desertores por su motivo de baja, por lo que conocemos los atributos de unos y de otros. El uso de estos algoritmos supervisados, permitió crear dos modelos predictivos construidos, uno por medio del algoritmo de regresión logística y el otro con la técnica *Random Forest*, mismos que se expondrán en este capítulo, con la finalidad de evaluarlos y medir su nivel de eficiencia y precisión.

3.6.1 Algoritmos de aprendizaje supervisados

Se aplicaron algoritmos lineales como la regresión logística para poder describir el fenómeno de la deserción en el campus Ensenada, y uno de clasificación como los árboles de decisión para construir un modelo predictivo del mismo; a partir de la decisión de baja indicada en el motivo de baja, aunque no todos los casos tienen la causa de baja registrada porque el estudiante puede simplemente no volver a inscribirse sin cerrar este proceso, cuando él desea cambiarse de campus tiene que definir la razón de baja como "cambio de campus", ya que para poder migrar su documentación y sus apoyos financieros es necesario hacer este cierre de proceso, es por eso que se hizo una clasificación binaria para identificar a los que desertaron del Sistema CETYS de los que no.

La decisión de utilizar los algoritmos de regresión logística (RL) fue por la condición categórica dicotómica de la variable dependiente utilizada, ya que este algoritmo indica la tasa de variación entre cero y uno de una respuesta (Ferri Ramirez y Ramirez Quintana, 2011), mostrándonos la probabilidad de que ocurra el hecho $Pr\ (y=1|\ x)$ en función de algunas variables influyentes o significativas , por lo que con este algoritmo construimos una función logística para clasificar a las dos poblaciones: los que desertan del Sistema CETYS (1) y los que se dan de baja del campus pero permanecen en el Sistema (1) . La RL también nos indica la fuerza de asociación que tienen esas variables independientes $X1,X2,..,Xn$, es decir sus coeficientes representado por $\beta1,\ \beta2,\ldots,\beta n$, los cuales se estiman utilizando el método de máxima verosimilitud (MLE,

Maximum Likelihood Estimation, por sus siglas en inglés) (Nwanganga, Fred Chapple, 2020) ya que este selecciona los coeficientes que hacen más probable que el valor de respuesta ocurra siendo este entre 0 y 1, $\beta 0$ es el término independiente o constante del modelo, esta función se representa $P(y = 1|x) = \frac{\exp(b0+\sum_{i=1}^{n} bixi)}{1+\exp(b0+\sum_{i=1}^{n} bixi)}$ para convertir esta en una expresión matemática más sencilla de manejar, se divide entre su complemento que es la probabilidad de que el evento no se dé, es decir, $\frac{P(y=1|x)}{1-P(y=1|x)}$, a estos complementos se les llama odds, finalmente se hace una transformación con el logaritmo natural, para obtener una ecuación lineal, llamada *logit*: $\log(\frac{P(y=1|x)}{1-P(y=1|x)}) = \beta 0+\sum_{i=1}^{n} bixi.$

Como la variable de respuesta no es continua sino categórica, con un valor finito de valores se usó árboles de clasificación (Gironés-Roig et al., 2017). Sin embargo, como las variables explicativas son semejantes entre una clase (desertor) y otra (no desertor) y con la intención de reducir el sesgo, la varianza o que el modelo sea incapaz de predecir correctamente los datos, ya que hay desbalanceo de clases en el set de datos ya que el 75% está representando a la clase desertor y el 25% a la que migró a otro campus, se utilizó un método de ensemble ya que estos combinan múltiples modelos (pequeños arboles individuales) en uno nuevo consiguiendo así mejores predicciones que cualquiera de los modelos individuales. El clasificado combinado más usado es el *Random Forest* (Louppe, 2014), que como su nombre lo indica, construye un bosque de árboles conformados de forma aleatoria por la técnica de bagging; que consiste en generar versiones diferentes de entrenamiento usando el muestreo con remplazo y al final define un solo árbol con la media de las predicciones de todos los árboles que lo forman Figura 3-8. Este algoritmo evita la correlación que pudiera existir entre los arboles ya que, al hacer una selección aleatoria de los n predictores antes de evaluar cada división, de esta forma permitirá que otros predictores puedan ser seleccionados

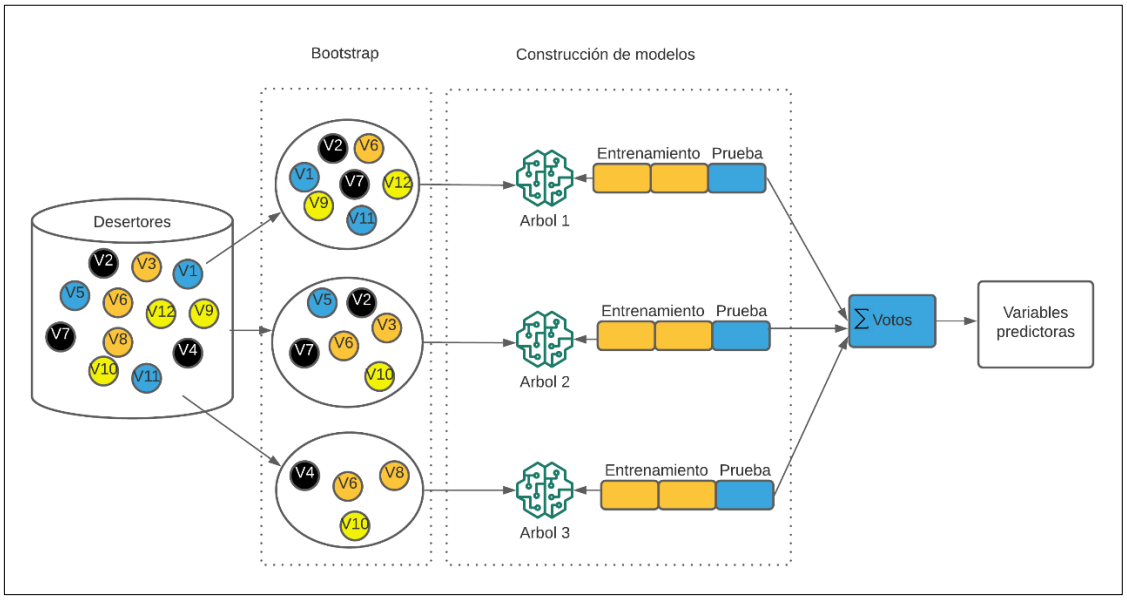


Figura 3-8 Algoritmo Random Forest (imagen modificada de (Orellana Alvear, n.d.)

3.6.2 Análisis estadístico y minería de datos

Las variables independientes utilizadas se analizaron de acuerdo a cada contexto y momento vivido por el estudiante hasta el momento de decidir darse de baja del campus, estos están representados en el modelo de la Figura 3-2 y descritas en las tablas 3-9, 3-10, 3-11, 3-12 y 3-13 las cuales pueden tener un valor cuantitativo, cualitativo o categórico, con la intención de predecir los valores de la variable dependiente, si deserta o permanece en el Sistema CETYS.

El total de instancias en el set de datos fueron 355 observaciones, cada una representando a un desertor con 102 características, es decir, 102 variables independientes que podrían predecir el fenómeno, como se mencionó algunas de ellas discretas, continuas y categóricas.

El set de datos tiene una alta dimensionalidad por el detalle de la información que se obtuvo, con una alta densidad ya que en el proceso de ETL se hizo una imputación de los únicos datos ausentes, como son algunos resultados de las áreas de verbal y matemático del examen de admisión, que como se explicó en el capítulo anterior, se obtuvieron por medio de la regresión lineal, con respecto a los datos personales del desertor no se cuenta con mucha resolución, ya que el sistema de promoción y de escolar de la Institución no guarda mucha información que se recaba cuando el estudiante quiere ingresar y solicita algún apoyo financiero.

Como parte del aprendizaje automático (machine learning), se creó un set de entrenamiento por medio de un proceso de muestreo, que consistió en dividir el set de datos (log.reg) en un 75%

como set de entrenamiento (log.train) y un 25% como set de pruebas (log.test), este proceso aleatorio garantiza que la distribución de las clases entre el set de entrenamiento y el de pruebas sean semejantes para garantizar que el modelo construido a partir del primer set sea igual a uno creado de todo el set de datos original, en la 3-14 se puede observar la misma distribución de clases entre los set de datos.

Tabla 3-14. Distribución de clases en los sets de datos

Set de datos	% Motivo de baja (0)	% Motivo de baja (1)
log.reg	25.07	74.93
log.train	25.09	74.91
log.test	25	75

3.7 Modelos predictivos

Por el principio de parsimonia se busca un modelo que pueda predecir la deserción con el menor número de variables (implica menores errores estándar) pero con un buen desempeño.

Se utilizaron dos herramientas para la creación y evaluación de modelos: RStudio e IBM SPSS Modeler, este último para evaluar todas las variables independientes en los diferentes contextos y momentos de la estancia en el campus del estudiante que desertó, con el fin de identificar las variables que mejor explicaban y predecían el fenómeno, así como los algoritmos que generaban un mejor modelo.

Modelos de regresión logística

Los modelos de regresión logística se construyeron por medio del método de eliminación hacia atrás (Backward Stepwise Regression) en donde se introducen todas las variables propuestas y se van excluyendo en cada iteración una por una, eliminando la menos influyente, para cada modelo. Se expondrán todos los modelos en base al momento que vivió el desertor y su contexto, así como la explicación y evaluación de cada uno de estos.

a) Al ingresar a la IES: contexto personal

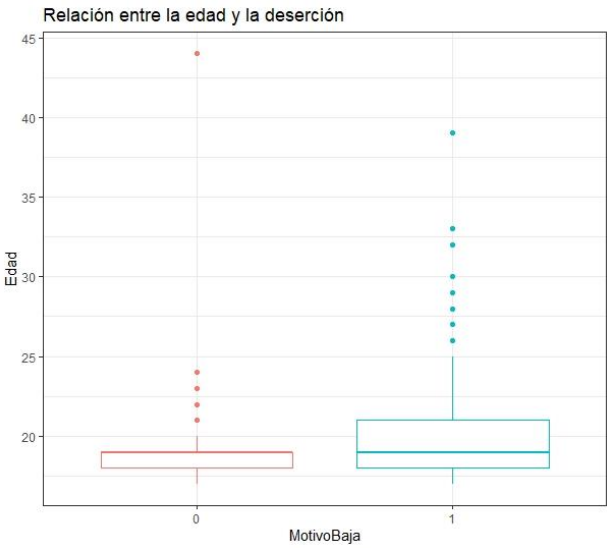


Figura 3-10 Edad y deserción

```
Call:
glm(formula = MotivoBaja ~ edad + VExperienciacInstitucion +
    NSE, family = binomial, data = log.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1456  -1.1340   0.6048   0.8237   1.2214

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -2.93050    1.62763  -1.800  0.07179 .
edad                      0.24413    0.08382   2.913  0.00358 **
VExperienciacInstitucion1 -0.52121    0.29733  -1.753  0.07961 .
NSE2                      -0.80154    0.38795  -2.066  0.03882 *
NSE3                      -0.03377    0.50993  -0.066  0.94720
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 299.67  on 264  degrees of freedom
Residual deviance: 276.98  on 260  degrees of freedom
AIC: 286.98

Number of Fisher Scoring iterations: 5
```

Figura 3-9 modelo.reg.log.d0: contexto personal al ingresar

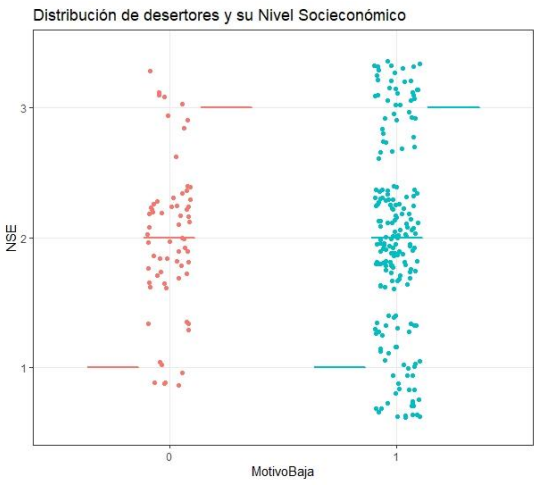


Figura 3-12 Distribución de desertores y su Nivel Socioeconómico

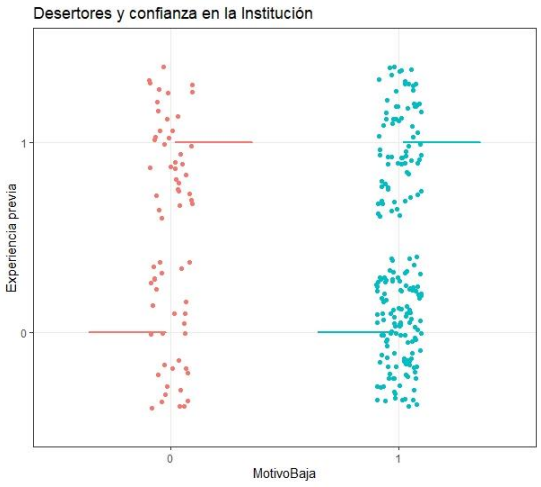


Figura 3-11 Distribución de desertores con experiencia previa en la Institución

La edad es una variable significativa para este modelo, por cada unidad de incremento en este valor, la probabilidad de que deserte aumenta por un factor de 1.2765 (es igual a exponenciar el log-odds de 0.24413 mostrado en la Figura 3-9), este supuesto se puede confirmar en la Figura, 3-10 en donde se puede observar una concentración de desertores con mayor edad, también se

puede ver que la variable independiente Nivel socioeconómico mostrada en la Figura 3-12 tiene una concentración importante en el nivel 2, el cual representa el NSE medio, en cuanto a la variable independiente Experiencia Institucional, la Figura 3-11 muestra que aquellos que no tenían ninguna experiencia previa (0), por no ser hijo de egresado o haber cursado su bachillerato en la Institución, es mucho mayor que los que si contaban con esa experiencia (1).

b) Al ingresar a la IES: contexto académico

```
Call:
glm(formula = MotivoBaja ~ promedio_ingreso + tipo_escuela +
    VExcelenciaacademica + VDeportista, family = binomial, data = log.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1984  -0.9303   0.6316   0.7804   1.1631

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)           6.6187     1.9598   3.377 0.000732 ***
promedio_ingreso     -0.5883     0.2294  -2.564 0.010344 *
tipo_escuela2        -0.8132     0.3594  -2.263 0.023643 *
VExcelenciaacademica1  0.9992     0.4507   2.217 0.026601 *
VDeportista1         -0.6534     0.3894  -1.678 0.093396 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 299.67  on 264  degrees of freedom
Residual deviance: 281.66  on 260  degrees of freedom
AIC: 291.66

Number of Fisher Scoring iterations: 4
```

Figura 3-13 modelo.reg.log.d1: contexto académico al ingresar

El modelo mostrado en la Figura 3-13 indica que las variables independientes *promedio de ingreso* es significativa para identificar a un desertor, aunque la información mostrada en la Figura 3-14 presenta a esta variable con una dispersión desde el siete de promedio lo cual implica que ingresó de forma condicionada de acuerdo al reglamento de selectividad y también se puede observar algunos promedios altos, es por ello que la variable *excelencia académica* (Figura 3-16) también se muestra como significativa en este modelo, ya que hay una concentración importante de desertores con promedios entre ocho y nueve, en cuanto al el *tipo de escuela de procedencia*: privada o pública, la Figura 3-15 presenta una concentración importante en los desertores que vienen de escuelas privadas, la Figura 3-17 muestra que hay más desertores con antecedentes *deportivos* aunque hay una gran concentración sin historial deportivo.

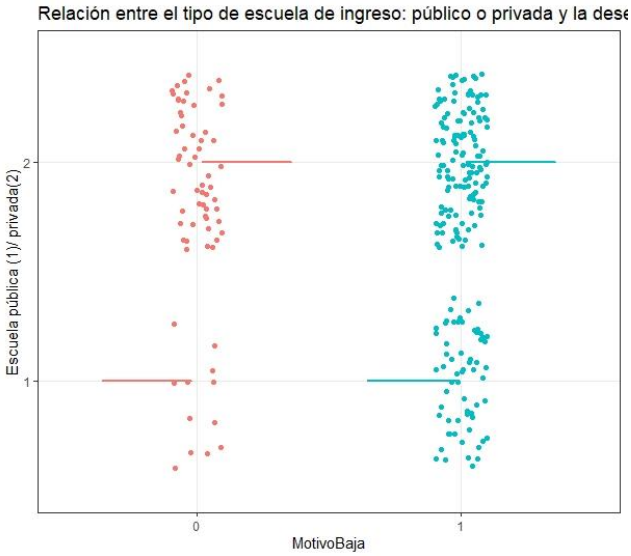


Figura 3-17 Relación de desertores y el tipo de escuela de ingreso

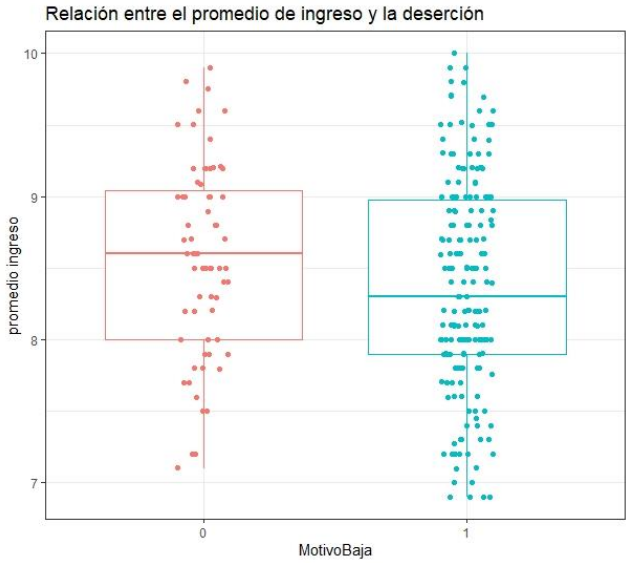


Figura 3-14 Relación de desertores y y su promedio de ingreso

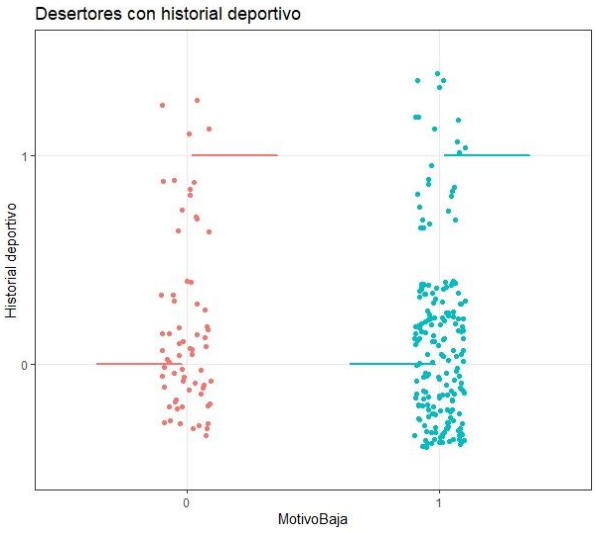


Figura 3-16 Desertores con historial deportivo

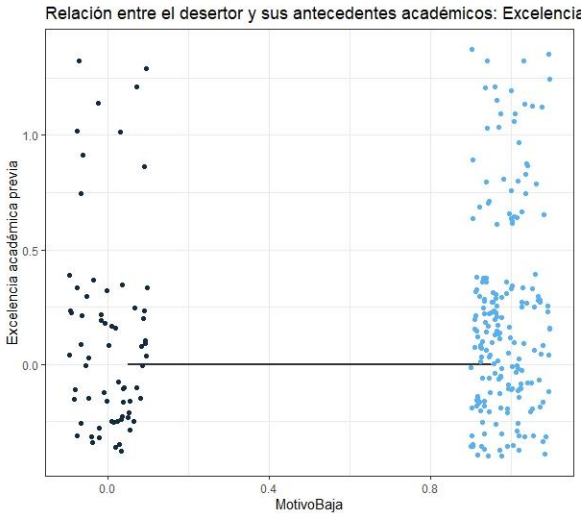


Figura 3-15 Desertores con historial de excelencia académica

c) Al ingresar a la IES: contexto institucional

En el modelo.reg.log.d2 se identifican los apoyos financieros más significativos (Figura 3-18), las becas *piafi*, *piaf* son becas que se otorgan a estudiantes que ingresan con aptitudes

intelectuales altas y una buena trayectoria académica en Bachillerato, la probabilidad de que un estudiante deserte incrementa dado un factor de 0.91 por cada unidad que disminuya la beca *piafi* y .93 por cada unidad de la beca *piaf*, la proporción que se muestra en la Figura 3-20, se puede identificar que no son muchos los desertores que tienen estas credenciales a su ingreso, las figuras 3-19 y 3-21 permiten comprobar lo ya identificado en el modelo.reg.d1, ya que no son muchos desertores que no tienen experiencia previa con la Institución, es decir, no cuentan con *beca de egresado* y si hay desertores con algún porcentaje de *beca deportiva*, respectivamente.

```
Call:
glm(formula = MotivoBaja ~ becaingreso + pafeni + pro_ing + piafi +
    piaf + egre + talento + depo + labo, family = binomial, data = log.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3500  -0.2371   0.5281   0.7738   1.8373

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.05255    0.33181   3.172 0.001513 **
becaingreso  0.05478    0.02055   2.666 0.007677 **
pafeni      -0.04524    0.02441  -1.854 0.063783 .
pro_ing     -0.05975    0.02388  -2.502 0.012345 *
piafi       -0.09387    0.02403  -3.907 9.35e-05 ***
piaf        -0.07537    0.02424  -3.109 0.001877 **
egre        -0.13210    0.03703  -3.567 0.000361 ***
talento     -0.04144    0.02200  -1.884 0.059600 .
depo        -0.06905    0.02629  -2.627 0.008613 **
labo        -0.04138    0.02217  -1.867 0.061925 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 299.67  on 264  degrees of freedom
Residual deviance: 249.58  on 255  degrees of freedom
AIC: 269.58

Number of Fisher Scoring iterations: 5
```

Figura 3-18 modelo.reg.log.d2 : contexto Institucional al ingresar

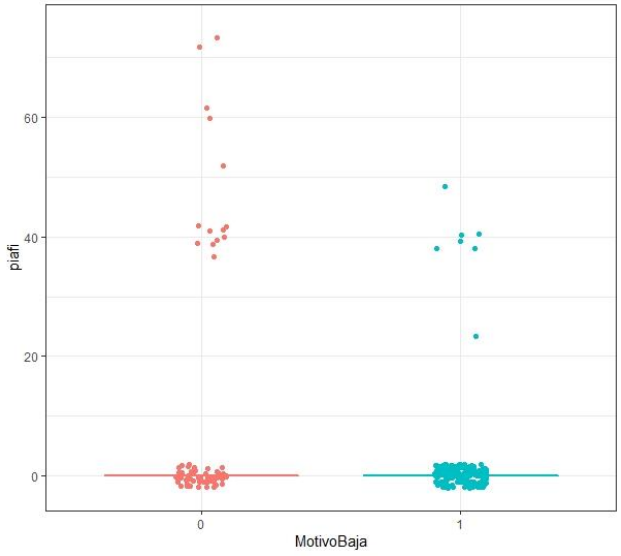


Figura 3-20 Porcentaje de beca a alumnos con talento y deserción

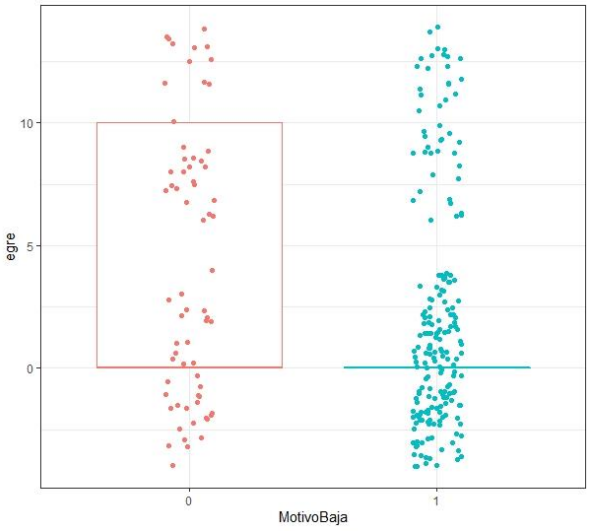


Figura 3-21 Porcentaje de beca a desertores egresados del bachillerato de CETYS

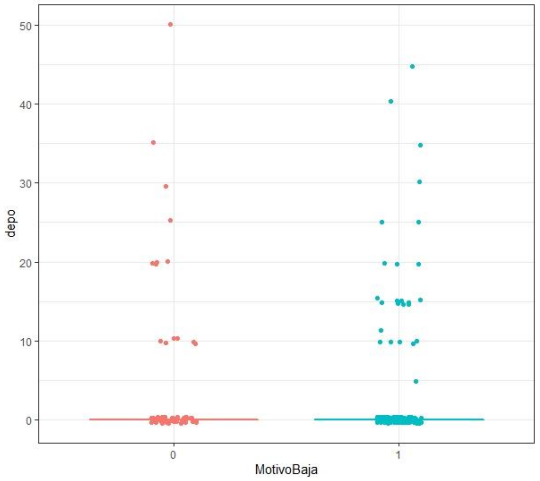


Figura 3-19 Porcentaje de beca deportiva a desertores con desempeño deportivo

d) Durante su estancia en la IES: contexto académico

Este modelo (Figura 3-22) muestra que los desertores que se van del Sistema CETYS son aquellos que tienen los promedios más bajos en su período de baja, aunque la probabilidad aumenta con un factor de .98 por cada unidad que disminuya esta calificación y los que tienen mayor reprobación en el área de formación general, son las materias que llevan todos los estudiantes de cualquier licenciatura en CETYS, que corresponden al área de humanidades. Figuras 3-23 y 3-24.

```
Call:
glm(formula = MotivoBaja ~ TotalFaltas + PromedioPeriodoSeleccionado +
    Reprobadasarea1, family = binomial, data = log.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4402  -1.3106   0.6381   0.8560   1.0309

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 1.676426   0.578189   2.899  0.00374 **
TotalFaltas                 0.003764   0.002797   1.346  0.17839
PromedioPeriodoSeleccionado -0.013854   0.006506  -2.129  0.03323 *
Reprobadasarea1             0.314991   0.170944   1.843  0.06538 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 299.67  on 264  degrees of freedom
Residual deviance: 282.22  on 261  degrees of freedom
AIC: 290.22

Number of Fisher Scoring iterations: 5
```

Figura 3-22 modelo.reg.log.d3: desempeño académico durante su estancia

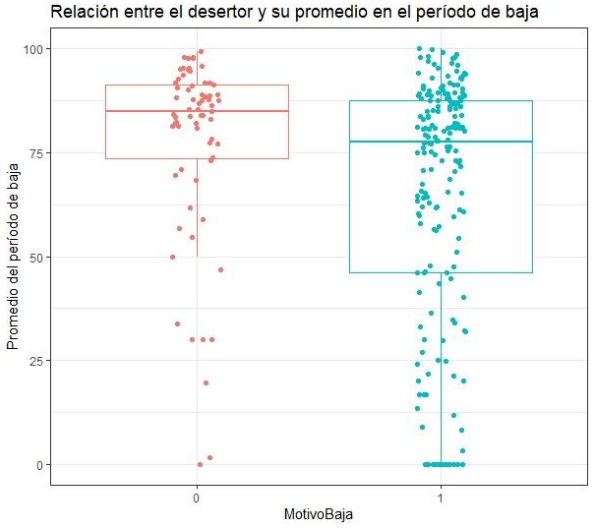


Figura 3-23 Desertores y el total de reprobación del área de formación general

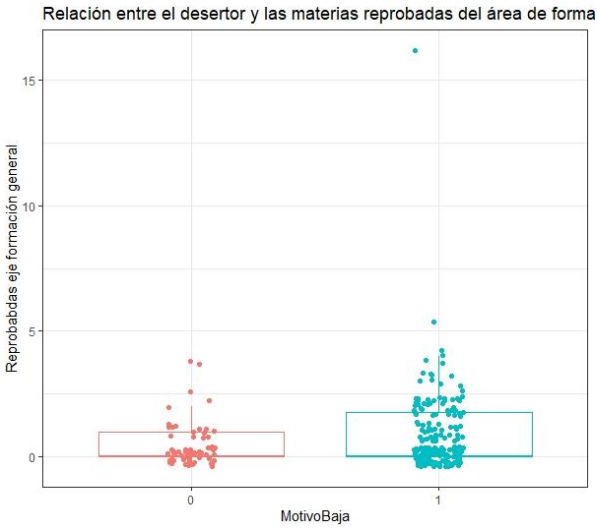


Figura 3-24 Desertores y su promedio en el periodo de baja

e) **Durante su estancia en la IES: contexto Institucional en el servicio**

Este modelo (Figura 3-25) muestra el comportamiento de las variables independientes de la evaluación de satisfacción en el área de servicios, identificando como servicio el *crédito educativo* (Figura 3-26), aunque este es otorgado a aquellos que de forma voluntaria lo solicitan y que además cumplen con los requisitos para su asignación, los desertores concentran su opinión en no tener bases para opinar (2) y bien (3) y en cuanto a la *beca promedio* (Figura 3-27) con la que el estudiante cuenta al momento de darse de baja es menor para los desertores que para los que se quedan en el Sistema CETYS, se puede observar que en ambos casos hay becas promedio del 100 %.

```
Call:
glm(formula = MotivoBaja ~ Becapromedio + VPagoencaja + VCreditoeducativo,
    family = binomial, data = log.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1896  -1.0043   0.5254   0.7809   1.5364

Coefficients: (1 not defined because of singularities)
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         2.419655   1.107835   2.184   0.0290 *
Becapromedio       -0.013046   0.006225  -2.096   0.0361 *
VPagoencaja4       -1.928173   1.126708  -1.711   0.0870 .
VPagoencaja5       -2.759249   1.239998  -2.225   0.0261 *
VCreditoeducativo2  1.810309   0.439361   4.120 3.78e-05 ***
VCreditoeducativo4  0.735545   0.387384   1.899   0.0576 .
VCreditoeducativo5        NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 299.67  on 264  degrees of freedom
Residual deviance: 275.60  on 259  degrees of freedom
AIC: 287.6

Number of Fisher Scoring iterations: 4
```

Figura 3-25 modelo.reg.log.d4: evaluación del servicio durante su estancia

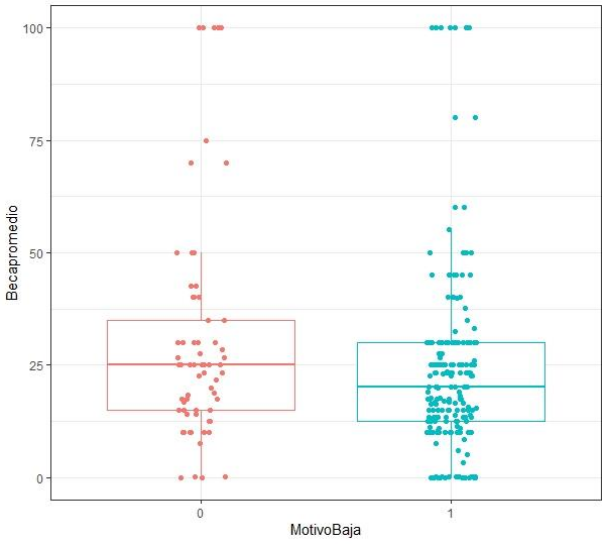


Figura 3-26 Porcentaje de beca total al momento de desertar

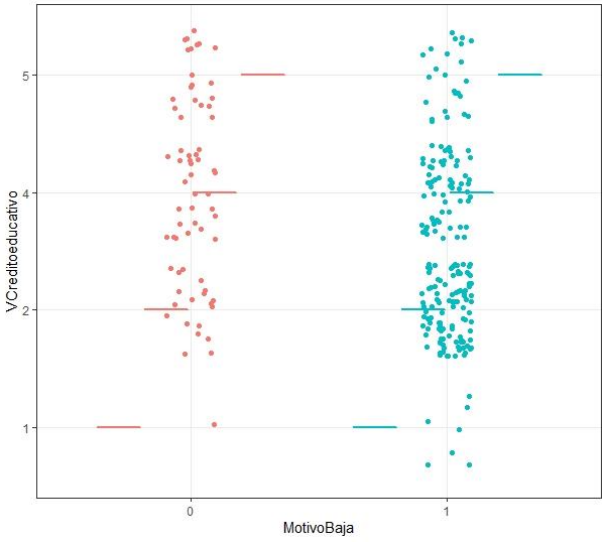


Figura 3-27 Evaluación del crédito educativo y la deserción

f) Durante su estancia en la IES: contexto Institucional en servicio educativo

El modelo no muestra de forma tan significativa la relación de estas variables independientes con la variable de respuesta (Figura 3-28), *los laboratorios de Ingeniería* son evaluados como regular (3) y bien (4), no se aprecia ningún valor de excelente (5), de acuerdo a la escala más frecuentemente señalada por los desertores, Figura 3-29.

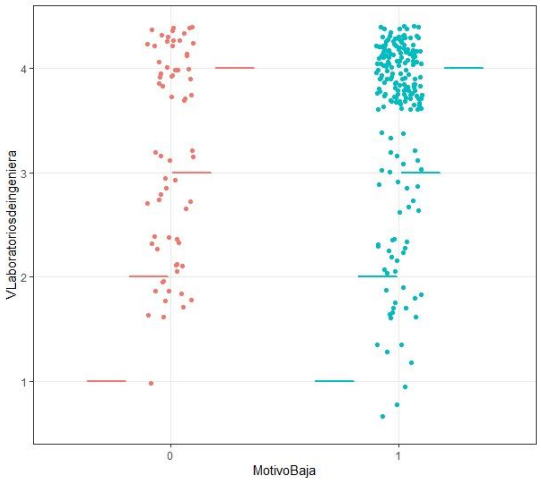


Figura 3-29 Evaluación de laboratorios de Ingeniería

```
Call:
glm(formula = MotivoBaja ~ VMaestrosenasesorasacadmicas + VLaboratoriosdeingeniera,
    family = binomial, data = log.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0393  -1.2557   0.5287   0.7337   1.1010

Coefficients: (2 not defined because of singularities)
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                      1.94591    1.06904   1.820   0.0687 .
VMaestrosenasesorasacadmicas2 -1.76359    1.11109  -1.587   0.1125
VMaestrosenasesorasacadmicas4 -0.77092    1.09780  -0.702   0.4825
VMaestrosenasesorasacadmicas5 -0.04879    1.11296  -0.044   0.9650
VLaboratoriosdeingeniera2           NA         NA      NA       NA
VLaboratoriosdeingeniera3     -0.79550    0.43804  -1.816   0.0694 .
VLaboratoriosdeingeniera4           NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 299.67  on 264  degrees of freedom
Residual deviance: 278.39  on 260  degrees of freedom
AIC: 288.39

Number of Fisher Scoring iterations: 4
```

Figura 3-28 modelo.reg.log.d5: evaluación del servicio educativo

g) **Durante su estancia en la IES: contexto Institucional en las actividades co-curriculares y extracurriculares**.

```
Call:
glm(formula = MotivoBaja ~ VEquiposrepresentativos + VExposiciones,
    family = binomial, data = log.train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.03933  -0.00097   0.57802   0.66805   1.00424

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)              1.946      1.069   1.820   0.0687 .
VEquiposrepresentativos2  16.271    882.743   0.018   0.9853
VEquiposrepresentativos4  14.988    882.743   0.017   0.9865
VExposiciones2           -16.512    882.744  -0.019   0.9851
VExposiciones4           -15.548    882.744  -0.018   0.9859
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 299.67  on 264  degrees of freedom
Residual deviance: 278.29  on 260  degrees of freedom
AIC: 288.29

Number of Fisher Scoring iterations: 13
```

Figura 3-30 modelo.reg.log.d6 : evaluación de las actividades co-curriculares y extracurriculares

Este modelo (Figura 3-30) después de 13 iteraciones no encuentra ninguna variable independiente como significativa, también se puede apreciar como el valor estimado muestra un error estándar que representa muchas veces más su valor, por ello no se tomará en cuenta ninguna de estas.

h) Durante su estancia en la IES: contexto Institucional evaluación de la satisfacción y orgullo de pertenencia

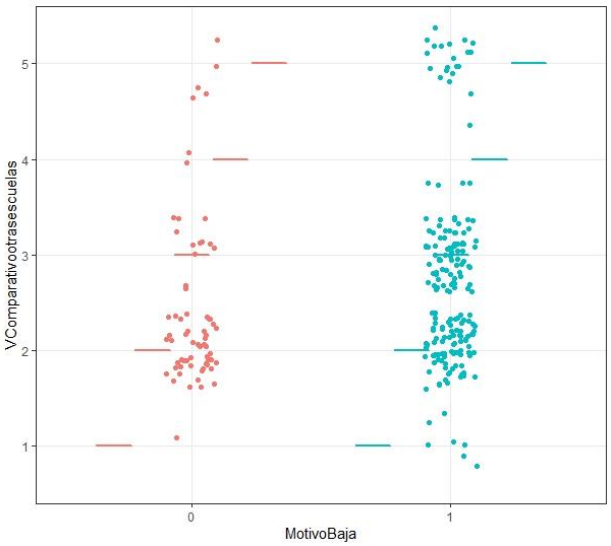


Figura 3-32 Evaluación al comparar a la Institución con otras escuelas

```
Call:
glm(formula = MotivoBaja ~ VComparativootrasescuelas, family = binomial,
    data = log.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0393  -1.4424   0.5287   0.9338   0.9338

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                1.94591    1.06904   1.820   0.0687 .
VComparativootrasescuelas2 -1.34171    1.08432  -1.237   0.2159
VComparativootrasescuelas3 -0.04879    1.11296  -0.044   0.9650
VComparativootrasescuelas4 -1.02962    1.35752  -0.758   0.4482
VComparativootrasescuelas5 -0.55962    1.18019  -0.474   0.6354
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 299.67  on 264  degrees of freedom
Residual deviance: 283.44  on 260  degrees of freedom
AIC: 293.44

Number of Fisher Scoring iterations: 4
```

Figura 3-31 modelo.reg.d7 : evaluación del sentido de satisfacción y orgullo de pertenencia

Aunque este modelo (Figura 3-31) no muestra significancia en sus variables independientes, se puede observar en la Figura 3-32 que cuando el desertor evalúa si al *comparar a la Institución* con otras escuelas la ubica entre no tengo bases para opinar (2) y regular (3).

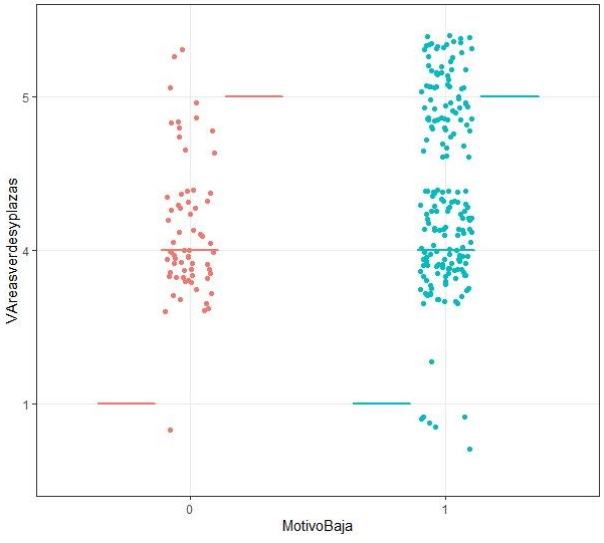


Figura 3-34 Evaluación de las áreas verdes y plazas del campus

```
Call:
glm(formula = MotivoBaja ~ VRedinalmbrica + VAreasverdesyplazas,
    family = binomial, data = log.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0393  -1.2557   0.5287   0.8503   1.1010

Coefficients: (1 not defined because of singularities)
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.94591    1.06904   1.820  0.06872 .
VRedinalmbrica2    -0.69777    1.17623  -0.593  0.55303
VRedinalmbrica3     0.39288    1.21437   0.324  0.74630
VRedinalmbrica4    -0.04879    1.11296  -0.044  0.96503
VAreasverdesyplazas4 -1.06582    0.38600  -2.761  0.00576 **
VAreasverdesyplazas5       NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 299.67  on 264  degrees of freedom
Residual deviance: 280.77  on 260  degrees of freedom
AIC: 290.77

Number of Fisher Scoring iterations: 4
```

Figura 3-33 modelo.reg.d8: evaluación de los espacios e infraestructura del campus

Este modelo (Figura 3-34) muestra que los desertores están contentos con los espacios del campus, sobre todo las *áreas verdes y plazas disponibles*, ya que la evalúan entre bien (4) y excelente (5) Figura 3-33.

A partir del análisis de cada modelo construido, después de identificar las variables independientes más significativas mostradas en la tabla 3-15, pensando en el principio de parsimonia que indica que se busca la menor cantidad de variables que expliquen los datos, se analizó la multicolinealidad que existe entre algunas de estas con la intención de elegir aquellas que predijeran el fenómeno con un mejor desempeño. En la Figura 3-35 se muestra la matriz de correlación, en la cual se puede observar la relación que existen entre algunas variables numéricas y a partir de ello, se decide eliminar el resultado del *área verbal* del *examen de admisión* por su fuerte correlación con el de *matemáticas*, la *beca ingreso* ya que está muy relacionada con el *promedio de ingreso* y nos interesa revisar si el desempeño académico del bachillerato puede predecir la deserción, también se eliminan las *becas talento y laboral* ya que están presentes con una significancia mayor otras becas que fueron asignadas por el *talento académico* de los desertores.

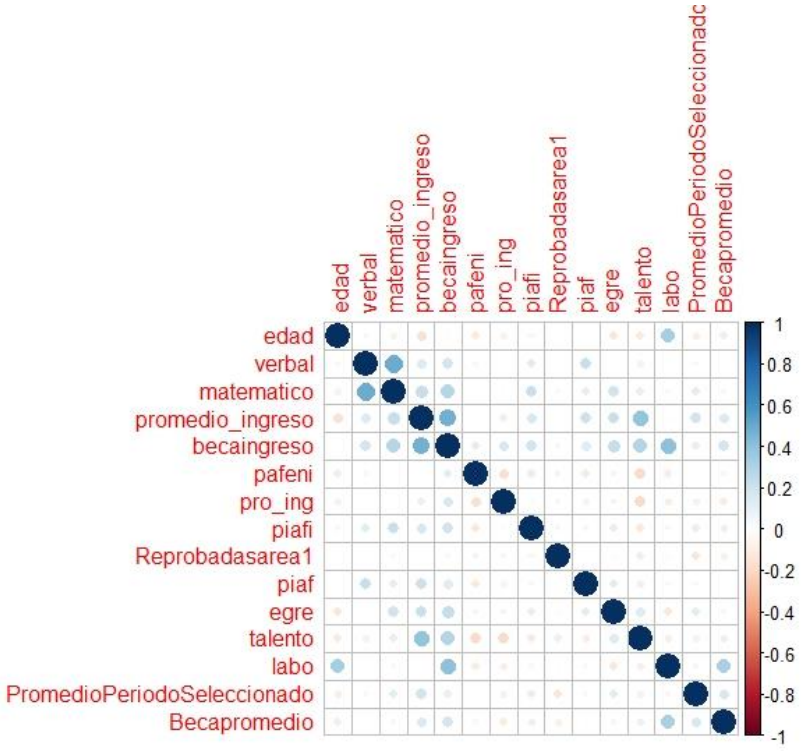


Figura 3-35 Matriz de correlación

Se construyeron algunos modelos con las variables seleccionadas, el valor del criterio de información de *Akaike (AIC),* es la cuantificación de que tan bien el modelo explica la variabilidad de los datos, y es usada para hacer una comparación entre varios modelos, el mejor es aquel que tiene el menor valor, este es calculado usando el valor de máxima verosimilitud y el número de variables del modelo: $AIC = 2K - 2\ln(L)$. Se hace un cálculo con los resultados de la Tabla 4-3 obteniendo la información mostrada en la tabla 3-16, se puede observar que el mejor modelo es el cuatro con ocho variables (k). Como es el mejor modelo al compararse con el mismo el valor del score, este es 0 (*Delta_AICc),* la proporción de la cantidad total del poder predictivo del modelo proporcionado por el conjunto total de modelos es de 85% *(AICcWt L*) También se puede observar el criterio de información Bayesiano (*BIC)* la cual es una medida de bondad de ajuste de un modelo estadístico y es usado como criterio de selección entre varios modelos, este criterio penaliza la cantidad de variables del modelo la fórmula es: BIC = número de parámetros del modelo (*k) x ln(n) – 2x ln(L*), donde k es el número de parámetros del modelo y *ln(L)* es la función de máxima verosimilitud para el modelo estadístico, en esta comparación se elige el menor valor de acuerdo al principio parsimonia mencionado, en la tabla 3-17 de comparación de modelos, se identifica el modelo cuatro con el menor valor.

Tabla 3-15 Variables independientes más significativas

Variable independiente	Modelo	Selecciona
edad	modelo.reg.d0	si
NSE	modelo.reg.d0	si
VexperienciacInstitucion1 (dummy)	modelo.reg.d0	si
verbal	modelo.reg.d1	no
matemático	modelo.reg.d1	si
tipo_escuela2	modelo.reg.d1	si
Vexcelenciaacademica1 (dummy)	modelo.reg.d1	si
promedio_ingreso	modelo.reg.d1	si
VDeportista1 (dummy)	modelo.reg.d1	si

Vpromedioacademico1 (dummy)	modelo.reg.d0	si
becaingreso	modelo.reg.d2	no
pafeni	modelo.reg.d2	si
pro_ing	modelo.reg.d2	si
piafi	modelo.reg.d2	si
piaf	modelo.reg.d2	si
egre	modelo.reg.d2	si
talento	modelo.reg.d2	no
labo	modelo.reg.d2	no
PromedioPeriodoSeleccionado	modelo.reg.d3	si
Reprobadasarea1	modelo.reg.d3	si
Becapromedio	modelo.reg.d4	si
Vcreditoeducativo2 (dummy)	modelo.reg.d4	si
VExposiciones2 (dummy)	modelo.reg.d6	si
VAreasverdesyplazas4 (dummy)	modelo.reg.d8	si

Tabla 3-16 Desempeño de los modelos de regresión logística finales

	Modelo 1	**Modelo 2**	**Modelo 3**	**Modelo 4**	**Modelo 5**
(Intercept)	7.23**	4.98*	2.75***	3.49***	1.35
	(2.79)	(1.97)	(0.51)	(.65)	(1.59)
edad	0.13				.14*
	(0.07)				(0.06)
NSE2	-1.08**			-0.94*	
	(0.41)			(0.43)	
NSE3	-0.40			-0.31	
	(0.50)			(0.52)	

promedio_ingreso	-0.62*	-0.40			
	(0.26)	(0.23)			
tipo_escuela2	-0.63				
	(0.36)				
VExcelenciaacademica1	0.62				
	(0.44)				
matemático	-0.003*				-0.00**
	(0.001)				(0.001)
piafi		-0.04***	-0.04***	-0.04***	
		(0.01)	(0.01)	(0.01)	
egre		-0.09**	-0.09**	-0.07*	
		(0.03)	(0.03)	(0.03)	
PromedioPeriodoSeleccionado			-0.01	-0.01*	
			(0.01)	(0.01)	
Becapromedio			-0.01	-0.01	-0.02 *
			(0.01)	(0.01)	(0.01)
AIC	**282.02**	**274.19**	**270.6**	**266.54**	**287.04**
BIC	**310.66**	**288.51**	**288.50**	**295.17**	**301.36**
Log Likelihood	**-133.01**	**-133.09**	**-130.30**	**-125.27**	**-139.52**
Deviance	**266.02**	**266.19**	**260.60**	**250.54**	**279.04**
Num. obs.	**265**	**265**	**265**	**265**	**265**

Tabla 3-17 AIC de los modelos finales

	K	AICc	Delta_AICc	AICcWt L	Cum.Wt	L
Modelo 4	8	267.10	0	.85	.85	-125.27
Modelo 3	5	270.81	3.73	0.13	.98	-130.30
Modelo 2	4	274.34	7.24	0.02	1.00	-133.09
Modelo 1	8	282.52	15.48	0.00	1.00	-133.01
Modelo 5	4	287.20	20.10	0.00	1.00	-139.52

El modelo final se muestra en la Figura 3-36, donde se aprecia que las variables predictoras más significativas son: las *becas piafi y egresado*, mismas que se otorgan al ingresar a la IES,

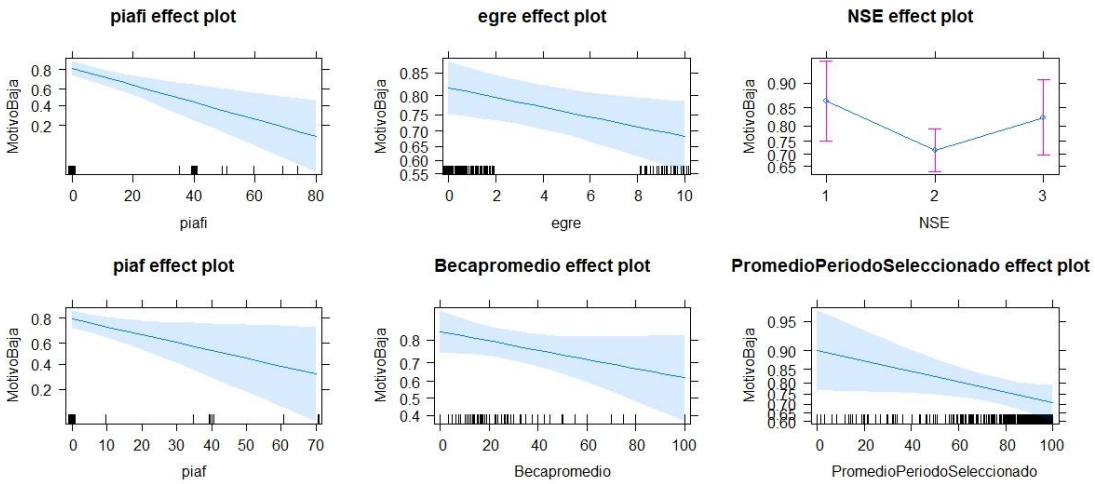
además de la *beca promedio* cuando se da de baja de la Institución; el *nivel socioeconómico* también es importante para los *niveles medio y alto*, también se puede ver el valor de la deviance nula y residual, en donde la primera indica que tan bien puede predecir el modelo solo con la constante de ajuste y la segunda incluyendo las variables predictoras, los valores indican que si mejora el resultado ya que es mayor que la deviance nula. Cuando se analizan los resultados y se convierte en una función lineal se obtiene el exponente del logaritmo, es decir, de los odds que se mencionaron en este capítulo de la siguiente forma: exp(coef(modelo 4)[para cada variable], dando como resultado la siguiente función de probabilidad:

ln (odds) = 3.49 -.042 (piafi) – 0.03(piaf) -0.07 (egre) - 0.01(Becapromedio) -.94(NSE2) -.31(NSE3)

algunos coeficientes son negativos por lo que esto explica que por cada unidad de reducción en el logaritmo de odds mayor probabilidad hay de que la variable explique la deserción, para poder traducir esta variablidad en términos de unidades se aplica la inversa del logaritmo natural por lo que el resultado sería ***exp(ln(odds) = 32.78 - 0.96 (piafi) - 0.97(piaf) - 0.93(egre) -0.99 (Becapromedio) – 0.39(NSE2) - 0.73(NSE3)***

```
Call:
glm(formula = MotivoBaja ~ piafi + egre + NSE + piaf + Becapromedio +
    PromedioPeriodoSeleccionado, family = binomial, data = log.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4520  -0.2822   0.4860   0.7161   1.8117

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 3.490337   0.647948   5.387 7.17e-08 ***
piafi                      -0.042365   0.011816  -3.585 0.000336 ***
egre                       -0.074309   0.033296  -2.232 0.025630 *
NSE2                       -0.940733   0.429883  -2.188 0.028644 *
NSE3                       -0.310395   0.523743  -0.593 0.553417
piaf                       -0.030564   0.013818  -2.212 0.026969 *
Becapromedio               -0.011027   0.006933  -1.590 0.111746
PromedioPeriodoSeleccionado -0.013372   0.006562  -2.038 0.041582 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 299.67  on 264  degrees of freedom
Residual deviance: 250.54  on 257  degrees of freedom
AIC: 266.54

Number of Fisher Scoring iterations: 5
```

Figura 3-36 Modelo de regresión logística final



Gráfica 3-13 Gráfica de efectos totales

En las gráficas de efectos totales (gráfica 3-13), se aprecia la influencia positiva o negativa que tienen cada variable independiente sobre la variable de respuesta, las variables de *beca piafi, piaf y egresado* y la *beca promedio al egresa*r, así como el *promedio del periodo de baja* indican que al momento de reducirse hay más probabilidad de desertar, en el caso de las variables dummy del nivel socioeconómico, cuando el desertor tiene un NSE medio tiende a quedarse en sistema CETYS.

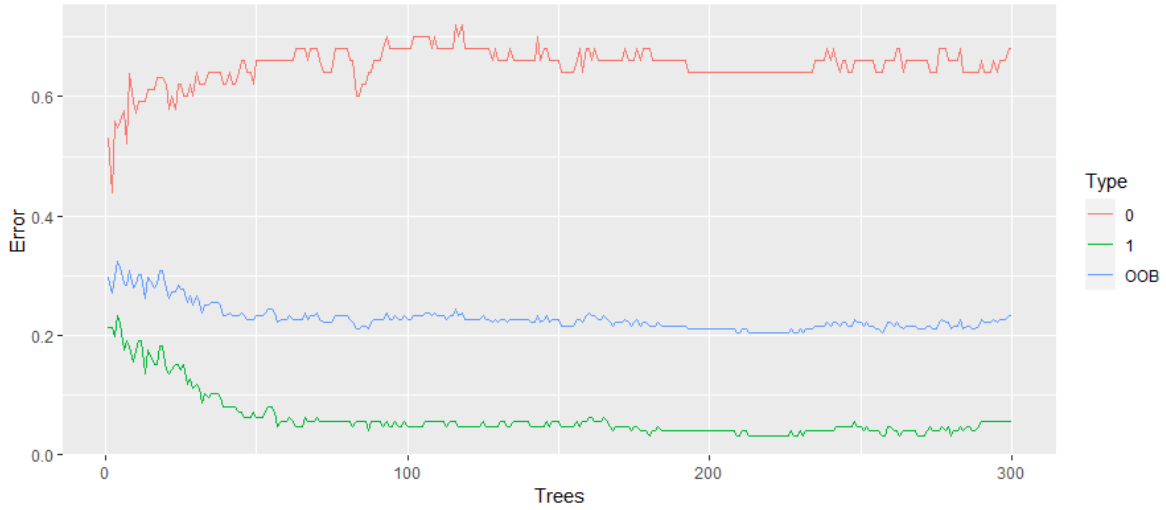
Modelos de bosque de árboles aleatorios (Random Forest)

Como se mencionó, cada árbol se va construyendo con un grupo de instancias quedando un tercio de ellas fuera de la muestra, estos datos son los que se clasifican como OOB (Out of bag, por sus siglas en inglés) mismos que son usados para evaluar si lo clasifica o no correctamente, para cada caso que va clasificando correctamente y va sumando las veces que esto sucede. Cada árbol o pseudo muestra de datos identifica las variables independientes más importantes o con mayor valor y va ajustando los modelos, a este proceso se le llama bagging, al final la predicción se obtiene con la media de todas las variables cuantitativas o la moda de las cuantitativas más importantes o con mayor valor.

Es importante también indicar el número total de predictores que se buscarán como parte de los parámetros de éste algoritmo , para problemas de clasificación como el nuestro, se recomienda iniciar con la raíz cuadrada del total de predictores o variables de entrada, para este caso el total es de 102 variables independientes, por lo que la $\sqrt{102}$ = 10. Sin embargo, se calcula el total de

predictores evaluando el OOB clasificación del error hasta encontrar el que presenta la tasa menor del error. En cuanto a la importancia de los predictores serán aquellas variables cuantitativas o la moda de las cuantitativas más importantes o con mayor valor, es decir, cuando un predictor es permutado el error incrementará por lo que el algoritmo presupone que el predictor si aportaba al modelo, se va cuantificando el promedio de los árboles en las que va participando el predictor, por lo que determinar el número de árboles que conformaran el bosque es importante, como se observa en la gráfica 3-14, alrededor de los 250-300 árboles ya se estabiliza OOB error; si visualizamos el modelo configurado con 10 variables de entrada y 50 árboles podemos observar el modelo mostrado en la Figura 3-38 con una tasa de OOB error del 0.21% , error obtenido al clasificar las observaciones OOB que no fueron incluidas en el conjunto de datos de entrenamiento Bagging, lo cual significa que la tasa de clasificación correcta de las observaciones OOB fue de 79%, (complemento del error OOB).

Podemos obtener los mejores parámetros utilizando el paquete Ranger de R (Wright y Ziegler, 2017), por medio del cual probaremos diferentes parámetros : #arboles (trees), #variables seleccionadas (mtry), profundidad de las particiones o el # máximo de nodos de cada árbol (depth) (Figura 3-37). Se busca identificar la importancia de las variables predictoras de acuerdo al índice de Gini (impurity), el cual es una medida de desorden (MeanDecreaseGini), que indica que a mayor medida mayor importancia en los modelos, ya que los valores más cercanos a la clase 0 (no desertor) tienen mayor desorden y valores más próximos a la clase 1 (desertor) tienen un menor desorden, por lo que se busca el mayor valor de decrecimiento para este indicador, el resultado de este análisis se puede observar en la Figura 3-38.



Gráfica 3-14 # de árboles y la estabilización del error estándar

```
Call:
 ranger(formula = MotivoBaja ~ ., data = train.arbol, num.trees = param_grid$num_trees[i],      mtry = param_grid$mtry[i], max.depth
 = param_grid$max_depth[i],      importance = "impurity", seed = 123)

Type:                             Classification
Number of trees:                  300
Sample size:                      176
Number of independent variables:  102
Mtry:                             5
Target node size:                 1
Variable importance mode:         impurity
Splitrule:                        gini
OOB prediction error:             21.02 %
> #oob_error
> menorError
[1] 0.2102273
> posicion
[1] 40
> arboles
[1] 300
> maxvariables
[1] 5
> nodos
[1] 102
```

Figura 3-38 Los mejores parámetros del modelo

```
Call:
 randomForest(formula = MotivoBaja ~ ., data = train.arbol, mtry = maxvariables,      importance = TRUE, ntree = arboles, max_nodes = depth)
               Type of random forest: classification
                     Number of trees: 300
No. of variables tried at each split: 5

        OOB estimate of  error rate: 22.16%
Confusion matrix:
   0   1 class.error
0 15  35  0.70000000
1  4 122  0.03174603
```

Figura 3-37 Mejor modelo de Random Forest

Es relevante entender que este algoritmo al ir construyendo de forma aleatoria los árboles, sin poda, va identificando las variables con más votos por lo que al final no construye un árbol sino de acuerdo al mejor valor, construyó 300. Para ejemplificar, en la Figura 3-39 se puede observar el árbol del modelo *random forest* con el menor número de nodos, utilizando el paquete *ggraph* de Thomas Lin Pedersen.

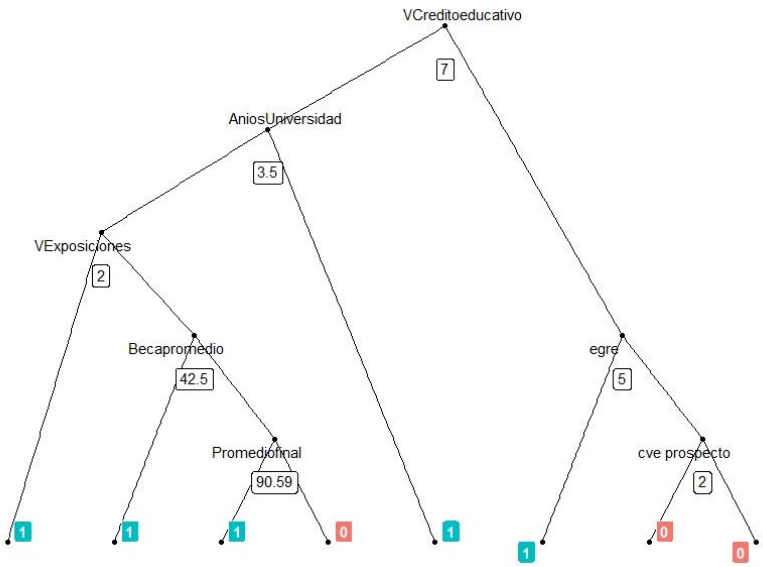


Figura 3-39 Árbol con el menor número de nodos

Las variables predictores más importantes de este modelo, de acuerdo a la pureza de nodos según el índice de Gini se muestran en la Figura 3-40. Entonces, las variables que más aportan a esta reducción son el *programa académico* al que se inscriben, *puntaje del examen de admisión*, el *puntaje de todas las áreas del examen admisión* por lo que estas variables están fuertemente relacionadas con el *puntaje final*, el *porcentaje de avance* en su programa académico, su *porcentaje total de beca al ingresar a la carrera*, su *edad*, la *escuela de procedencia*, la *beca asociada al talento académico* que demuestra el estudiante al ingresar a la Institución, y su *promedio de ingreso.*

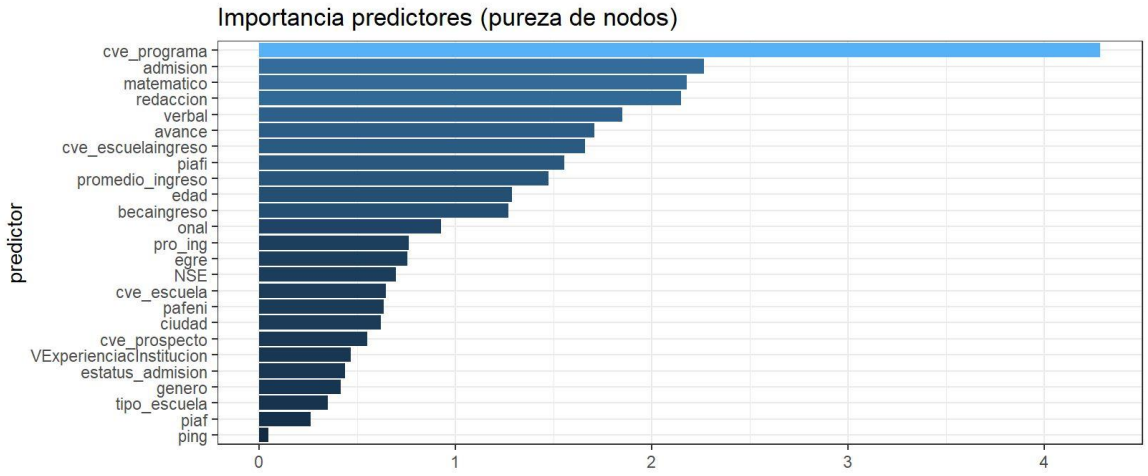


Figura 3-40 Mejores predictores de acuerdo a la pureza de Gini

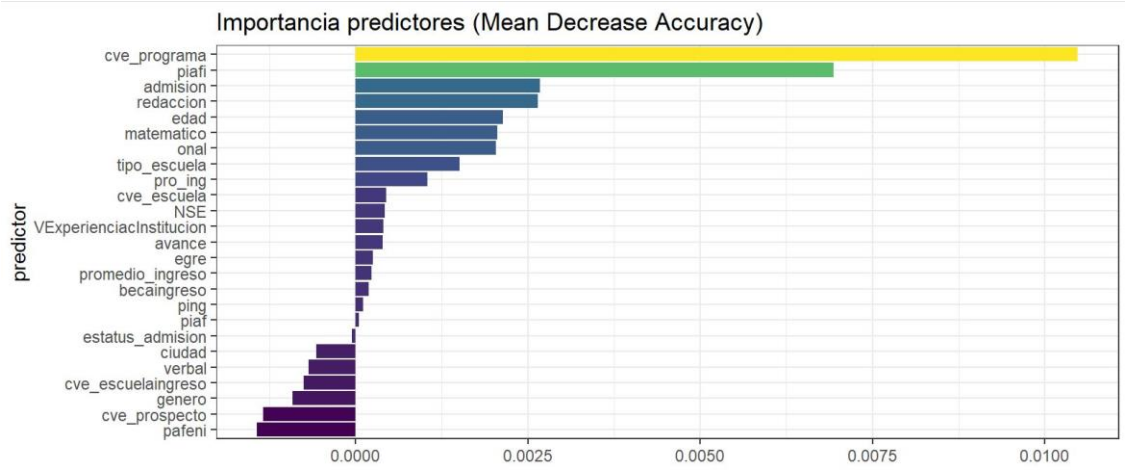


Figura 3-41 Mejores predictores de acuerdo a la exactitud o precisión del modelo

Se construyeron y evaluaron cinco modelos (Figuras 1 a la Figura 5 del anexo 2), a partir del análisis de las variables de los diferentes contextos de la vida del estudiante (Gráficas 1 a la 8 del anexo 2). Al observar las variables predictoras más importantes del modelo según la precisión del mismo, lo cual significa que si se quitan algunas de estas variables la precisión disminuirá, podemos observar en la Figura 3-41 que el *programa académico* y el *examen de admisión* vuelve a presentarse y la *edad*, además de las áreas correspondientes al e*xamen de admisión como la de redacción y matemática,* se utilizó la librería rpart de R, para construir un árbol de clasificación con estas variables predictoras con la finalidad de ilustrar las reglas de clasificación definidas para identificar al desertor del que no deserta en la Figura 3-42.

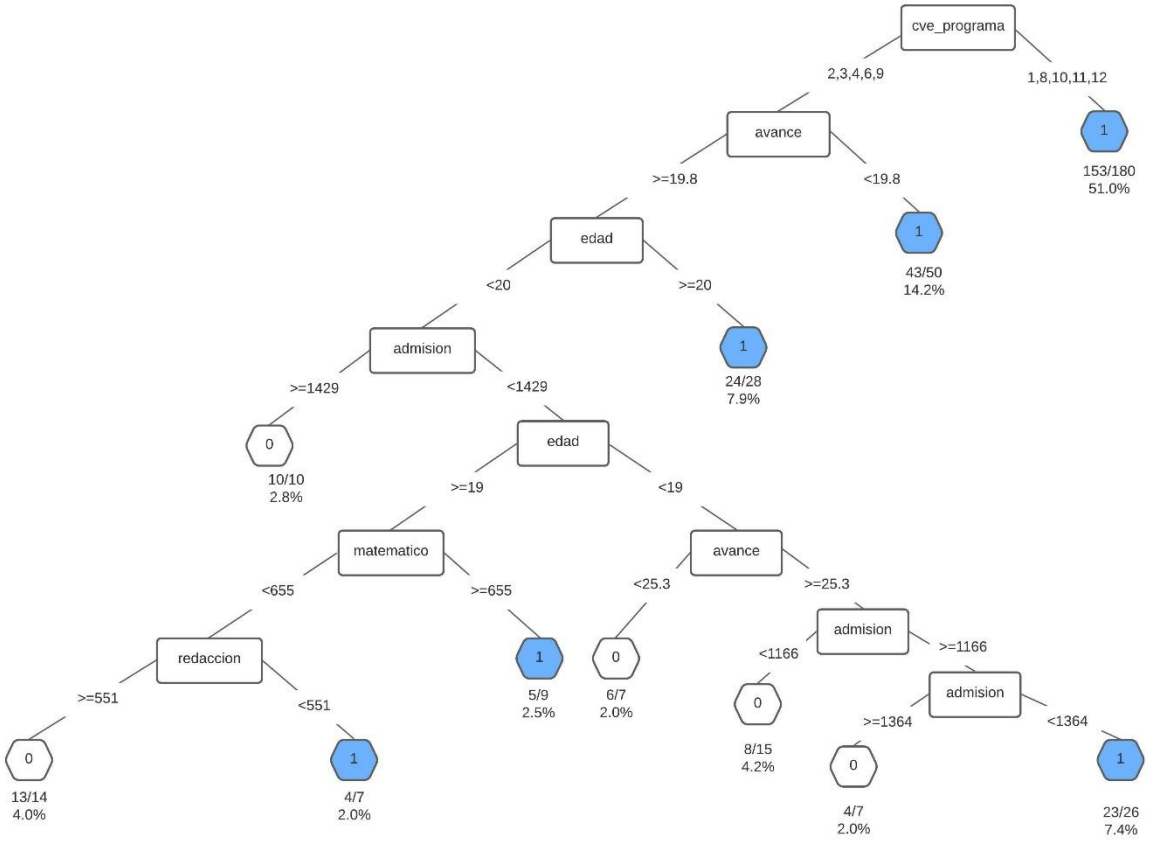


Figura 3-42 Árbol de clasificación con rpart

3.8 Prueba experimental de modelos

En esta sección se presentará la evaluación de desempeño de los modelos elegidos y una prueba experimental con el set de pruebas que se preparó de acuerdo a la distribución de instancias seleccionadas.

3.8.1 Evaluación del modelo de regresión logística

Primero se evaluó la predicción del modelo con el set de entrenamiento indicando que tiene el 60% de probabilidad de predecir a los que no van a desertar del Sistema y del 80% a los que van a desertar, en cuanto a la medición que se hizo en el set de prueba, el cual cuenta con el 25% de las instancias del total del set de datos, indica que la probabilidad de predecir a los que se

quedaran en el Sistema es de 73% y el de predecir a los que se van del 78%, sin embargo usaremos otros indicadores de desempeño para evaluar el modelo y su capacidad de predicción.

La matriz de confusión

Es una herramienta muy utilizada en el aprendizaje de máquina para valor la capacidad de predicción de un modelo, ya que muestra explícitamente cuando las clases del set de datos son clasificadas de forma correcta o no de forma independiente, lo cual nos sirve para calcular otras métricas importantes asociadas.(Nwanganga, Fred Chapple, 2020)

Se utilizó el lenguaje de programación R para realizar todas las operaciones asociadas a la evaluación de desempeño del modelo, calculando en el set de prueba cualquier predicción de respuesta con un valor menor a 0.5 como falsa (negativo) y las mayores a este valor como verdaderas (positivo), obteniendo la matriz de confusión (Figura 3-43). Se muestra como el modelo clasifica correctamente (en verde) los que van a desertar (63) y los que no lo harán (2), también podemos ver también que hay algunas clases que no fueron clasificadas correctamente (en rojo) como los casos que fueron clasificados como que iban a desertar (20) y no lo hicieron, así como los que se predijeron como no desertores y si lo fueron (3).

Figura 3-43 Matriz de confusión

Esta distribución entre los que el modelo predijo incorrectamente: falsos positivos (FP) y falsos negativos (FN), y los que predijo de forma correcta: verdaderos negativos (VN) y verdaderos positivos (VP), sirven para calcular otras métricas. La exactitud (Accuracy), representa el sesgo de una estimación, es decir, que tan cerca está el modelo de predecir correctamente un valor verdadero *((VP+VN) / (VP+FP+FN+VN))*, para nuestro modelo el resultado es 73.86%, el cual es un valor aceptable pero no suficiente para indicar que es un buen modelo.

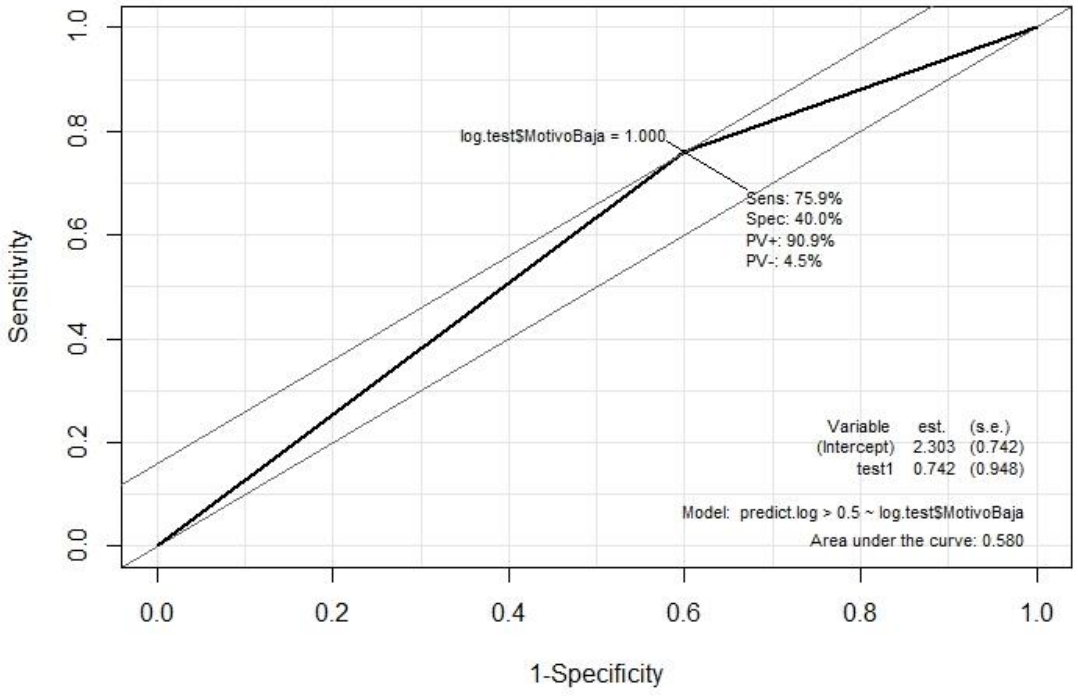
La precisión (Precision), representa la dispersión de los valores que se obtienen a partir de la medición realizada con el modelo y es la proporción de verdaderos positivos que se predijeron entre todos los positivos predichos (VP)/(VP+FP), el resultado obtenido es del 75.90%, por lo que el modelo presenta una buena precisión. La sensibilidad (Recall o sensivity), es la tasa de verdaderos positivos (VP) que el modelo puede clasificar correctamente, es decir, indica que tan bien predice a aquellos que se irán del Sistema CETYS, representado por VP/(VP+FN), nuestro modelo indica que tiene una sensibilidad del 95.45% este valor es bueno ya que de 10 casos que serán desertores, 9 seran predichos de forma correcta.

La especificidad (Especifity), es otro valor importante que complementa la sensibilidad, ya que esta métrica indica la capacidad del modelo de predecir correctamente a los que se quedarán en el Sistema CETYS (VN), representado por VN/(VN+FP), el resultado obtenido es de 40%, lo cual es un valor no muy bueno.

En este momento puede decirse que el modelo de regresión construido detecta bien a los desertores ya que este modelo presenta una alta sensibilidad y una buena precisión, la prueba de F1 score es la métrica que resume la precisión y la sensibilidad en una sola, esta medida es útil porque la distribución de clases no es equitativa, el set de datos muestra una proporción más alta de desertores, lo cual representa un desbalance, la F1 se calcula de la siguiente manera: $2 * (Recall) * Precision/(Recall + Precision)$, el resultado de esta métrica es de 85.13% por lo que esto nos permite concluir que es un buen modelo para predecir la deserción,.

En la gráfica 3-15, se puede observar el área bajo la curva ROC (Receiver Operator Characteristic Curve, por sus siglas en inglés) la cual muestra la capacidad de diagnóstico del modelo de 58%, el cual no es tan relevante, aunque es importante identificar que esta medida mide la capacidad de discriminación del modelo no necesariamente de la calibración del mismo, es decir, que tanta sensibilidad o exactitud tiene. La sensibilidad y la especificidad son medidas

de desempeño con las que se busca tener un balance adecuado entre que el modelo identifique de forma certera a quien va a desertar con respecto a si el modelo también identifica al que se va a quedar en el Sistema CETYS, siendo más relevante el primero para este caso, por otro lado la precisión, conocida también como el valor predictivo positivo, es muy alto por lo que el modelo predice de forma correcta a los estudiantes probables de ser desertores y esto lo hace muy confiable (Brett Lanz, 2019), y de acuerdo al estudio realizado por Urbina-Nájera, A.B. et. al (2020) el mejor porcentaje de predicción reportado con el uso de regresión lineal y logísica identificando las causas de la deserción escolar, fue de 66.59%; por lo que nuestro modelo muestra un mejor valor de desempeño.



Gráfica 3-15 Desempeño del modelo de regresión logística

3.8.2 Evaluación del modelo de bosque de árboles aleatorios

Al evaluar el modelo final construido utilizando el algoritmo de RF en el set de entrenamiento, el cual fue dividido en un 50% de las instancias del total del set de datos, se obtienen los resultados mostrados en la Figura 3-44, aquí se pueden identificar las variables predictoras mas importantes. Las variables predictoras son: el *programa académico*, el puntaje

en *el área de matemáticas y verbal del examen de admisión*, lo cual significa que se calculó el número de veces que esta variable fue predicha correctamente (OOB) restándoselo al número de clases correctas en la partición de datos a evaluar, obteniendo un promedio de este valor; así es como se calcula la importancia del predictor. Otras predictoras importantes son: el *promedio al ingresar*, su *beca de ingreso*, siendo más relevante la *beca egresado y la piafi* la cual está relacionada con el desempeño académico del estudiante, también se puede observar la clasificación de clases a través de la predicción realizada por el modelo.

Al evaluar el modelo en términos de probabilidad, se puede observar como el modelo muestra una probabilidad relevante de clasificar correctamente al desertor como lo observamos en la tabla 3-18.

```
Call:
 randomForest(formula = MotivoBaja ~ cve_programa + avance + admision +      redaccion + matematico + piafi + ver
bal + edad, data = train.arbol,      mtry = maxvariables, importance = TRUE, ntree = arboles,      max_nodes = de
pth)
               Type of random forest: classification
                     Number of trees: 300
No. of variables tried at each split: 5

        OOB estimate of  error rate: 24.43%
Confusion matrix:
    0   1 class.error
0 15  35  0.70000000
1  8 118  0.06349206
```

Figura 3-44 Modelo de bosque de árboles aleatorio final

Tabla 3-18 Evaluación del modelo random forest en términos de probabilidad

# observaciones	No desertor	Desertor
160	0.21	0.79
167	0.25	0.75
275	0.43	0.57
141	0.05	0.95
265	0.00	1.00
322	0.41	0.59

La matriz de confusión

Así como evaluamos los modelos de regresión logística, utilizamos la matriz de confusión para evaluar el desempeño del modelo de bosques aleatorios (RF).

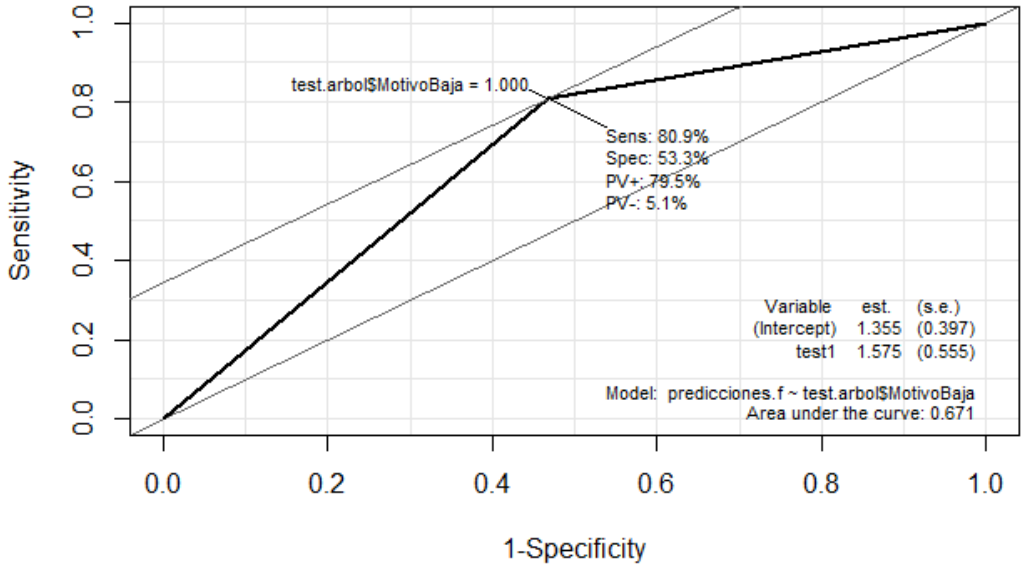
La Figura 3-45 muestra como el modelo clasifica correctamente (en verde) los que van a desertar (131) y los que no lo harán (8), hay algunas clases que no fueron clasificadas correctamente como aquellos que fueron identificados como desertores y no eran (31), así como aquellos que se predijeron como no desertores y si salieron del Sistema CETYS (7).

Figura 3-45 Matriz de confusión del modelo de bosque de árboles aleatorios

La medida de exactitud (*Accuracy*) para este modelo es de 78.53%, indicando la capacidad del modelo para predecir correctamente. La precisión (*Precision*) representa la dispersión de los valores que se obtienen a partir de la medición realizada con el modelo, y es del 80.86%, por lo que el modelo presenta una buena precisión, lo que puede indicar que está detectando 8 de cada 10 casos que pueden desertar del campus. La sensibilidad (*Recall o sensivity*), indica que tan bien predice a aquellos que se irán del Sistema CETYS, representado por VP/(FN+VP), nuestro modelo indica que tiene una sensibilidad del 94.93% este valor es bueno ya que detectará 9 de cada 10 de los casos que serán desertores. La especificidad (*Especifity*), es otro valor importante que complementa la sensibilidad, ya que esta métrica indica, como se había explicado, la capacidad del modelo de predecir correctamente a los que se quedaran en el Sistema CETYS, el resultado obtenido es de 53.3%, el cual no es un buen valor, ya que tiene un poco menos del 50% de probabilidad e fallar en la predicción de este tipo de desertor. Aplicamos la prueba de F1 score

para revisar la precisión y la sensibilidad en una sola, el resultado es de 87.33% por lo que esto nos permite concluir que es un buen modelo para predecir la deserción.

En la gráfica 3-16 se muestra el área bajo la curva ROC con una capacidad de diagnóstico del modelo de 67%, lo que demuestra este es un mejor modelo que el de regresión logística.



Gráfica 3-16 Desempeño del modelo de bosque de árboles aleatorios

3.9 Análisis de los resultados y discusión

Los modelos coinciden en su capacidad de predecir aquellos estudiantes que se convertirán en desertores, mejor que la capacidad para detectar a los que se dieron de baja para migrar a otros campus y por consecuencia se quedarán en el Sistema CETYS. Esto se puede deber a que las clases de la variable de salida no están balanceadas, ya que el 75% de las instancias representan a los desertores y el resto a los que se quedan en el Sistema.

La métrica de armonía F1, es del 85.13% para el modelo de RL y de 87.33% para el modelo de RF, es una buena métrica para ambos modelos.

Se considera la métrica de Recall o sensibilidad como la más relevante ya que los modelos tienen la capacidad de predecir a los que se irán del Sistema CETYS, es decir a los desertores, con un valor por encima del 90%, esto es muy valioso porque las estrategias que se implementen

para asegurar la permanencia y éxito estudiantil serán positivas, aunque el estudiante no sea un candidato a desertar, esta medida es del 95.45% para el modelo de RL y de 94.93% para el modelo de RF.

Las variables predictoras coinciden en aquellos atributos académicos al momento de ingresar a la Institución, como el *examen de admisión*, siendo este el más significativo según el modelo de RL. La información que arroja este modelo indica que la magnitud de los coeficientes que lo conforman, son negativos, por lo que es muy relevante percatarnos que por cada unidad a la baja de estas predictoras la variable de respuesta será que el estudiante se convertirá en un desertor.

Indicando, que cada área que conforma este examen tiene un nivel de importancia: el área de redacción, matemático y verbal (en ese orden). Esto se puede observar con detalle en la gráfica 3-7, donde se muestra que hay muchos desertores con un puntaje bajo en el área de redacción además que el área de matemáticas muestra una distribución concentrada en los puntajes por debajo de la media.

Hay una gran cantidad de desertores con puntajes bajos en el examen de admisión en las carreras de la escuela de Administración y Negocios, lo cual coincide ya que el *programa académico* una variable predictora identificada en el modelo RF.

Si agregamos a este análisis a las variables predictoras de RF coinciden en aquellos atributos académicos al momento de ingresar a la Institución, como el *examen de admisión*, y sobresale considerablemente en importancia, el *programa académico* que eligieron al entrar a la IES *su promedio al desertar*, y *su porcentaje de avance* indicando que estos estudiantes desertaron en los primeros semestres de su programa académico.

Se puede afirmar que el primer año de estudios es importante, ya que la mayor concentración de la deserción se da en los dos primeros semestres, siendo esta del 54.9% como se aprecia en la tabla 3-8, esto se confirma en el modelo RF ya que el porcentaje de avance es una variable predictora importante para este modelo y el árbol de clasificación de la Figura 3-42 muestra en sus reglas de clasificación que a menor avance hay una mayor probabilidad de ser un desertor. Con estos datos y lo dicho del primer año de estudios, permite inferir que es en el primer año cuando los estudiantes pudieran cambiar sus intenciones de estudio o las competencias y las habilidades con las que entran a la IES no son suficientes para lograr un buen desempeño académico (Cu Balán, 2008). El examen de admisión y el desempeño académico ya habían sido

mencionados en otros estudios como los de (Huesca Ramírez y Castaño Corvo, 2007), (Pérez et al., 2018) y por (Eckert y Suénaga, 2015)

También el *promedio de ingreso* es muy importante y se puede inferir que guarda una gran relación con las becas asociadas al desempeño académico previo como son las becas *piafi* y *piaf*. También se puede apreciar que la *beca egresado*, es una buena variable predictora esto lo validamos cuando analizamos las escuelas de procedencia, ya que muchos de los desertores vienen de la escuela preparatoria del Sistema CETYS.

El rendimiento académico durante su estancia en la Institución es significativo, ya que el promedio que tiene el estudiante cuando se da de baja está representado por el atributo *PromedioperiodoSeleccionado*, siendo este más bajo para los alumnos desertores que para los que se quedan en el Sistema CETYS.

Spady (1970), señaló que las variables significativas de la deserción eran el desempeño académico pasado y las bajas calificaciones. Tinto (1975) y Bean (1985) encontraron también que el primer año de estudios era relevante y además indicaron que el apoyo financiero era un factor importante para los estudiantes y podía ser motivo de deserción.

Cabrera et al (1993) en el "Modelo teórico de la retención del estudiante", también enfatizaron en los apoyos financieros en cuanto a la relación costo-beneficio como una influencia directa sobre la decisión de permanencia o deserción, se pudo confirmar por medio del modelo RL ya que este muestra que las variables predictoras más significativas son la "beca piafi" con un valor de - 0.96, la "piaf" con un valor de -0.97; siendo estas becas académicas otorgadas al entrar a la IES producto del promedio obtenido en el bachillerato y el puntaje en el examen de admisión, y la de "egre" con - 0.93 la cual es una beca que no importando los valores del examen o del promedio se otorgan a todos los estudiantes que provienen de alguna preparatoria del Sistema CETYS. La mayoría de estos estudiantes becados provienen del mismo campus (gráfica 3.5.

Cuando Spady (1970) mencionó en el "Modelo sociológico explicativo del proceso de deserción", que son diversas las fuentes a las están expuestos los estudiantes dentro de la IES, por ejemplo: cursos, profesores, administradores, procesos académicos y administrativos y Pascarella & Terenzini (1980) en su "Modelo causal", también enfatizaron en las variables académicas previas al ingreso, el programa académico elegido y las características de la IES, como variables que guardan una correlación relevante. Además, incluyeron otras variables personales como el género, la edad, el nivel socioeconómico entre otros. En el modelo de RF la edad es importante y

en el de RL el nivel socioeconómico se considera una variable predictora cuando este es medio o bajo (gráfica 3-4), aunque se pudo identificar que la edad más frecuente de los desertores es de 18 años, y se pudo notar (gráfica 3-3) que la escuela de Administración y Negocios es la que muestra más estudiantes desertores mayores de 20 años.

Es importante hacer notar que estos estudiantes tienden a perder apoyos financieros. La gráfica 3-9 muestra que algunos de estos desertores entran con porcentajes de beca por encima del 50%. Se observa como la variable "beca-promedio", que representa el promedio total de beca del estudiante al momento de desertar (Figura 3 26) muestra que este porcentaje se concentra por debajo del 25-30% en los desertores. Esta variable se presenta con un valor de -0.99 en el modelo de RL, lo cual indica que por cada unidad que se reduzca, aumenta la probabilidad de que el estudiante deserte.

Se identificó además que el promedio de reprobación y aprobación de los estudiantes es muy semejante entre las dos escuelas (gráfica 3-10). La reprobación de estos desertores es aproximadamente del 33%. Las áreas de reprobación con mayor concentración (gráfica 3-11), son las materias que desarrollan las competencias de formación integral y aquellas que trabajan en las competencias de formación básica de colegio, las de ciencias de la ingeniería y las de ciencias administrativas según sea el caso. Lo anterior tiene mucho sentido ya que estas materias reprobadas están concentradas en los primeros cuatro semestres de cada programa académico.

Existe una correlación evidente entre las variables predictoras, ya que las capacidades académicas con las que ingresa el estudiante, tienen un impacto en los apoyos financieros, y su desempeño durante su estancia en la Universidad termina por reducirlo aún mas. Las becas no son muy altas y además se siguen reduciendo a muy temprana edad académica ya que desertan muy pronto, hay una gran representación de desertores de la Escuela de Administración y Negocios.

Capitulo 4 Conclusiones, trabajo futuro y recomendaciones

Esta investigación partió de la recopilación de datos históricos, la preparación de estos y su pre procesamiento; esto se hizo desde diversas fuentes de datos, formales e informales, con la intención de obtener información de los estudiantes que desertaron de la Institución del período del 2008 al 2018. Se tuvo que acudir al archivo físico del campus para recolectar y validar mucha de la información que no se pudo identificar en las bases de datos de la Institución, invirtiendo un tiempo de trabajo de más de seis meses ya que la información no se encontraba íntegra o contenía datos ausentes que en su tiempo no habían sido registrados en los Sistemas de información. El proceso de extracción, transformación y carga de los datos fue una etapa muy relevante para asegurar la integridad y la calidad de los datos que se usarían en este estudio.

La literatura y lo que se ha dicho sobre la deserción y el uso de minería de datos en la educación, sirvió como fundamento para identificar que datos eran importantes de considerar en la investigación. Se construyó un modelo entidad-relación con todos los atributos que pudiera caracterizar a un desertor antes, durante y cuando decide darse de baja de la IES. Sin embargo, no se pudieron incluir todos los que pudieran representar una experiencia o algún atributo cualitativo que permitiera caracterizar a estos estudiantes desertores durante su estancia en la IES ni al momento de entrar a esta, ya que la Institución no concentra todos los datos que pudieran ser significativos al no contar con un registro oficial de las actividades extracurriculares, cocurriculares, de atención y seguimiento. También, la encuesta de satisfacción donde se podrían evaluar las percepciones y experiencias cualitativas del servicio educativo, del sentido de pertenencia, de su experiencia en CETYS entre otros; por su carácter anónimo no se puede identificar el sentir de los alumnos desertores de forma individual.

Diaz Pedroza et al., (2019), en su investigación documental sobre técnicas, herramientas, algoritmos y atributos para la minería de datos utilizados en la deserción, mencionaban que los atributos más utilizados por los autores se clasifican en variables académicas, incluyendo las calificaciones durante su estancia como su promedio al ingresar a la IES así como el examen de admisión. Este estudio incluye algunas variables cualitativas obtenidas por la encuesta de satisfacción. Los autores recomiendan recolectar la información relacionada con aspectos familiares al ingresar, y aquello que puede incidir en la deserción como las aspiraciones del estudiante, su percepción de la Institución académica y sus intereses vocacionales, además de

documentar y recolectar toda la experiencia durante su estancia en la IES. Importante también obtener su participación en la vida estudiantil: actividades extracurriculares, liderazgo estudiantil, voluntariado entre otros. Solo se puede inferir por la encuesta de satisfacción lo que percibe del servicio educativo, su relación con el colectivo universitario: profesores, administrativos, compañeros, etc.

Los modelos coinciden en su capacidad de predecir aquellos estudiantes que se convertirán en desertores. No tienen un buen desempeño en general para detectar a los que se dieron de baja para migrar a otros campus y por consecuencia se quedarán en el Sistema CETYS. Este fenómeno es importante revisarlo, ya que no hay un seguimiento de los estudiantes que migran a los otros campus, debe hacerse una investigación especial ya que estos estudiantes entran como nuevo ingreso a los otros campus con un número de matrícula distinto.

La condición de pandemia actual de Covid-19 cambia la deserción como fue concebida durante el período de tiempo observado, por lo que no se pudo aplicar el modelo a los estudiantes de un semestre en curso, validando al final del semestre si el modelo predijo la deserción de forma precisa, ni se pudo probar en otro campus del Sistema CETYS. Se puede concluir que si existe una relación multivariada entre las variables que influyen en la deserción y que la minería de datos sirvió para encontrar los atributos más significativos de los desertores de este caso de estudio y las variables predictoras de la misma.

Esta investigación también comprobó que el fenómeno de la deserción es multifactorial coincidiendo como se indicó con algunos estudios antes realizados (Urbina-Nájera et al., 2020), (Londoño, 2013), (Peralta et al., 2017), algunos de estos coinciden que el desempeño académico previo y durante su estancia en la IES influye de forma importante, por lo que considerar estrategias de acompañamiento y seguimiento en los primeros años de su desarrollo académico es sustantivo, se puede comprobar que el examen de admisión es un instrumento valioso para medir las competencias de un estudiante al entrar a la IES, por lo que es relevante profundizar en estos resultado para diseñar estrategias asertivas y puntuales en el acompañamiento académico mencionado.

En cuanto a la implementación de algoritmos de aprendizaje automático para hacer el análisis predictivo, se identificaron como los más adecuados para este fenómeno los de aprendizaje supervisado y entre estos los de clasificación, la literatura también ayudó a reconocer los más pertinentes y eficientes según el problema planteado y las características del set de datos.

Aplicar algoritmos de aprendizaje automático fue muy importante, porque permitió reconocer desde los datos patrones de comportamiento que van más allá de la información que se tenía sobre la baja de estudiantes, ya que lo único que se registraba era el motivo de baja, y muchos de los desertores no habían cerrado su proceso formal de baja por lo que el motivo tampoco estaba registrado con la causa verdadera.

Los algoritmos de aprendizaje de máquina supervisados utilizados en este estudio muestran su efectividad desde dos perspectivas, por un lado, la capacidad del algoritmo de regresión logística de traducir la influencia de una variable predictora en la variable de salida, así como la correlación que existe entre estas variables de entrada del modelo, ya que pudimos identificar como valida que los factores económicos son muy representativos de la deserción. Sin embargo, no incluye otras variables que, gracias al algoritmo de bosque de árboles aleatorio, se identificaron como importantes complementando este estudio, y que además en el contexto de este caso de estudio, si guardan una relación significativa entre estas, ya que los atributos académicos tienen una influencia directa en los factores económicos.

El algoritmo de Random Forest, fue muy valioso para identificar las variables que predicen al desertor con un mejor desempeño que el de Regresión Logística, sin embargo, por su diseño es difícil visualizar como se fueron construyendo cada uno de los árboles que consideró en cada ensamble y así ir siguiendo la pista de cada regla de clasificación, pero puede ser un excelente algoritmo para reducir las variables de entrada de un modelo y así considerar las más significativas como finalmente lo pudimos observar en el árbol de clasificación construido en la Figura 3-42. La combinación de ambos algoritmos pudo demostrar que hay una correlación clara entre las variables que estos modelos identificaron como predictoras, aunque ambos mostraron un desempeño bajo en la identificación de aquellos que no desertarán del Sistema CETYS, lo cual no es tan relevante dado el contexto de este estudio, porque cualquier estrategia que se implemente para disminuir la deserción impactará de forma positiva en aquellos estudiantes que ya consideran el Sistema CETYS como una IES donde pueden realizar sus estudios de licenciatura.

Finalmente, se puede concluir que los desertores del Sistema CETYS entraron con un desempeño regular, la mayoría eligieron programas académicos de la escuela de Administración y Negocios y obtuvieron una beca no asociada a la excelencia académica, una gran proporción de estos desertores son de los primeros semestres como se mostró en el análisis descriptivo de los datos.

Se logró el propósito de la investigación ya que se pudieron identificar los atributos y caracterizar a los desertores, en el período de agosto de 2008 a agosto de 2018, identificando la correlación o asociación entre estos atributos. Además se construyeron modelos predictivos, con los que se pudo obtener información que permita intervenir de forma estratégica y temprana la deserción, se puede recomendar a la Institución que implemente estrategias de retención ya que los estudiantes con mayor riesgo son estudiantes de niveles socioeconómicos bajos y altos, además que estos estudiantes ingresan con un desempeño académico no muy alto, ya que no trae promedio altos y los resultados del examen de admisión tampoco son altos, es muy importante el acompañamiento académico ya que tienden a perder beca durante su estancia en la Institución desde los primeros semestres, ya que van incumpliendo con los requisitos de no reprobación o de promedio académico, y abandonan sus estudios a una edad académica temprana. Siendo una Institución privada, el tema de apoyo financiero es importante, aunque no se tiene información recolectada a la salida de la Institución, solamente se cuenta con el motivo de baja, y muchas veces el estudiante simplemente no se inscribe y nunca realiza su proceso de baja formal.

Se recomienda a la Institución recabar y consolidar la información que un estudiante genera desde que es un prospecto y durante su estancia en la Institución en un repositorio de bases de datos, e implementar una encuesta de salida en donde se pueda recoger información cualitativa sobre la experiencia del estudiante en la Institución tomando como referencia algunas de las variables consideradas por los teóricos de la deserción (tabla 3-1).

Es muy importante reconocer que hay un fenómeno de migración estudiantil a otros campus del mismo Sistema CETYS, ya que a pesar de que se permanece en el Sistema, al final representa una baja que impacta a los indicadores del campus de origen, como son las metas de crecimiento, la eficiencia terminal, los programas de acreditación y finalmente en su presupuesto. En México el sistema multicampus es común para las IES públicas y privadas, por lo que es un fenómeno de interés público y de gran relevancia educativa.

Los estudios de deserción universitaria coinciden en muchas de sus predicciones, sin embargo, es muy difícil atribuirle esta capacidad predictiva de forma absoluta a un grupo específico de variables, por lo que seguramente será motivo de futuros estudios, sin embargo, identificar estas variables predictoras en este caso de estudio, permite enfatizar en la relevancia del examen de admisión y los resultados de cada área de evaluación, el acompañamiento a los

estudiantes en los primeros años de estudio, y el aseguramiento de que las competencias no demostradas puedan alcanzarse en estos primeros semestres.

En el futuro se podrían analizar datos cualitativos recabados al ingreso y al momento de salir de la Institución para contar con más elementos que permitan revisar este fenómeno, además de estudiar la migración multicampus que se observó al momento de clasificar a los desertores de los que permanecieron en el Sistema CETYS, pero en otro campus, si lograron o no graduarse o simplemente desertaron tiempo después.

Referencias

Abdelrahman, N., & Farah, A. (2018). Big Data: A Review. *International Journal of Computer and Information Technology*, *7*(3), 2279–0764. www.ijcit.com

Adelman, M., & Székely, M. (2016). *School Dropout in Central America An Overview of Trends, Causes, Consequences, and Promising Interventions*. http://econ.worldbank.org.

Alban, M., & Mauricio, D. (2019). Predicting University Dropout through Data Mining: A Systematic Literature. *Indian Journal of Science and Technology*, *12*(4), 10. https://doi.org/10.17485/ijst/2019/v12i4/139729

Aljohani, O. (2016). A Comprehensive Review of the Major Studies and Theoretical Models of Student Retention in Higher Education. *Higher Education Studies*, *6*(2), 1–18. https://doi.org/10.5539/hes.v6n2p1

Alnoukari, M. and Hanano, A. (2017). Integration of business intelligence with corporate strategic management. *Journal of Intelligence Studies in Business*, *7*(2), 5–16.

Andrea, P., Cardona, N., Adolfo, G., Brand, Q., Armando, L., & Marín, U. (2016). Visualizaciones analíticas para la toma de decisiones en pequeñas y medianas empresas utilizando Data Mining *. *Cuaderno Activa*, *8*, 31–39.

Apache Hive TM. (n.d.). Retrieved November 24, 2020, from https://hive.apache.org/index.html

Astin, A. W. (1999). Student involvement: A developmental theory for higher education. *Journal of College Student Development*, *25*(4), 297–308. https://doi.org/10.1016/0263-

Astin, A. W. (1999). Student Involvement: A Developmental Theory for Higher Education. *Journal of College Student Development*, *40*(5), 518–529.

Bāliņa, S., Žuka, R., & Krasts, J. (2016). Opportunities for the Use of Business Data Analysis Technologies. *Economics and Business*, *28*(1), 20–26. https://doi.org/10.1515/eb-2016-0003

Bean, J. P. (1985). Interaction Effects Based on Class Level in an Explanatory Model of College Student Dropout Syndrome. *American Educational Research Journal*, *22*(1), 35–64. https://doi.org/10.2307/1162986

Bean, J. P. ., & Metzner, B. S. (1985). *A Conceptual Model of Nontraditional Undergraduate Student Attrition Author*. *55*(4), 485–540. https://doi.org/10.2307/1170245

Berger, J. B., Blanco Ramirez, G., & Lyons, S. (2012). A Historical Look at Retention. In *College Student Retention : Formula for Student Success* (pp. 7–34). Lanham: Rowman & Littlefield Publishers. http://ezproxy.upaep.mx:2062/login.aspx?direct=true&db=nlebk&AN=437448&lang=es&site=eds-live

Board, C. (2014). *Prueba de Aptitud Académica*.

Boticario, J. G., Santos, O. C., Romero, C., Pechenizkiy, M., Merceron, A., Mitros, P., Luna, J.

M., Mihaescu, C., Moreno, P., Hershkovitz, A., & Ventura, S. (2015). *Educational Data Mining 2015: 8th International Conference on Educational Data Mining*. *June*.

Brett Lanz. (2019). *Machine Learning with R : Expert Techniques for Predictive Modeling* (Third edit). https://web.s.ebscohost.com/ehost/ebookviewer/ebook/ZTAwMHh3d19fMjEwNjMwNF9fQU41?sid=8856667c-38d4-47a8-9b02-4d6eac2c3d26@redis&vid=4&format=EB&rid=3

Brinkmann, D. (2015). *Strategic capability through business intelligence applications Daniel Brinkmann A thesis submitted to The University of Gloucestershire in accordance with the requirements of the degree of Doctor of Business Administration in the Faculty of Business , Ed*. *August*.

Cabrera, A. F., Nora, J. A., & Castaneda, M. B. (1993). College Persistence : Structural Equations Modeling Test of an Integrated Model of Student Retention. *The Journal of Higher Education*, *64*(2), 123–139. https://www.jstor.org/stable/2960026 Accessed:

Casella, G., Fienberg, S., & Olkin, I. (2013). An Introduction to Statistical Learning. In *Springer Texts in Statistics*. https://doi.org/10.1016/j.peva.2007.06.006

CETYS. (2011). *Plan de Desarrollo CETYS 2020*.

CETYS. (2018a). *SICU: Reporte de retención y eficiencia terminal por campus*.

CETYS, S. de I. (2018b). *Cohor Graduation, Retention and Transfer Rates by Programas Undergraduate*.

CETYS Universidad. (2014). *Reglamento-CETYS-Alumnos-Profesional-2017-2018*. 2414.

Cu Balán, G. (2008). El impacto de la escuela de procedencia del Nivel Medio Superior en el desempeño de los alumnos en el Nivel Universitario. *Revista Electrónica Iberoamericana Sobre Calidad Eficacia y Cambio En Educación*, *6*(2), 59–99. https://doi.org/2152

Curto Diaz, J. (2016). *introducción al business intelligence* (UOC). https://elibro.net/es/ereader/cetys/101030?prev=bf&page=73

Data Integration | IBM. (n.d.). Retrieved November 24, 2020, from https://www.ibm.com/analytics/data-integration?p1=Search&p4=p50329209090&p5=e&cm_mmc=Search_Google-_-1S_1S-_-WW_NA-_-ibm data integration_e&cm_mmca7=71700000060952642&cm_mmca8=kwd-323927610430&cm_mmca9=Cj0KCQjwmpb0BRCBARIsAG7y4zZqzhyS84ta3HYTkQGTVomG9Z4BF

Devasia, T., Vinushree, T. P., & Hegde, V. (2016). Prediction of students performance using Educational Data Mining. *Proceedings of 2016 International Conference on Data Mining and Advanced Computing, SAPIENCE 2016*, 91–95. https://doi.org/10.1109/SAPIENCE.2016.7684167

Diaz Pedroza, K. Y., Chindoy Chasoy, B. Y., & Rosado Gómez, A. A. (2019). Review of techniques, tools, algorithms and attributes for data mining used in student desertion. *Journal of Physics: Conference Series*, *1409*(1). https://doi.org/10.1088/1742-6596/1409/1/012003

Donoso, S., & Schiefelbein, E. (2007). ANÁLISIS DE LOS MODELOS EXPLICATIVOS DE RETENCIÓN DE ESTUDIANTES EN LA UNIVERSIDAD : UNA VISIÓN DESDE LA DESIGUALDAD SOCIAL. *Estudios Pedagógicos*, *XXXIII*(1), 7–27. https://www.redalyc.org/articulo.oa?id=173514133001

Eccles, J. S., & Wigfield, A. (2002). MOTIVATIONAL BELIEFS, VALUES, AND GOALS. *Annual Review of Psychology*, *53*(1), 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Eckert, K. B., & Suénaga, R. (2015). Analysis of Attrition-Retention of College Students Using Classification Technique in Data Mining. *Formación Universitaria*, *14*(5), 3–12. https://doi.org/10.4067/S0718-50062015000500002

Education at a Glance 2021: OECD Indicators. (2021). OECD Publishing. https://doi.org/10.1787/b35a14e5-en

ENCUESTA SATISFACCIÓN 2017-2 CAMPUS ENSENADA. (2017).

Ethington, C. A. (1990). A PSYCHOLOGICAL MODEL OF STUDENT PERSISTENCE. *Research in Higher Education*, *31*(3), 279–293.

Fayyad, U., PiatetskySaphiro, G., & Smyth, P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications of the ACM*, *39*(11), 27–34. http://delivery.acm.org/10.1145/250000/240464/p27-fayyad.pdf?ip=189.204.168.59&id=240464&acc=ACTIVE SERVICE&key=4D4702B0C3E38B35.4D4702B0C3E38B35.CFB69864B5C19376.4D4702B0C3E38B35&__acm__=1541901442_f0748bc05eac91ea24de97fb8c908b92

Ferri Ramirez, C., & Ramirez Quintana, M. J. (2011). *Introducción a la Minería de Datos* (P. Educacion (Ed.)).

Fonseca, G., & García, F. (2016). Permanencia y abandono de estudios en estudiantes universitarios: un análisis desde la teoría organizacional. *Revista de La Educacion Superior*, *45*(179), 25–39. https://doi.org/10.1016/j.resu.2016.06.004

Fu, K. S. (1968). *Sequential Methods in Pattern Recognition and Machine Learning*. Academic Press. http://ezproxy.upaep.mx:2048/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=e000xww&AN=297029&lang=es&site=eds-live

Ge, Z., Song, Z., Ding, S. X., & Huang, B. (2017). Data Mining and Analytics in the Process Industry: The Role of Machine Learning. *IEEE Access*, *5*, 20590–20616. https://doi.org/10.1109/ACCESS.2017.2756872

Gironés-Roig, J., Casas-Roma, J., Minguillón-Alfonso, J., & Caihuelas-Quiles, R. (2017). Minería de datos: Modelos y Algoritmos. In *Manuales <<Tecnología>> Editorial UOC* (Issue 1). https://www.cambridge.org/core/product/identifier/CBO9781139058452A007/type/book_part

Gironés, J., Casas, J., & Minguillón, J. (2017). *Minería de datos: modelos y algoritmos*.

Editorial UOC. http://ebookcentral.proquest.com/lib/cetyssp/detail.action?docID=5045398

Goldman, B. L. (2015). El Big Data y la Analítica de Negocios en el capitalismo. *XI Jornadas de Sociología*. http://www.aacademica.org/000-061/993%0AActa

Gutiérrez, J. A., & Molina, B. (2016). Identificación de técnicas de minería de datos para apoyar la toma de decisiones en la solución de problemas empresariales. *Revista Ontare*, *3*(2), 33–51. https://doi.org/10.21158/01208160.N2.2015.1440

Hastie, T., Tibshirani, R., & Friedman, J. (2013). The Elements of Statistical Learning: Data Mining, Inference and Prediction. In 2009 Springer (Ed.), *Springer* (Second). Springer. https://doi.org/10.1109/SITIS.2013.106

Hernández, A. R., Melendez, L. A., Morales, A., Garcia, A., Tecpanecatl, J. L., & Algredo, I. (2016). Comparative Study of Algorithms to Predict the Desertion in the Students at the ITSM-Mexico Artificial Intelligence in Education View project Dependable ESB systems based on self-healing and checkpointing principles View project. *IEEE LATIN AMERICA TRANSACTIONS*, *14*(11), 4573–4578. https://doi.org/10.1109/TLA.2016.7795831

Huesca Ramírez, M. G. E., & Castaño Corvo, M. B. (2007). Causas de Deserción de Alumnos de Primeros Semestres de una Universidad Privada. *Revista Mexicana de Orientación Educativa*, *V*(12), 34–40. http://www.alfaguia.org/alfaguia/files/1319582164causas de desercion en una universidad privada.pdf

INEGI. (2021). *Tasa de abandono escolar por entidad federativa según nivel educativo, ciclos escolares seleccionados de 2000/2001 a 2020/2021*. https://www.inegi.org.mx/app/tabulados/interactivos/?pxq=9171df60-8e9e-4417-932e-9b80593216ee

Khademolqorani, S., & Hamadani, A. Z. (2013). An Adjusted Decision Support System through Data Mining and Multiple Criteria Decision Making. *Procedia - Social and Behavioral Sciences*, *73*, 388–395. https://doi.org/10.1016/j.sbspro.2013.02.066

Lee, P. M. (2013). Use Of Data Mining In Business Analytics To Support Business Competitiveness. *The Review of Business Information Systems (Online)*, *17*(2), 53. http://search.proquest.com/docview/1418721911?accountid=8144%5Cnhttp://sfx.aub.aau.dk/sfxaub?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&genre=article&sid=ProQ:ProQ:abiglobal&atitle=Use+Of+Data+Mining+In+Business+Analytics+To+Support+Busi

Limsathitwong, K., Tiwatthanont, K., & Yatsungnoen, T. (2018). Dropout prediction system to reduce discontinue study rate of information technology students. *Proceedings of 2018 5th International Conference on Business and Industrial Research: Smart Technology for Next Generation of Information, Engineering, Business and Social Science, ICBIR 2018*, 110–114. https://doi.org/10.1109/ICBIR.2018.8391176

Londoño, L. (2013). Factores de riesgo presentes en la deserción estudiantil en la Corporación Universitaria Lasallista. *Revista Virtual Universidad Católica Del Norte*, *38*, 183–194.

Louppe, G. (2014). *Understanding Random Forests: From Theory to Practice*. http://arxiv.org/abs/1407.7502

Manyanga, F., Sithole, A., & Hanson, S. M. (2017). Comparison of Student Retention Models in Undergraduate Education From the Past Eight Decades. *Journal of Applied Learning in Higher Education*, *7*(Cdm), 30–42. http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1188373&lang=es&site=eds-live

Marco General de Referencia para los Procesos de Acreditación de Programas Académicos de Tipo Superior ver 3.0. (2016). https://www.copaes.org/assets/docs/Marco-de-Referencia-V-3.0_.pdf

Miranda, M. A., & Guzmán, J. (2017). Análisis de la Deserción de Estudiantes Universitarios usando Técnicas de Minería de Datos. *Formación Universitaria*, *10*(3), 61–68. https://doi.org/10.4067/S0718-50062017000300007

Mishra, B. K., Hazra, D., Tarannum, K., & Kumar, M. (2016). Business Intelligence using Data Mining techniques and Business Analytics. *2016 International Conference System Modeling & Advancement in Research Trends (SMART), January 2016*, 84–89. https://doi.org/10.1109/SYSMART.2016.7894496

Morrison, L., & Silverman, L. (2012). Retention Theories, Models, and Concepts. In *College Student Retention : Formula for Student Success* (pp. 61–80). Lanham: Rowman & Littlefield Publishers. http://ebookcentral.proquest.com

MySQL. (n.d.). Retrieved November 24, 2020, from https://www.mysql.com/

Navarro Charris, N. E., Redondo Bilbao, O. E., Contreras Salinas, J. A. |, Romero Diaz, C. H., & D.Andreis Zapata, A. C. (2017). Permanencia y deserción versus autoeficacia de estudiantes universitarios: Un desafío de la calidad educativa. *Revista Lasallista de Investigacion*, *14*(1), 198–206. https://doi.org/10.22507/rli.v14n1a17

Nwanganga, Fred Chapple, M. (2020). *Practical Machine Learning in R*. John Wiley & Sons. https://ebiblio.cetys.mx:2118/web/view/pdf/show.v/rcid:kpPMLR0005/cid:kt012EWJK1/viewerType:pdf//root_slug:5-logistic-regression/url_slug:logistic-regression?cid=kt012EWJK1&b-toc-cid=kpPMLR0005&b-toc-title=Practical Machine Learning in R&b-toc-url-slug=lo

OECD. (2019). *Educación superior en México: Resultados y relevancia para el mercado laboral*. OECD. https://doi.org/10.1787/A93ED2B7-ES

Pascarella, E. T., Smart, J. C., & Ethington, C. A. (1986). Long-Term Persistence of Two-Year College Students. *Association for the Study of Higher Education*.

Pascarella, E. T., & Terenzini, P. T. (1980). Model Predicting Freshman Persistence and Voluntary Dropout Decisions from a Theoretical Model. *The Journal of Higher Education*, *51*(1), 60–75. https://doi.org/10.2307/1981125

Patiño, L. & Cardona, A. (2012). Revisión de algunos estudios sobre la deserción estudiantil universitaria en Colombia y latinoámerica. *Revista Theoria*, *21*(1), 9–20.

Peralta, B., Poblete, T., & Caro, L. (2017). Automatic feature selection for desertion and graduation prediction: A chilean case. *Proceedings - International Conference of the*

Chilean Computer Science Society, SCCC. https://doi.org/10.1109/SCCC.2016.7836055

Pérez, A., Grandón, E. E., Caniupán, M., & Vargas, G. (2018). *Análisis Comparativo de Técnicas de Predicción para Determinar la Deserción Estudiantil: Regresión Logística vs Árboles de Decisión*. https://www.sap.com/latinamerica/products/predictive-

Piatetsky-Shapiro, G., Djeraba, C., Getoor, L., Grossman, R., Feldman, R., & Zaki, M. (2006). What Are The Grand Challenges for Data Mining? *SIGKDD Explorations*, 70–77. http://delivery.acm.org/10.1145/1240000/1233330/p70-piatetsky-shapiro.pdf?ip=189.204.168.59&id=1233330&acc=ACTIVE SERVICE&key=4D4702B0C3E38B35.4D4702B0C3E38B35.CFB69864B5C19376.4D4702B0C3E38B35&__acm__=1541901466_f441410a34530369bb829767e90e5daa

Proano, J. P. Z., & Villamar, V. C. P. (2018). Systematic mapping study of literature on educational data mining to determine factors that affect school performance. *Proceedings - 3rd International Conference on Information Systems and Computer Science, INCISCOS 2018, 2018-Decem*, 239–245. https://doi.org/10.1109/INCISCOS.2018.00042

Quiénes somos - CETYS Universidad. (n.d.). Retrieved December 6, 2018, from http://www.cetys.mx/quienes-somos/

R: What is R? (n.d.). Retrieved November 25, 2020, from https://www.r-project.org/about.html

Ramírez, P. E., & Grandón, E. E. (2018). Predicción de la Deserción Académica en una Universidad Pública Chilena a través de la Clasificación basada en Árboles de Decisión con Parámetros Optimizados. *Formación Universitaria*, *11*(3), 3–10. https://doi.org/10.4067/S0718-50062018000300003

Reglamento de becas y descuentos del Sistema CETYS Universidad (pp. 1–17). (2010).

Reyes Dixson, Y., & Nuñez Maturel, L. (2015). La inteligencia de negocio como apoyo a la toma de decisiones en el ámbito académico. *Revista Internacional de Gestión Del Conocimiento y La Tecnología*, *3*(2), 63–73. https://doi.org/10.1007/BF01530759

Rodriguez Maya, N. E., Jimenez Alfaro, A. J., Reyes Hernandez, L. A., Suarez Carranza, B. A., & Ruiz Garduno, J. K. (2017). *Data mining: a scholar dropout predictive model*. *186*, 89–93. https://doi.org/10.1109/mhtc.2017.8006421

Rosasco, L. (2017). *Introductory Machine Learning Notes*. http://lcsl.mit.edu/courses/ml/1718/MLNotes.pdf

Secretaria de Educación Pública. (2018). *Sistema Educativo de los Estados Unidos Mexicanos, Principales Cifras 2016-2017*.

Secretaria de Educación Pública. (2020). Principales cifras, Sistema Educativo de los Estados Unidos Mexicanos. *Secretaría de Educación Pública*, 288. http://planeacion.sep.gob.mx/assets/images/principales_cifras/PRINCIPALESCIFRAS14_15.pdf

Shalev-Shwartz, S., & Ben-David, S. (2013). Understanding machine learning: From theory to algorithms. In *Understanding Machine Learning: From Theory to Algorithms* (Vol. 9781107057). https://doi.org/10.1017/CBO9781107298019

Sosa, J. (2016). *Evaluación del nivel de madurez en los procesos de administración de información sobre indicadores de calidad educativa en una institución de educación superior desde la perspectiva de los procesos de acreditación*. Universidad Popular Autónoma del Estado de Puebla.

Spady, W. G. (1970). *Dropouts from Higher Education: An Interdisciplinary Review and Synthesis*. http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ024800&site=ehost-live

SPSS Modeler - Overview | IBM. (n.d.). Retrieved November 25, 2020, from https://www.ibm.com/products/spss-modeler

SPSS Software | IBM. (n.d.). Retrieved August 19, 2021, from https://www.ibm.com/analytics/spss-statistics-software

Tableau Prep Builder y Tableau Prep Conductor: Una solución de preparación de datos de autoservicio. (n.d.). Retrieved November 23, 2020, from https://www.tableau.com/es-es/products/prep#Builder

Tamez, R., Zúñiga, L., & Martínez, F. (2006). *SISTEMA DE INDICADORES EDUCATIVOS DE LOS ESTADOS UNIDOS MEXICANOS* (Primera). Dirección General de Planeación y programación, Secretaria de Educación Pública. http://www.snie.sep.gob.mx/descargas/indicadores/SININDE.pdf

Tinto, V. (1975). Dropout from Higher Education : A Theoretical Synthesis of Recent Research. *Review of Educational Research*, *45*(1), 89–125. http://search.ebscohost.com/login.aspx?direct=true&db=edsjsr&AN=edsjsr.1170024&lang=es&site=eds-live

Tinto, V. (1989). Definir la desercion: Una cuestion de perspectiva. *Revista de Educación Superior*, *18*(71), 160. https://doi.org/10.1017/CBO9781107415324.004

Tollenar, N., & Van der Heijden, M. (2013). Which method predicts recidivism best ?: a comparison of statistical , machine learning and data mining predictive models Author (s): N . Tollenaar and P . G . M . van der Heijden Published by : Wiley for the Royal Statistical Society Stable URL : http: *12 Royal Statistical*, *176*(2), 565–584.

Urbina-Nájera, A. B., Camino-Hampshire, J. C., & Cruz Barbosa, R. (2020). Deserción escolar universitaria: Patrones para prevenirla aplicando minería de datos educativa. *RELIEVE - Revista Electrónica de Investigación y Evaluación Educativa*, *26*(1). https://doi.org/10.7203/relieve.26.1.16061

Urbina, A. B., & De la Calleja, J. (2017). Brief review of educational applications using data mining and machine learning. *Revista Electronica de Investigacion Educativa*, *19*(4), 84–96. https://doi.org/10.24320/redie.2017.19.4.1305

Vargas, H. (2015). *Conceptos retencion y eficiencia terminal en CETYS*.

Velayutham, S. (2020). 18.2.15.2 Unsupervised Learning. In *Handbook of Research on Applications and Implementations of Machine Learning Techniques* - (pp. 350–359). IGI

Global. https://app.knovel.com/hotlink/pdf/id:kt01262XM1/handbook-research-applications/introducti-partitioning BT - Handbook of Research on Applications and Implementations of Machine Learning Techniques

Vera Noriega, J. A., Ramos Estrada, D. Y., Sotelo Castillo, M. A., Echeverria Castro, S., Serrano Encinas, D. M., & Vales García, J. J. (2012). Factores asociados al rezago en estudiantes de una institución de educación superior en México. *Revista Iberoamericana de Educación Superior*, *III*(7), 41–56.

Weka 3 - Data Mining with Open Source Machine Learning Software in Java. (n.d.). Retrieved November 25, 2020, from https://www.cs.waikato.ac.nz/ml/weka/

What Is ETL? | SAS. (n.d.). Retrieved November 24, 2020, from https://www.sas.com/en_us/insights/data-management/what-is-etl.html

Why RapidMiner | RapidMiner. (n.d.). Retrieved November 25, 2020, from https://rapidminer.com/why-rapidminer/

Wright, M. N., & Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1). https://doi.org/10.18637/JSS.V077.I01

Anexo 1. Diccionario de datos

Nombre de la tabla: Desertores

Descripción: Tabla de hechos, alumnos que desertaron.

Campo	Tamaño	Tipo de Dato	Descripción
matricula	7	String	Identificador único para cada estudiante.
cve_nivel	4	String	Identificador de la escuela a la que pertenece el programa académico.
cve_programa	6	String	Identificador del programa académico
cve_plan	7	String	Identificador del plan de estudios correspondiente al programa académico.
cohorte	4	integer	Cohorte del estudiante
fecha_inscripcion	10	date	Fecha de la última inscripción.
fecha_baja	10	date	Fecha en la que el estudiante se dio de baja.

Nombre de la tabla: Finanzas

Descripción: Tabla de apoyos financieros

Campo	Tamaño	Tipo de Dato	Descripción
matricula	7	String	Identificador único para cada estudiante
cve_periodo	10	String	Periodo de inicio de la beca
cve_beca	10	String	Tipo de apoyo financiero
monto	4	integer	Cantidad porcentual de beca

Nombre de la tabla: Apoyos_financieros

Descripción: Claves de apoyos financieros

Campo	Tamaño	Tipo de Dato	Descripción
cve_beca	10	String	Clave del tipo de apoyo financiero
descripción	30	String	Descripción del apoyo financiero

Nombre de la tabla: Estatus_alumnos

Descripción: Tabla que muestra el estatus del estudiante inscrito o dado de baja.

Campo	Tamaño	Tipo de Dato	Descripción
matricula	7	String	Identificador único para cada estudiante.
cve_nivel	4	String	Identificador de la escuela a la que pertenece el programa académico.
cve_programa	6	String	Identificador del programa académico
cve_periodo	20	String	Período académico: ordinario, extraordinario, verano, invierno, intercambio
semestre	4	String	Semestre en el que se encuentra el estudiante
motivo_baja	5	String	Clave de la razón de la baja
fecha_baja	10	date	Fecha en la que el estudiante se dio de baja.

Nombre de la tabla: Motivosbaja

Descripción: Descripciones de los motivos de baja

Campo	Tamaño	Tipo de Dato	Descripción
cve_baja	6	String	Clave del motivo de la baja del estudiante.
descripción	30	String	Descripción del motivo de baja

Nombre de la tabla: Programas

Descripción: Descripciones de los programas académicos

Campo	Tamaño	Tipo de Dato	Descripción
cve_nivel	5	Integer	Clave de la escuela
cve_programa	5	String	Clave del programa académico
descripción	30	String	Descripción programa académico.

Nombre de la tabla: Planes de estudio

Descripción: Descripciones de los planes de estudio

Campo	Tamaño	Tipo de Dato	Descripción
nivel	5	Integer	Clave de la escuela
cve_programa	5	String	Clave del programa académico
planes_estudio	10	String	Clave de planes de estudio
descripción	30	String	Descripción del plan de estudio.

Nombre de la tabla: Niveles

Descripción: Descripciones de las escuelas de educación superior.

Campo	Tamaño	Tipo de Dato	Descripción
cve_nivel	4	integer	Clave de la escuela.
descripción	30	String	Descripción de la escuela.

Nombre de la tabla: Periodos.

Descripción: Descripciones de los períodos académicos.

Campo	Tamaño	Tipo de Dato	Descripción
cve_periodo	20	String	Clave del período académico.
descripción	30	String	Descripción período académico.

Nombre de la tabla: Perfil_ingreso

Descripción: Estatus de los alumnos prospectos a ingresar a la Institución.

Campo	Tamaño	Tipo de Dato	Descripción
matricula	7	String	Identificador único para cada estudiante.
sexo	1	Char	Género del estudiante.
escuela_procedencia	30	String	Nombre de la escuela de procedencia.
ciudad_escuela_procedencia	30	String	Nombre de la ciudad de la escuela de procedencia.
tipo_prospecto	4	String	Si el estudiante es foráneo o no.
cve_periodo	20	String	Periodo en la que estudiante ingresa a la institución
promedio	6	Integer	Calificación promedio del nivel académico anterior
cve_programa	6	String	Identificador del programa académico al que desea ingresar.
nivel_socioeconomico	4	Integer	Nivel socioeconómico de ingreso (se toma en cuenta ingreso mensual familiar, zona de la vivienda, tipo de trabajo del padre o tutor)
carrera_interes1	30	String	Primera opción de carrera.
carrera_interes2	30	String	Segunda opción de carrera.

Nombre de la tabla: Resultados_admision.

Descripción: Tabla de resultados de cada una de las áreas del conocimiento evaluadas en el examen de admisión.

Campo	Tamaño	Tipo de Dato	Descripción
matricula	7	String	Identificador único para cada estudiante
fecha_examen	12	date	Fecha de presentación del exámen de admisión.
resultado_RM_examen	6	integer	Resultado de razonamiento matemático.
resultado_RV_examen	6	integer	Resultado de razonamiento verbal.
resultado_REI_examen	6	integer	Resultado de razonamiento español e inglés.

resultado_integrado_examen	6	Integer	Resultado integrado
estado_examen	10	String	Aprobado o no.

Nombre de la tabla: Desempeño_academico.

Descripción: Tabla información de reprobación por área de competencia, cambios de programas y presentación de extraordinarios por período.

Campo	Tamaño	Tipo de Dato	Descripción
matricula	7	String	Identificador único para cada estudiante
cve_periodo	20	String	Periodo en la que estudiante ingresa a la institución
cve_plan	7	String	Identificador del plan de estudios correspondiente al programa académico.
reprobadas	4	integer	Numero de materias reprobadas
area	4	integer	Clave del área de competencia reprobada según la tipología del programa académico
cambio_programa	2	String	Si el estudiante tuvo un cambio de programa académico durante su estancia en la IES
Extraordinarios	2	Integer	Numero de materias presentadas en extraordinario

Nombre de la tabla: Área de formación.

Descripción: Tabla áreas de formación según la tipología de competencias.

Campo	Tamaño	Tipo de Dato	Descripción
area	4	Integer	Clave del área de competencia reprobada según la tipología del programa académico
descripción	20	String	Área de competencia según la tipología del programa académico

Nombre de la tabla: Satisfacción.

Descripción: Tabla de resultados de la evaluación semestral de satisfacción del estudiante.

Campo	Tamaño	Tipo de Dato	Descripción
area	4	Integer	Clave del área de competencia reprobada según la tipología del programa académico
descripción	20	String	Área de competencia según la tipología del programa académico

Nombre de la tabla: Categorías

Descripción: Tabla de categorías o áreas de la evaluación de satisfacción.

Campo	Tamaño	Tipo de Dato	Descripción
cve_categoria	4	Integer	Clave de la categoría o área de la evaluación de satisfacción
descripción	20	String	Descripción de la categoría o área de evaluación de satisfacción

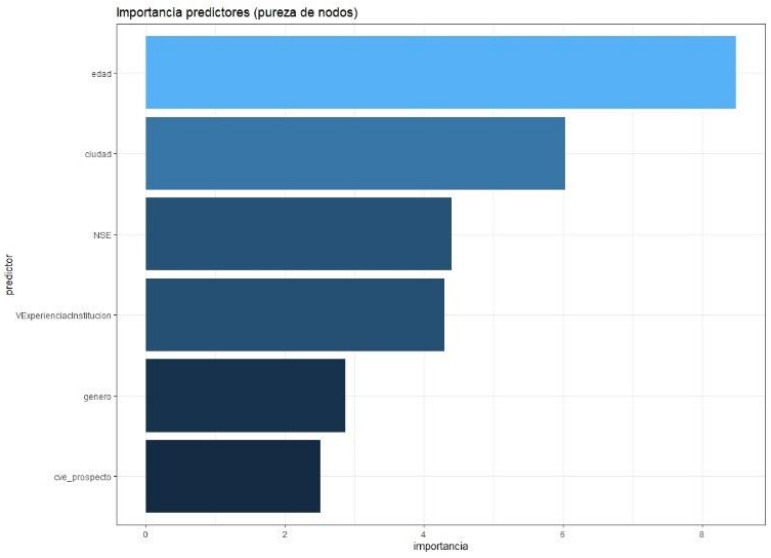
Nombre de la tabla: Subcategorías

Descripción: Tabla de subcategorías de la evaluación de satisfacción.

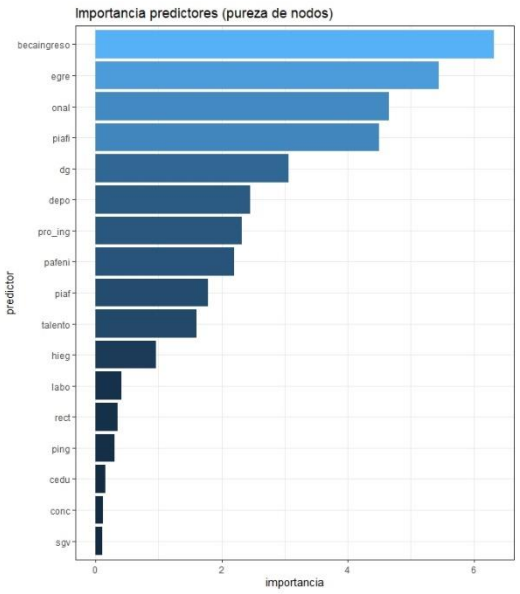
Campo	Tamaño	Tipo de Dato	Descripción
cve_categoria	4	String	Clave de la categoría o área de la evaluación de satisfacción
cve_subcategoria	4	String	Clave de la subcategoría de evaluación
descripción	20	String	Descripción de la subcategoría de evaluación de satisfacción

Anexo 2. Construcción de modelos y gráficos durante el procesamiento de datos

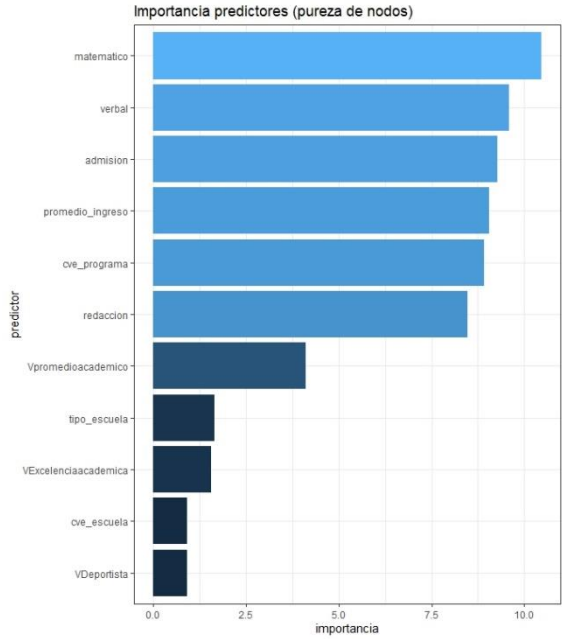
a) Al ingresar a la IES



Predictores importantes del perfil personal del desertor al ingresar a la Universidad

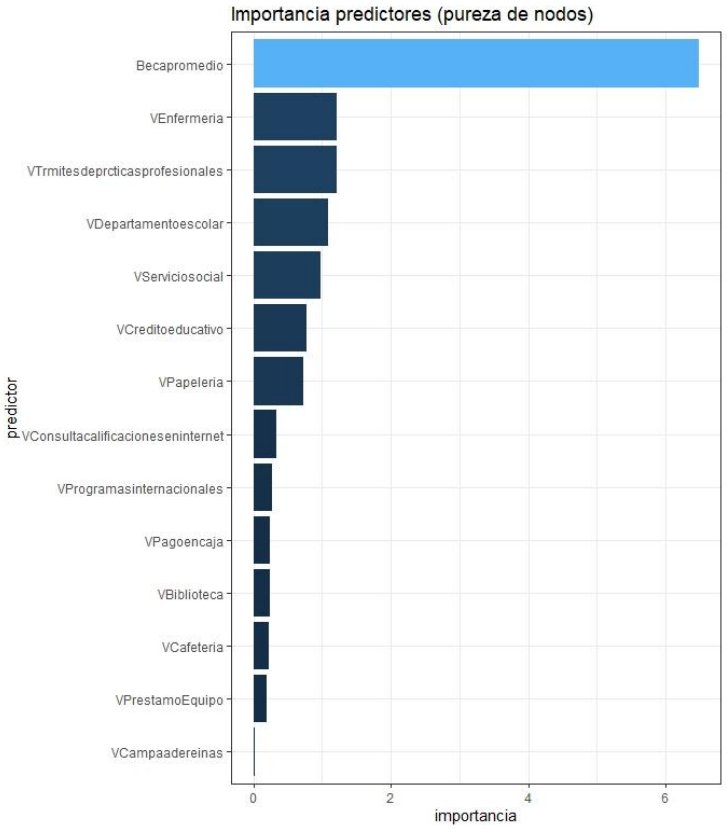


Predictores importantes de la institución hacia el desertor al ingresar a la Universidad

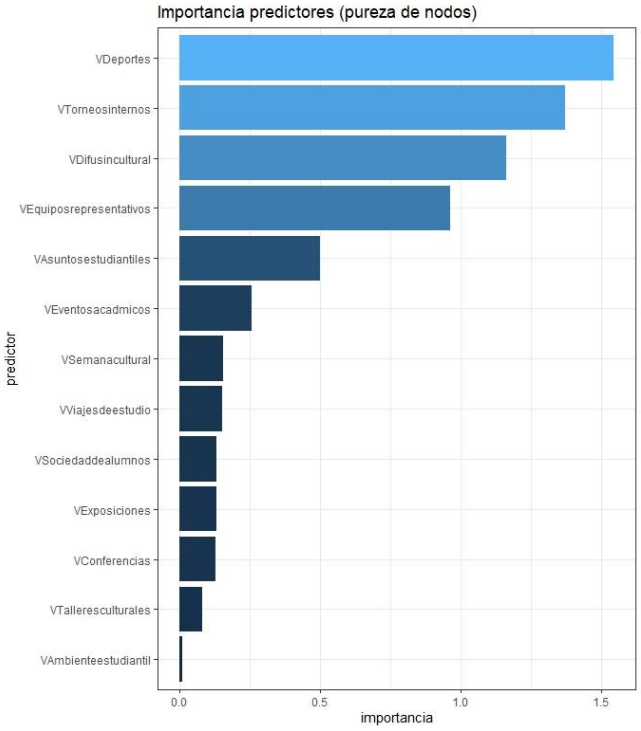


Predictores importantes del perfil académico del desertor al ingresar a la Universidad

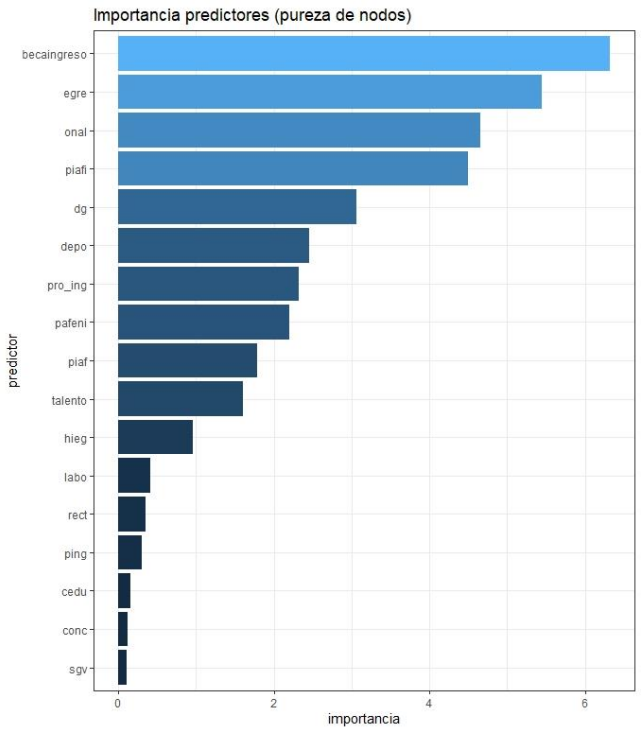
b) Durante su estancia en la IES



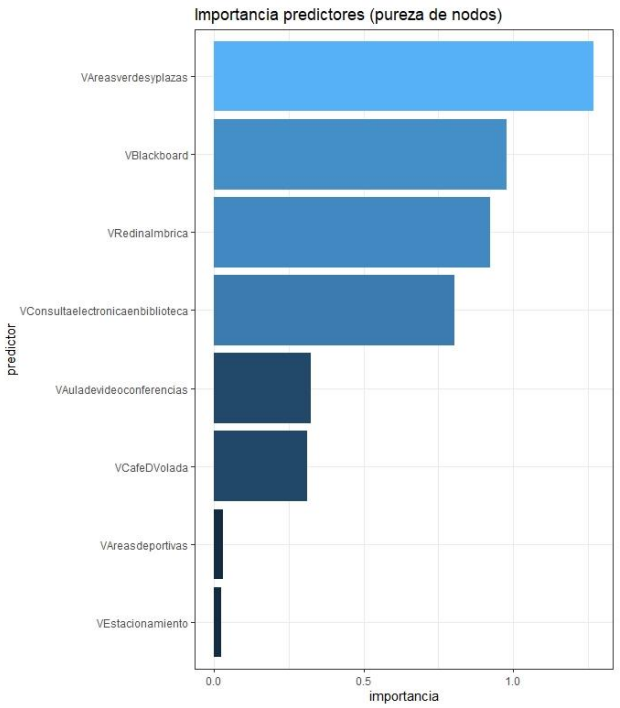
Predictores importantes de los servicios de la institución desde la perspectiva del desertor durante su estancia en la Universidad



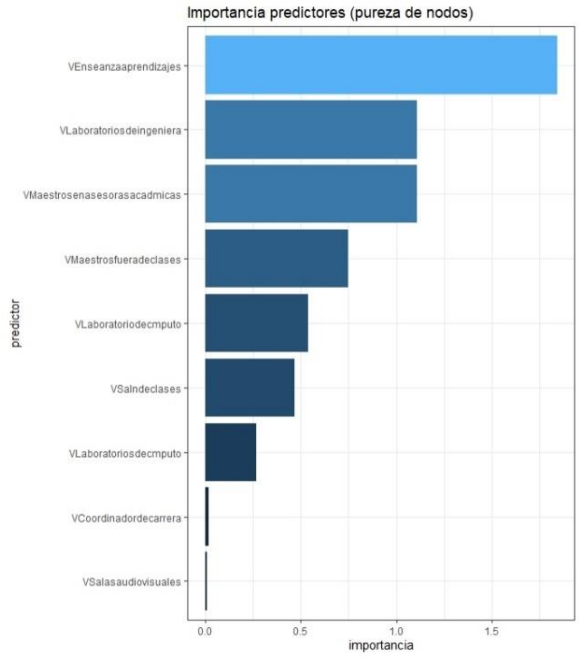
Predictores importantes de las actividades extracurriculares desde la perspectiva del desertor durante su estancia en la Universidad



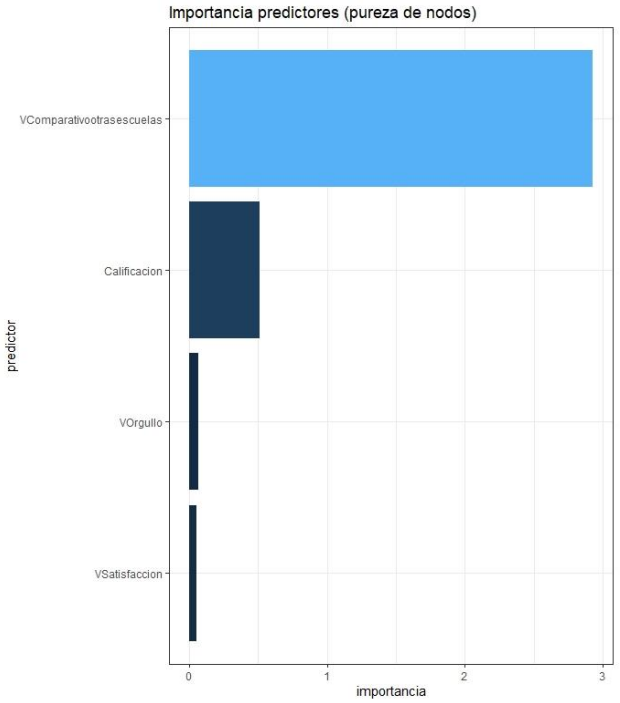
Predictores importantes financieras de la institución desde la perspectiva del desertor durante su estancia en la Universidad



Predictores importantes de la infraestructura de la institución desde la perspectiva del desertor durante su estancia en la Universidad



Predictores importantes del servicio educativo de la institución desde la perspectiva del desertor durante su estancia en la Universidad



Predictores importantes de la satisfacción del desertor durante su estancia en la Universidad

	MeanDecreaseGini
edad	6.290309
ciudad	5.215734
matematico	14.949164
verbal	12.554088
admision	13.241628
promedio_ingreso	12.681663

Figura 2 Modelo 2 RF

	MeanDecreaseGini
becaingreso	10.348847
egre	5.992433
piafi	5.035742
dg	3.897527
Becapromedio	13.947360

Figura 1 Modelo 1 RF

	MeanDecreaseGini
VEnseanzaaprendizajes	1.3807419
VLaboratoriosdeingeniera	0.7726758
VDeportes	1.3212317
VTorneosinternos	1.4456116
VComparativootrasescuelas	1.0314582
VAreasverdesyplazas	0.2395091
VBlackboard	0.3578002

Figura 4 Modelo 3 RF

	MeanDecreaseGini
matematico	10.486391
verbal	10.023669
admision	9.677519
promedio_ingreso	9.214012
Becapromedio	9.055131
becaingreso	7.099659
egre	4.882031
piafi	3.272594

Figura 3 Modelo 4 RF

	MeanDecreaseGini
cve_programa	9.674925
Promediofinal	8.806715
edad	4.741048
matematico	9.212183
PromedioPeriodoSeleccionado	8.431633
TotalFaltas	7.155741
AniosUniversidad	2.460150
NSE	1.861475
Becapromedio	6.562608
promedio_ingreso	6.794137

Figura 4-5 Modelo 5 RF

Anexo 3. Código fuente en programación R

```
# ************CONSTRUCCIÓN DEL MODELO DE REGRESIÓN LOGÍSTICA

#Directorio de trabajo
setwd("~/LUCIA.BELTRAN.PERSONAL/CDC/DOCTORADO/UPAEP/ACADEMICO/Tesis/Tesis IV")

#Paquetes
library(caTools)
library(ggplot2)
library(Epi)
library(rpart, lib.loc = "C:/Program Files/R/R-4.1.0/library")
library(factoextra)
library(rpart.plot)
library(caret)
library(cowplot)
library(car)
library(ROCR)
library("rfVarImpOOB")
library(texreg)
library(coefplot)
library(effects)
library(AICcmodavg)
library(cvAUC)
library(randomForest)
library(MASS)
library(tidyverse)
library(stats)
library(corrplot)

#********** TRABAJO DE PREPARACIÓN DE DATOS
#set de datos
log.reg =read.csv ('Analisis final desertores MB.csv')
#str(log.reg)

#Convertir las variables categóricas
log.reg$cve_escuela<-as.factor(log.reg$cve_escuela)
log.reg$VExperienciacInstitucion<-as.factor(log.reg$VExperienciacInstitucion)
```

```
log.reg$cve_programa<-as.factor(log.reg$cve_programa)
log.reg$cve_escuelaingreso<-as.character(log.reg$cve_escuelaingreso)
log.reg$tipo_escuela<-as.factor(log.reg$tipo_escuela)
log.reg$cve_prospecto<-as.factor(log.reg$cve_prospecto)
log.reg$genero<-as.factor(log.reg$genero)
log.reg$estatus_admision<-as.factor(log.reg$estatus_admision)
log.reg$NSE<-as.factor(log.reg$NSE)
log.reg$MotivoBaja<-as.factor(log.reg$MotivoBaja)
log.reg$Vpromedioacademico<-as.factor(log.reg$Vpromedioacademico)
log.reg$VDeportista<-as.factor(log.reg$VDeportista)
log.reg$VExcelenciaacademica<-as.factor(log.reg$VExcelenciaacademica)

log.reg$VPrestamoEquipo<-as.factor(log.reg$VPrestamoEquipo)
log.reg$VDepartamentoescolar<-as.factor(log.reg$VDepartamentoescolar)
log.reg$VPagoencaja<-as.factor(log.reg$VPagoencaja)
log.reg$VCreditoeducativo<-as.factor(log.reg$VCreditoeducativo)
log.reg$VBiblioteca<-as.factor(log.reg$VBiblioteca)
log.reg$VCafeteria<-as.factor(log.reg$VCafeteria)
log.reg$Vcafeteria<-as.factor(log.reg$Vcafeteria)
log.reg$VBlackboard<-as.factor(log.reg$VBlackboard)
log.reg$VRedinalmbrica<-as.factor(log.reg$VRedinalmbrica)
log.reg$VConsultaelectronicaenbiblioteca<-as.factor(log.reg$VConsultaelectronicaenbiblioteca)
log.reg$VLaboratoriodecmputo<-as.factor(log.reg$VLaboratoriodecmputo)
log.reg$VPapeleria<-as.factor(log.reg$VPapeleria)
log.reg$VEnfermeria<-as.factor(log.reg$VEnfermeria)
log.reg$VLaboratoriosdecmputo<-as.factor(log.reg$VLaboratoriosdecmputo)
log.reg$VDeportes<-as.factor(log.reg$VDeportes)
log.reg$VDifusincultural<-as.factor(log.reg$VDifusincultural)
log.reg$VMaestrosenasesorasacadmicas<-as.factor(log.reg$VMaestrosenasesorasacadmicas)
log.reg$VMaestrosfueradeclases<-as.factor(log.reg$VMaestrosfueradeclases)
log.reg$VAsuntosestudiantiles<-as.factor(log.reg$VAsuntosestudiantiles)
log.reg$VServiciosocial<-as.factor(log.reg$VServiciosocial)
log.reg$VProgramasinternacionales<-as.factor(log.reg$VProgramasinternacionales)
log.reg$VTrmitesdeprcticasprofesionales<-as.factor(log.reg$VTrmitesdeprcticasprofesionales)
log.reg$VEnseanzaaprendizajes<-as.factor(log.reg$VEnseanzaaprendizajes)
log.reg$VConsultacalificacioneseninternet<-as.factor(log.reg$VConsultacalificacioneseninternet)
log.reg$VAreasdeportivas<-as.factor(log.reg$VAreasdeportivas)
log.reg$VAreasverdesyplazas<-as.factor(log.reg$VAreasverdesyplazas)
log.reg$VAuladevideoconferencias<-as.factor(log.reg$VAuladevideoconferencias)
log.reg$VBaos<-as.factor(log.reg$VBaos)
log.reg$VCafeDVolada<-as.factor(log.reg$VCafeDVolada)
log.reg$VEstacionamiento<-as.factor(log.reg$VEstacionamiento)
log.reg$VLaboratoriosdeingeniera<-as.factor(log.reg$VLaboratoriosdeingeniera)
log.reg$VSalasaudiovisuales<-as.factor(log.reg$VSalasaudiovisuales)
log.reg$VSalndeclases<-as.factor(log.reg$VSalndeclases)
log.reg$VCampaadereinas<-as.factor(log.reg$VCampaadereinas)
log.reg$VConferencias<-as.factor(log.reg$VConferencias)
log.reg$VEquiposrepresentativos<-as.factor(log.reg$VEquiposrepresentativos)
log.reg$VEventosacadmicos<-as.factor(log.reg$VEventosacadmicos)
log.reg$VExposiciones<-as.factor(log.reg$VExposiciones)
log.reg$VSemanacultural<-as.factor(log.reg$VSemanacultural)
log.reg$VSociedaddealumnos<-as.factor(log.reg$VSociedaddealumnos)
log.reg$VTalleresculturales<-as.factor(log.reg$VTalleresculturales)
log.reg$VTorneosinternos<-as.factor(log.reg$VTorneosinternos)
log.reg$VViajesdeestudio<-as.factor(log.reg$VViajesdeestudio)
log.reg$VCoordinadordecarrera<-as.factor(log.reg$VCoordinadordecarrera)
```

```
log.reg$VOrgullo<-as.factor(log.reg$VOrgullo)
log.reg$VAmbienteestudiantil<-as.factor(log.reg$VAmbienteestudiantil)
log.reg$VOrgullo<-as.factor(log.reg$VOrgullo)
log.reg$Calificacion<-as.factor(log.reg$Calificacion)
log.reg$VSatisfaccion<-as.factor(log.reg$VSatisfaccion)
log.reg$VComparativootrasescuelas<-as.factor(log.reg$VComparativootrasescuelas)
# Crear variables dummy

#str(log.reg)
summary(log.reg)


# Revision de los datos
glimpse(log.reg)

# Descripcion datos categoricos
log.reg %>%
keep(is.factor) %>%
summary()

# Descripcion datos numéricos
log.reg %>%
keep(is.numeric) %>%
summary()


#Visualización de los datos y su ditribución
histogram(x=log.reg$MotivoBaja, main ="Histograma dela distribución de desertores", xlab="Desertor(1) Migración otro campus(0)", ylab = "Frecuencia")


histogram(x=log.reg$AniosUniversidad, main ="Histograma de Años en la Universidad", xlab="# Años", ylab = "Frecuencia")
histogram(x=log.reg$cve_escuela, main ="Distribución de desertores por escuela", xlab="# Desertores", ylab = "Frecuencia")
histogram(x=log.reg$no_semestres, main ="Histograma de los números de semestres", xlab="#Semestres", ylab = "Frecuencia")

#Evaluar proporción de missing values
log.reg %>%
select(cve_escuelaingreso) %>%
table(exclude=NULL) %>%
prop.table()


set.seed(88)
split = sample.split(log.reg$MotivoBaja,SplitRatio=.75)
log.train= subset(log.reg,split==T)
log.test= subset(log.reg,split==F)

#Distribución de clases entre los set de datos
select(log.reg,MotivoBaja)

round(prop.table(table(select(log.reg, MotivoBaja), exclude = NULL)), 4) * 100
round(prop.table(table(select(log.train, MotivoBaja), exclude = NULL)), 4) * 100
round(prop.table(table(select(log.test, MotivoBaja), exclude = NULL)), 4) * 100
```

```
##************************MODELOS CON REGRESION LOGISTICA
library(MASS)

# Ajuste de un modelo logístico.
modelo_logistico <- glm(MotivoBaja ~ admision, data = log.train, family = "binomial")

# Representación gráfica del modelo.
ggplot(data = log.train, aes(x = admision, y = MotivoBaja)) +
  geom_point(aes(color = as.factor(MotivoBaja)), shape = 1) +
  stat_function(fun = function(x){predict(modelo_logistico,
                                  newdata = data.frame(MotivoBaja = x),
                                  type = "response")}) +
  theme_bw() +
  labs(title = "Regresión logística",
       y = "Probabilidad resultado examen admision") +
  theme(legend.position = "none")


# Con geom_smooth se puede obtener el gráfico directamente.
ggplot(data = log.reg, aes(x =admision, y = MotivoBaja)) +
  geom_point(aes(color = as.factor(MotivoBaja)), shape = 1) +
  geom_smooth(method = "glm",
              method.args = list(family = "binomial"),
              color = "gray20",
              se = FALSE) +
  theme_bw() +
  theme(legend.position = "none")


#Al ingresar a la Universidad
#CONTEXTO PERSONAL
modelo.reg.log.d0 = glm(MotivoBaja ~ edad + genero + VExperienciacInstitucion + NSE + ciudad +
cve_prospecto , data = log.train, family =  binomial) %>%
  stepAIC(race = 1, direction = "backward")
summary(modelo.reg.log.d0)

#Grafico edad y Deserción
ggplot(data = log.train, aes(x = MotivoBaja, y = edad, colour = MotivoBaja)) +
  geom_boxplot() +
  theme_bw() +
  theme(legend.position = "bottom") +
  labs(title = "Relación entre la edad y la deserción",
       y = "Edad") +
  theme(legend.position = "none")


#Grafico NSE y deserción
ggplot(data = log.train, aes(x = MotivoBaja, y = NSE, color = MotivoBaja)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.1) +
  theme_bw() +
  theme(legend.position = "bottom") +
  labs(title = "Distribución de desertores y su Nivel Socieconómico",
       y = "NSE"  ) +
  theme(legend.position = "none")
```

```
#Grafico Experiencia Institucional  y deserción
ggplot(data = log.train, aes(x = MotivoBaja, y = VExperienciacInstitucion, color = MotivoBaja)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.1) +
  theme_bw() +
  theme(legend.position = "bottom") +
  labs(title = "Desertores y confianza en la Institución",
       y = "Experiencia previa"  ) +
  theme(legend.position = "none")



#coeficiente el cambio en el log-odds de la respuesta como resultado de un cambio unitario en la variable predictora,
exp(coef(modelo.reg.log.d0)["edad"])
exp(coef(modelo.reg.log.d0)["NSE"])
exp(coef(modelo.reg.log.d0)["VExperienciacInstitucion"])



#CONTEXTO ACADÉMICO

modelo.reg.log.d1 = glm(MotivoBaja ~ cve_escuela + cve_programa + admision + verbal + matematico + redaccion + promedio_ingreso
                        +  tipo_escuela + VExcelenciaacademica + Vpromedioacademico + VDeportista, data = log.train, family =  binomial) %>%
  stepAIC(race = 1, direction = "backward")
summary(modelo.reg.log.d1)

#Grafico cve_escuela con todo el set de datos
ggplot(data = log.reg, aes(x = MotivoBaja, y = cve_escuela, color = MotivoBaja)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.1) +
  theme_bw() +
  theme(legend.position = "bottom") +
  labs(title = "Distribución de desertores y la Escuela que pertenecía su programa académico",
       y = "Escuela: Ingeniería (1)/ Administración(2)"  ) +
  theme(legend.position = "none")

#Grafico cve_ programa con todo el set de datos
ggplot(data = log.reg, aes(x = MotivoBaja, y = cve_programa, color = MotivoBaja)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.1) +
  theme_bw() +
  theme(legend.position = "bottom") +
  labs(title = "Distribución de desertores y su programa académico",
       y = "Escuela: Ingeniería (1 al 8)/ Administración(9 al 12)"  ) +
  theme(legend.position = "none")

#Grafico de dispersión con todo el set de datos
plot(x =log.reg$MotivoBaja, y = log.reg$cve_programa, col= log.reg$MotivoBaja)

#Promedio de ingreso
ggplot(data = log.train, aes(x = MotivoBaja, y = promedio_ingreso, color = MotivoBaja)) +
  geom_boxplot(outlier.shape = NA) +
```

```
  geom_jitter(width = 0.1) +
  theme_bw() +
  theme(legend.position = "null")  +
  theme_bw() +
  labs(title = "Relación entre el promedio de ingreso y la deserción",
      y = "promedio ingreso") +
  theme(legend.position = "none")

ggplot(data = log.train, aes(x = MotivoBaja, y = admision, color = MotivoBaja)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.1) +
  theme_bw() +
  theme(legend.position = "null") +
  labs(title = "Relación entre el desertor y su examen de admision",
      y = "Excelencia académica previa") +
  theme(legend.position = "none")

ggplot(data = log.train, aes(x = MotivoBaja, y = matematico, color = MotivoBaja)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.1) +
  theme_bw() +
  theme(legend.position = "bottom") +
  labs(title = "Relación entre el area matemática del examen de admisión y  la deserción",
      y = "Escuela pública (1)/ privada(2)") +
  theme(legend.position = "none")

ggplot(data = log.train, aes(x = MotivoBaja, y = VDeportista, color = MotivoBaja)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.1) +
  theme_bw() +
  theme(legend.position = "null") +
  labs(title = "Desertores con historial deportivo",
      y = "Historial deportivo") +
  theme(legend.position = "none")


#coeficiente el cambio en el log-odds de la respuesta como resultado de un cambio unitario en la variable predictora,
exp(coef(modelo.reg.log.d1)["promedio_ingreso"])
exp(coef(modelo.reg.log.d1)["tipo_escuela"])
exp(coef(modelo.reg.log.d1)["VExcelenciaacademica"])

# Institucional
modelo.reg.log.d2 = glm(MotivoBaja ~  becaingreso  + pafeni + pro_ing + ping + piafi +  piaf + egre + cedu + onal + talento + dg +
                        depo + conc +  rect + labo + hieg + sgv , data = log.train, family =  binomial) %>%
  stepAIC(race = 1, direction = "backward")
summary(modelo.reg.log.d2)

ggplot(data = log.train, aes(x = MotivoBaja, y = piafi, color = MotivoBaja)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.1) +
  theme_bw() +
  theme(legend.position = "null")

ggplot(data = log.train, aes(x = MotivoBaja, y = egre, color = MotivoBaja)) +
```

```
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.1) +
  theme_bw() +
  theme(legend.position = "null")

ggplot(data = log.train, aes(x = MotivoBaja, y = becaingreso, color = MotivoBaja)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.1) +
  theme_bw() +
  theme(legend.position = "null")

ggplot(data = log.train, aes(x = MotivoBaja, y = depo, color = MotivoBaja)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.1) +
  theme_bw() +
  theme(legend.position = "null")

ggplot(data = log.train, aes(x = MotivoBaja, y = piaf, color = MotivoBaja)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.1) +
  theme_bw() +
  theme(legend.position = "null")

exp(coef(modelo.reg.log.d2)["piafi"])
exp(coef(modelo.reg.log.d2)["piaf"])
exp(coef(modelo.reg.log.d2)["egre"])

#Durante su estancia en la Universidad
#Datos academicos
modelo.reg.log.d3 = glm(MotivoBaja ~ Promediofinal +  TotalFaltas + PromedioPeriodoSeleccionado +
Materiascursadas +
                        TotalReprobadas + Reprobadasarea1 + Reprobadasarea2 + Reprobadasarea3 +
TotalAprobadas +
                        Extraordinarios + ultimosemestre + avance, data = log.train, family =  binomial)  %>%
  stepAIC(race = 1, direction = "backward")
summary(modelo.reg.log.d3)

ggplot(data = log.train, aes(x = MotivoBaja, y = PromedioPeriodoSeleccionado, color = MotivoBaja)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.1) +
  theme_bw() +
  theme(legend.position = "null") +
  labs(title = "Relación entre el desertor y su promedio en el período de baja",
       y = "Promedio del período de baja") +
  theme(legend.position = "none")

ggplot(data = log.train, aes(x = MotivoBaja, y = avance, color = MotivoBaja)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.1) +
  theme_bw() +
  theme(legend.position = "null") +
  labs(title = "Relación entre el desertor y su % de avance",
       y = "% de Avance") +
  theme(legend.position = "none")
```

```
      ggplot(data = log.train, aes(x = MotivoBaja, y = Reprobadasarea1, color = MotivoBaja)) +
        geom_boxplot(outlier.shape = NA) +
        geom_jitter(width = 0.1) +
        theme_bw() +
        theme(legend.position = "null") +
        labs(title = "Relación entre el desertor y las materias reprobadas del área de formación general",
             y = "Reprobabdas eje formación general") +
        theme(legend.position = "none")

      exp(coef(modelo.reg.log.d3)["PromedioPeriodoSeleccionado"])
      exp(coef(modelo.reg.log.d3)["Reprobadasarea1"])

      #Institucional
      #Areas de servicio
      modelo.reg.log.d4 = glm(MotivoBaja ~ Becapromedio + VPrestamoEquipo + VDepartamentoescolar +
VPagoencaja + VCreditoeducativo +
                        VBiblioteca + VCafeteria + VPapeleria + VEnfermeria + VServiciosocial +
VProgramasinternacionales +
                        VTrmitesdeprcticasprofesionales + VConsultacalificacioneseninternet + VCampaadereinas,
data = log.train, family =  binomial) %>%
        stepAIC(race = 1, direction = "backward")
      summary(modelo.reg.log.d4)

      ggplot(data = log.train, aes(x = MotivoBaja, y = Becapromedio, color = MotivoBaja)) +
        geom_boxplot(outlier.shape = NA) +
        geom_jitter(width = 0.1) +
        theme_bw() +
        theme(legend.position = "null")

      ggplot(data = log.train, aes(x = MotivoBaja, y = VCreditoeducativo, color = MotivoBaja)) +
        geom_boxplot(outlier.shape = NA) +
        geom_jitter(width = 0.1) +
        theme_bw() +
        theme(legend.position = "null")

      exp(coef(modelo.reg.log.d4)["Becapromedio"])
      exp(coef(modelo.reg.log.d4)["VCreditoeducativo"])

      # equípamiento e infraestructura
      modelo.reg.log.d5 = glm(MotivoBaja ~ VLaboratoriodecmputo + VMaestrosenasesorasacadmicas +
VMaestrosfueradeclases + VEnseanzaaprendizajes +
                  VLaboratoriosdecmputo + VLaboratoriosdeingeniera + VSalasaudiovisuales + VSalndeclases +
VCoordinadordecarrera, data = log.train, family =  binomial)  %>%
        stepAIC(race = 1, direction = "backward")
      summary(modelo.reg.log.d5)

      ggplot(data = log.train, aes(x = MotivoBaja, y = VLaboratoriosdeingeniera , color = MotivoBaja)) +
        geom_boxplot(outlier.shape = NA) +
        geom_jitter(width = 0.1) +
        theme_bw() +
        theme(legend.position = "null")

      # Evaluación de las actividades cocurriculaes y extracurriculares
      modelo.reg.log.d6 = glm(MotivoBaja ~ VEquiposrepresentativos + VEventosacadmicos + VExposiciones +
VSemanacultural + VSociedaddealumnos +
```

```
                    VTalleresculturales + VTorneosinternos + VViajesdeestudio + VConferencias +
VAmbienteestudiantil + VDeportes +
                    VDifusincultural + VAsuntosestudiantiles, data = log.train, family =  binomial)  %>%
        stepAIC(race = 1, direction = "backward")
      summary(modelo.reg.log.d6)

      #Evaluación de la Institución, satisfacción y orgullo de pertenencia
      modelo.reg.log.d7 = glm(MotivoBaja ~ VOrgullo + Calificacion + VSatisfaccion +
VComparativootrasescuelas, data = log.train, family =  binomial)  %>%
        stepAIC(race = 1, direction = "backward")
      summary(modelo.reg.log.d7)
      ggplot(data = log.train, aes(x = MotivoBaja, y = VComparativootrasescuelas, color = MotivoBaja)) +
        geom_boxplot(outlier.shape = NA) +
        geom_jitter(width = 0.1) +
        theme_bw() +
        theme(legend.position = "null")
      ggplot(data = log.train, aes(x = MotivoBaja, y = Calificacion, color = MotivoBaja)) +
        geom_boxplot(outlier.shape = NA) +
        geom_jitter(width = 0.1) +
        theme_bw() +
        theme(legend.position = "null")
      ggplot(data = log.train, aes(x = MotivoBaja, y = VOrgullo, color = MotivoBaja)) +
        geom_boxplot(outlier.shape = NA) +
        geom_jitter(width = 0.1) +
        theme_bw() +
        theme(legend.position = "null")



      #Evaluación de la infraestructura de la institución
      modelo.reg.log.d8 = glm(MotivoBaja ~VBlackboard + VRedinalmbrica + VConsultaelectronicaenbiblioteca
+ VAreasdeportivas + VAreasverdesyplazas +
                    VAuladevideoconferencias + VCafeDVolada + VEstacionamiento, data = log.train, family =
binomial) %>%
        stepAIC(race = 1, direction = "backward")
      summary(modelo.reg.log.d8)
      ggplot(data = log.train, aes(x = MotivoBaja, y = VAreasverdesyplazas, color = MotivoBaja)) +
        geom_boxplot(outlier.shape = NA) +
        geom_jitter(width = 0.1) +
        theme_bw() +
        theme(legend.position = "null")


      # Comparación de los modelos
      screenreg(list(modelo.reg.log.d0, modelo.reg.log.d1, modelo.reg.log.d2, modelo.reg.log.d3,
modelo.reg.log.d4, modelo.reg.log.d5,modelo.reg.log.d6,modelo.reg.log.d7,modelo.reg.log.d8),
caption="Comparación de modelos logit")
      # EVALUACION DE LOS MODELOS FINALES CON LAS VARIABLES MAS SIGNIFICATIVAS DE
LOS DIFERENTES MOMENTOS DEL DESERTOR

      #matriz de correlacion
      log.reg %>%
select(edad,verbal,matematico,promedio_ingreso,becaingreso,pafeni,pro_ing,piafi,Reprobadasarea1, piaf,
egre,talento,labo,
          PromedioPeriodoSeleccionado,Becapromedio) %>%
        cor() %>%
```

```
corrplot()


#Modelo final 1 personal y academico
modelo.reg.log.df1 = glm(MotivoBaja ~   VExperienciacInstitucion +  edad +  NSE + promedio_ingreso +
tipo_escuela +
                         VExcelenciaacademica +  matematico +VDeportista, data = log.train, family = binomial)
%>%
  stepAIC(race = 1, direction = "backward")
summary(modelo.reg.log.df1)

modelo.reg.log.df1 = glm(MotivoBaja ~   cve_programa + admision + promedio_ingreso + avance +
becaingreso + piafi, data = log.train, family = binomial) %>%
  stepAIC(race = 1, direction = "backward")
summary(modelo.reg.log.df1)


#Modelo final 2
modelo.reg.log.df2 = glm(MotivoBaja ~   piafi + egre + becaingreso + piaf + pafeni + pro_ing + matematico
+
                         promedio_ingreso ,data = log.train, family = binomial) %>%
  stepAIC(race = 1, direction = "backward")
summary(modelo.reg.log.df2)


#Modelo final 3
modelo.reg.log.df3 = glm(MotivoBaja ~     piafi +  egre + Becapromedio+
                           PromedioPeriodoSeleccionado , data = log.train, family = binomial) %>%
  stepAIC(race = 1, direction = "backward")
summary(modelo.reg.log.df3)

#Modelo final 4
modelo.reg.log.df4 = glm(MotivoBaja ~     piafi + egre + NSE + piaf +
                         Becapromedio + PromedioPeriodoSeleccionado,data = log.train, family = binomial) %>%
  stepAIC(race = 1, direction = "backward")
summary(modelo.reg.log.df4)


#Modelo final 5
modelo.reg.log.df5 = glm(MotivoBaja ~    matematico + NSE + piafi + egre + Becapromedio +
                         PromedioPeriodoSeleccionado, data = log.train, family = binomial) %>%
  stepAIC(race = 1, direction = "backward")
summary(modelo.reg.log.df5)

# Comparación de los modelos
  screenreg(list(modelo.reg.log.df1, modelo.reg.log.df2, modelo.reg.log.df3, modelo.reg.log.df4,
modelo.reg.log.df5), caption="Comparación de modelos logit")

#Comparacion del AIC de los modelos
models <- list(modelo.reg.log.df1, modelo.reg.log.df2, modelo.reg.log.df3, modelo.reg.log.df4,
modelo.reg.log.df5)
model.names <- c('Modelo 1', 'Modelo 2', 'Modelo 3','Modelo 4' ,'Modelo 5')

# k numero de parametros, AICc score del modelo, Delta_AiCc diferencia entre el AIC y el AIC mejor
modelo,
aictab(cand.set = models, modnames = model.names)
```

```
#Evaluando la multicolinearidad
library(stats)
library(corrplot)
log.reg %>%
  keep(is.numeric) %>%
  cor() %>%
  corrplot()
vif(modelo.reg.log.df4)

######################PROBANDO CON REGRESION LOGISTICA LOS PREDICTORES QUE ME ARROJO EL FOREST RANDOM##############

modelo.reg.log.df1 = glm(MotivoBaja ~   admision +  matematico +  verbal + promedio_ingreso + Becapromedio +
                         becaingreso, data = log.train, family = binomial) %>%
  stepAIC(race = 1, direction = "backward")
summary(modelo.reg.log.df1)

## agregando las becas del modelo 4
modelo.reg.log.df1 = glm(MotivoBaja ~   matematico +  verbal + piaf+ egre + promedio_ingreso + Becapromedio +
                         becaingreso, data = log.train, family = binomial) %>%
  stepAIC(race = 1, direction = "backward")
summary(modelo.reg.log.df1)
#############################################################################################################################


#*******************************EVALUACION DE DESEMPEÑO DEL MODELO FINAL
modelo.reg.log.df = modelo.reg.log.df4
summary(modelo.reg.log.df)

#bondad de ajuste evaluada con la deviance
dev <- modelo.reg.log.df$deviance
nullDev <- modelo.reg.log.df$null.deviance
modelChi <- nullDev - dev
modelChi
pv<-1-(modelChi)
pv
#la probabilidad asociada al estadístico chi-cuadrado
chigl <- modelo.reg.log.df$df.null - modelo.reg.log.df$df.residual
chisq.prob <- 1 - pchisq(modelChi, chigl)
chisq.prob


summary(modelo.reg.log.df1)$coefficients


#El primer paso es quitar el logaritmo y convertir al coeficiente en una razón de probabilidad:
#exponenciar el logaritmo. En R lo hacemos con la función exp(), que nos regresa el exponencial de un logaritmo
exp(coef(modelo.reg.log.df)["piafi"])
exp(coef(modelo.reg.log.df)["piaf"])
exp(coef(modelo.reg.log.df)["egre"])
```

```
exp(coef(modelo.reg.log.df)["PromedioPeriodoSeleccionado"])
exp(coef(modelo.reg.log.df)["Becapromedio"])
exp(coef(modelo.reg.log.df)["NSE2"])
exp(coef(modelo.reg.log.df)["NSE3"])

exp(3.49)
0.03918834/(10.03918834)



coefplot(modelo.reg.log.df)  +
  theme_minimal() +
  labs(title="Estimación de coeficientes con error estandar",
       x="Estimación",
       y="Variable",
       caption="Elaboración propia")


#Intervalos de confianza s inferiores y superiores de nuestro intervalo de confianza estén
#por encima de 1 nos da la confianza de que la dirección de la relación que hemos observado es cierta en la población.
confint(modelo.reg.log.df)
coefplot(modelo.reg.log.df)  +
  theme_minimal() +
  labs(title="Intervalos de confianza",
       x="Estimación",
       y="Variable", guide = "none",
       caption="Deserción")

exp(confint(modelo.reg.log.df, level=0.99))


#OBTENCION DE RESIDUOS, VALORES PREDICHOS Y ESTADISTICOS
tabla<-(4:9)
tabla$probabilidades.predichas <- fitted(modelo.reg.log.df)
tabla$studentized.residuals <- rstudent(modelo.reg.log.df)
tabla$dfbeta <- dfbeta(modelo.reg.log.df)
tabla$dffit <- dffits(modelo.reg.log.df)
tabla$leverage <- hatvalues(modelo.reg.log.df)
head(tabla[c("Desertor","piafi egre  NSE  piaf Becapromedio
PromedioPeriodoSeleccionado","probabilidades predichas")])


#curve(predict(modelo.reg.log.df, data.frame((piafi + egre + NSE + piaf +
 #                                  Becapromedio + PromedioPeriodoSeleccionado = x), type = "response"),
  #    col = "firebrick", lwd = 2.5, add = TRUE)



#Tabla y gráfico de efectos totales promedio
allEffects(modelo.reg.log.df)
plot(allEffects(modelo.reg.log.df))
labs(title="Variables dependientes del modelo",
       caption="Modelo predictivo de deserción")
```

```
#Estimacion de la prueba Wald() nos dice para cada predictora si es significativamente diferente de 0
z=b/SEb
z <- summary(modelo.reg.log.df)$coefficients[,1]/summary(modelo.reg.log.df)$coefficients[,2]
z
p <- (1 - pnorm(abs(z), 0, 1)) * 2  #Usamos pnorm() para estimar la probabilidad, dado que es una función de
prob acumulada usamos 1-pnorm()
p

#chi cuadrada
#p-value = 1 - pchisq(deviance, degrees of freedom)


#Predicciones del valor esperado en el set de entrenamiento
predict.train.log = predict(modelo.reg.log.df,newdata = log.train, type = "response")
tapply(predict.train.log, log.train$MotivoBaja,mean)

#Predicciones del valor esperado en el set de prueba
predict.log = predict(modelo.reg.log.df, newdata = log.test, type = "response")
tapply(predict.log, log.test$MotivoBaja,mean)

predict.data = predict

#Tabla de confusion de las predicciones
confusion<-table(log.test$MotivoBaja,predict.log > .5)
confusion



#Accuracy
pred<-ifelse(modelo.reg.log.df$fitted.values<0.5,0,1)
pred
Accuracy(pred,  log.train$MotivoBaja, dig = 8)


fourfoldplot(confusion, color = c("black", "yellow"),
          conf.level = 0, margin = 1, main = "Matriz de confusión")

#confusionMatrix((predict.log),(log.test))


n<-sum(confusion) # numero de instancias
nc<-sum(nrow(confusion)) # numero de clases
diag <- sum(diag(confusion)) # numero de instancias clasificadas correctamente
rowsums <- (apply(confusion, 1, sum)) #numero de instancias por clase
colsums <- (apply(confusion, 2, sum)) # numero de predicciones por clase
p <- rowsums / n # distribucion de instancias entre las clases
q <- colsums / n # distribucion de instancias entre las clases predichas
p
q

oneVsAll = lapply(1 : nc,
              function(i){
                v = c(confusion[i,i],
                     rowsums[i] - confusion[i,i],
                     colsums[i] - confusion[i,i],
                     n-rowsums[i] - colsums[i] + confusion[i,i]);
```

```
                return(matrix(v, nrow = 2, byrow = T))})
oneVsAll
s = matrix(0, nrow = 2, ncol = 2)
for(i in 1 : nc){s = s + oneVsAll[[i]]}
s
avgAccuracy = sum(diag(s)) / sum(s)
avgAccuracy

micro_prf = (diag(s) / apply(s,1, sum))[1];
micro_prf

#ROC
tot<-colSums(confusion)# Number salidas w/ each test result
truepos<-unname(rev(cumsum(rev(confusion[2,2])))) # Number of true positives
falsepos<-unname(rev(cumsum(rev(confusion[2,1])))) # Number of false positives
falseneg<- unname(rev(cumsum(rev(confusion[1,2])))) # Number of false negativos
totneg<-sum(confusion[1,1]+confusion[1,2])  # The total number of negatives (one number)
trueneg<- unname(rev(cumsum(rev(confusion[1,1])))) # Number of true negativos
totpos<-sum(confusion[2,1]+confusion[2,2])  # The total number of positives (one number)
sens=truepos/(truepos+falseneg) # Sensitivity (fraction true positives)
sens
omspec=trueneg/(trueneg+falsepos) # specificity (false positives)
sens=c(sens,0); omspec=c(omspec,0) # Numbers when we classify all as normal
sens
plot(omspec, sens, type="b", xlim=c(0,1), ylim=c(0,1), lwd=2,
    xlab="Especificidad", ylab="Sensibilidad") # perhaps with xaxs="i"
grid()
abline(0,1, col="red", lty=2)

height = (sens[-1]+sens[-length(sens)])/2
width = -diff(omspec) # = diff(rev(omspec))
sum(height*width)


ROC(log.test$MotivoBaja,predict.log > .5,plot = "ROC")

#Accuracy
#Exactitud (acurracy) se refiere a cuán cerca del valor real se encuentra el valor medido.

# (VP+VN)/(VP+FP+FN+VN)
exactitud <- sum(diag)/ n
exactitud
(truepos+trueneg)/n

#Precision
#precisión (precision) se refiere a la dispersión del conjunto de valores obtenidos
#de mediciones repetidas de una magnitud.  Cuanto menor es la dispersión mayor la precisión
#(VP)/(VP+FP)
precision<- diag/colsums
precision
precision<-truepos/(truepos+falsepos)
precision

#Sensibilidad
#Sensibilidad (recall) se refiere a la respuesta que el instrumento de medición
#tenga para medir una variable y que tan rápida sea este para estabilizar su medida.
```

```
#VP/(VP+FN)
sensibilidad <- diag/ rowsums
sensibilidad
sensibilidad<-truepos/(truepos+falseneg)
sensibilidad

#Especificidad
#También conocida como la Tasa de Verdaderos Negativos, ("true negative rate") o TN. Se
#trata de los casos negativos que el algoritmo ha clasificado correctamente.  Expresa cuan bien puede el
modelo detectar esa clase.
#Se calcula:  VN/(VN+FP)
especificidad<-trueneg/(trueneg+falsepos)
especificidad



#F-valor
#F-Valor (F-measure) medida de precisión que tiene una prueba.
#F1 =2*(precision*recall)/(precision+recall)=tp/(tp+1/2(fp+fn) ),  siendo un modelo perfecto cuando F1=1.
#2*(precision*sensibilidad)/(precision+sensibilidad)
f1 <- (2*(precision*sensibilidad) / (precision+sensibilidad))
f1
f1<- truepos/(truepos+(1/(2*(falsepos+falseneg))))
f1
f1<- 2*(sensibilidad*precision)/(sensibilidad+precision)
f1


#Coeficiente phi asociación entre dos variables binarias, va de -1 a 1.


#Tabla de evaluacion del modelo
data.frame(precision,sensibilidad,f1)

#Area bajo la curva
height = (sens[-1]+sens[-length(sens)])/2
width = -diff(omspec) # = diff(rev(omspec))
sum(height*width)


# PROMEDIO
macroPrecision <- mean(precision)
macroExactitud <- mean(exactitud)
macroF1 <- mean(f1)
data.frame(macroPrecision, macroExactitud, macroF1)


pred <- ifelse(modelo.reg.log.df$fitted.values < 0.5, 0, 1)
Accuracy(log.train,pred,2)



rocr.pred = prediction(predict.log, log.test$MotivoBaja)
rocr.perf = performance(rocr.pred, "tpr", "fpr")
```

```
height = (sensibilidad[-1]+sensibilidad[-length(sensibilidad)])/2
width = -diff(omspec) # = diff(rev(omspec))
height
width
auc<-sum(height*width)
auc
#auc<-AUC(predict.log,log.test)
plot(rocr.perf, print.cutoffs.at=seq(0,1,0.1), text.adj = c(-0.2,1.7),colorize=TRUE,main = "ROC AUC  Modelo Regresión Logística")


plot(rocr.perf, avg = "threshold", colorize=TRUE, lwd= 3,main = "ROC AUC  Modelo Regresión Logística")
plot(rocr.perf, lty=3, col="grey78", add=TRUE)


rocr.perf <- performance(pred, "prec", "rec")
plot(rocr.perf, avg= "threshold", colorize=TRUE, lwd= 3,main= "... Precision/Recall graphs ...")
plot(rocr.perf, lty=3, col="grey78", add=TRUE)

rocr.perf <- performance(pred, "sens", "spec")
plot(rocr.perf, avg= "threshold", colorize=TRUE, lwd= 3, main="... Sensitivity/Specificity plots ...")
plot(rocr.perf, lty=3, col="grey78", add=TRUE)


rocr.perf <- performance(pred, "lift", "rpp")
plot(rocr.perf, avg= "threshold", colorize=TRUE, lwd= 3,main= "... and Lift charts.")
plot(perf, lty=3, col="grey78", add=TRUE)


plot(rocr.perf, colorize=TRUE, lwd=2,main='ROC curves from 10-fold cross-validation')

plot(rocr.perf, avg='vertical', spread.estimate='stderror',lwd=3,main='Vertical averaging + 1 standard error',col='blue')

plot(rocr.perf, avg='horizontal', spread.estimate='boxplot',lwd=3,main='Horizontal averaging + boxplots',col='blue')

plot(rocr.perf, avg='threshold', spread.estimate='stddev',lwd=2, main='Threshold averaging + 1 standard deviation',colorize=TRUE)
```

```
# ************CONSTRUCCIÓN DEL MODELO DE CON RANDOM FOREST

#Directorio de trabajo
setwd("~/LUCIA.BELTRAN.PERSONAL/CDC/DOCTORADO/UPAEP/ACADEMICO/Tesis/Tesis IV")
library(ranger)
library(tibble)
library(dplyr)
library(MASS)
library(dplyr)
library(tidyr)
library(ggplot2)
library(ggpubr)
library(tidymodels)
```

```
library(ranger)
library(doParallel)
library(randomForest)
library(ROCR)
library(forcats)
library(cowplot)
library(corrplot)
library(stats)
library(caTools)
library(ggplot2)
library(Epi)
library(rpart, lib.loc = "C:/Program Files/R/R-4.1.0/library")
library(factoextra)
library(rpart.plot)
library(caret)
library(cowplot)
library(car)
library(ROCR)
library("rfVarImpOOB")
library(texreg)
library(coefplot)
library(effects)
library(AICcmodavg)
library(cvAUC)
library(randomForest)
library(MASS)
library(tidyverse)
library(stats)
library(corrplot)
library(devtools)
library(ggraph)
library(igraph)


# Bosques aleatorios
arbol.D <- read.csv ('Analisis final desertores MB.csv')
str(arbol.D)

#Convertir la variable de salida en un factor
arbol.D$MotivoBaja<-as.character(arbol.D$MotivoBaja)
arbol.D$MotivoBaja<-as.factor(arbol.D$MotivoBaja)

#Convertir la variable numéricas a categóricas
arbol.D$VExperienciacInstitucion<-as.factor(arbol.D$VExperienciacInstitucion)
arbol.D$Vpromedioacademico<-as.factor(arbol.D$Vpromedioacademico)
arbol.D$VPrestamoEquipo<-as.factor(arbol.D$VPrestamoEquipo)
arbol.D$VDeportes<-as.factor(arbol.D$VDeportes)
arbol.D$VDeportista<-as.factor(arbol.D$VDeportista)
arbol.D$VComparativootrasescuelas<-as.factor(arbol.D$VComparativootrasescuelas)
arbol.D$VCreditoeducativo<-as.factor(arbol.D$VCreditoeducativo)
arbol.D$cve_escuela<-as.character(arbol.D$cve_escuela)
arbol.D$genero<-as.factor(arbol.D$genero)
arbol.D$cve_programa<-as.factor(arbol.D$cve_programa)
arbol.D$cve_escuelaingreso<-as.character(arbol.D$cve_escuelaingreso)
arbol.D$tipo_escuela<-as.factor(arbol.D$tipo_escuela)
arbol.D$cve_prospecto<-as.factor(arbol.D$cve_prospecto)
```

```
arbol.D$cve_genero<-as.factor(arbol.D$genero)
arbol.D$estatus_admision<-as.factor(arbol.D$estatus_admision)
arbol.D$NSE<-as.factor(arbol.D$NSE)
arbol.D$MotivoBaja<-as.factor(arbol.D$MotivoBaja)
arbol.D$Vpromedioacademico<-as.factor(arbol.D$Vpromedioacademico)
arbol.D$VDeportista<-as.factor(arbol.D$VDeportista)
arbol.D$VExcelenciaacademica<-as.factor(arbol.D$VExcelenciaacademica)
arbol.D$VPrestamoEquipo<-as.factor(arbol.D$VPrestamoEquipo)
arbol.D$VDepartamentoescolar<-as.factor(arbol.D$VDepartamentoescolar)
arbol.D$VPagoencaja<-as.factor(arbol.D$VPagoencaja)
arbol.D$VCreditoeducativo<-as.factor(arbol.D$VCreditoeducativo)
arbol.D$VBiblioteca<-as.factor(arbol.D$VBiblioteca)
arbol.D$VCafeteria<-as.factor(arbol.D$VCafeteria)
arbol.D$Vcafeteria<-as.factor(arbol.D$Vcafeteria)
arbol.D$VBlackboard<-as.factor(arbol.D$VBlackboard)
arbol.D$VRedinalmbrica<-as.factor(arbol.D$VRedinalmbrica)
arbol.D$VConsultaelectronicaenbiblioteca<-as.factor(arbol.D$VConsultaelectronicaenbiblioteca)
arbol.D$VLaboratoriodecmputo<-as.factor(arbol.D$VLaboratoriodecmputo)
arbol.D$VPapeleria<-as.factor(arbol.D$VPapeleria)
arbol.D$VEnfermeria<-as.factor(arbol.D$VEnfermeria)
arbol.D$VLaboratoriosdecmputo<-as.factor(arbol.D$VLaboratoriosdecmputo)
arbol.D$VDeportes<-as.factor(arbol.D$VDeportes)
arbol.D$VDifusincultural<-as.factor(arbol.D$VDifusincultural)
arbol.D$VMaestrosenasesorasacadmicas<-as.factor(arbol.D$VMaestrosenasesorasacadmicas)
arbol.D$VMaestrosfueradeclases<-as.factor(arbol.D$VMaestrosfueradeclases)
arbol.D$VAsuntosestudiantiles<-as.factor(arbol.D$VAsuntosestudiantiles)
arbol.D$VServiciosocial<-as.factor(arbol.D$VServiciosocial)
arbol.D$VProgramasinternacionales<-as.factor(arbol.D$VProgramasinternacionales)
arbol.D$VTrmitesdeprcticasprofesionales<-as.factor(arbol.D$VTrmitesdeprcticasprofesionales)
arbol.D$VEnseanzaaprendizajes<-as.factor(arbol.D$VEnseanzaaprendizajes)
arbol.D$VConsultacalificacioneseninternet<-as.factor(arbol.D$VConsultacalificacioneseninternet)
arbol.D$VAreasdeportivas<-as.factor(arbol.D$VAreasdeportivas)
arbol.D$VAreasverdesyplazas<-as.factor(arbol.D$VAreasverdesyplazas)
arbol.D$VAuladevideoconferencias<-as.factor(arbol.D$VAuladevideoconferencias)
arbol.D$VBaos<-as.factor(arbol.D$VBaos)
arbol.D$VCafeDVolada<-as.factor(arbol.D$VCafeDVolada)
arbol.D$VEstacionamiento<-as.factor(arbol.D$VEstacionamiento)
arbol.D$VLaboratoriosdeingeniera<-as.factor(arbol.D$VLaboratoriosdeingeniera)
arbol.D$VSalasaudiovisuales<-as.factor(arbol.D$VSalasaudiovisuales)
arbol.D$VSalndeclases<-as.factor(arbol.D$VSalndeclases)
arbol.D$VCampaadereinas<-as.factor(arbol.D$VCampaadereinas)
arbol.D$VConferencias<-as.factor(arbol.D$VConferencias)
arbol.D$VEquiposrepresentativos<-as.factor(arbol.D$VEquiposrepresentativos)
arbol.D$VEventosacadmicos<-as.factor(arbol.D$VEventosacadmicos)
arbol.D$VExposiciones<-as.factor(arbol.D$VExposiciones)
arbol.D$VSemanacultural<-as.factor(arbol.D$VSemanacultural)
arbol.D$VSociedaddealumnos<-as.factor(arbol.D$VSociedaddealumnos)
arbol.D$VTalleresculturales<-as.factor(arbol.D$VTalleresculturales)
arbol.D$VTorneosinternos<-as.factor(arbol.D$VTorneosinternos)
arbol.D$VViajesdeestudio<-as.factor(arbol.D$VViajesdeestudio)
arbol.D$VCoordinadordecarrera<-as.factor(arbol.D$VCoordinadordecarrera)
arbol.D$VOrgullo<-as.factor(arbol.D$VOrgullo)
arbol.D$VAmbienteestudiantil<-as.factor(arbol.D$VAmbienteestudiantil)
arbol.D$VOrgullo<-as.factor(arbol.D$VOrgullo)
arbol.D$Calificacion<-as.factor(arbol.D$Calificacion)
arbol.D$VSatisfaccion<-as.factor(arbol.D$VSatisfaccion)
```

```
arbol.D$VComparativootrasescuelas<-as.factor(arbol.D$VComparativootrasescuelas)

set.seed(123)
str(arbol.D)

# Matriz de correlacion

arbol.D%>%
  dplyr::select(admision,matematico, verbal,
redaccion,promedio_ingreso,avance,becaingreso,piafi,edad,onal,pro_ing,
        pafeni,depo,egre,ping)%>%
  cor() %>%
  corrplot()

arbol.D %>%
  keep(is.numeric) %>%
  cor() %>%
  corrplot()

#Se eliminan los campos de redaccion, verbal y matematico
arbol.D<-arbol.D[,-(14:16)]
str(arbol.D)

ggplot(data = arbol.D, aes(x = MotivoBaja, y = avance, color = MotivoBaja)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.1) +
  theme_bw() +
  theme(legend.position = "null") +
  labs(title = "Relación entre el desertor y su % de avance",
       y = "% de Avance") +
  theme(legend.position = "none")


# Datos de entrenamiento y de prueba
total <- nrow(arbol.D)
entreno <- round(total *.5)
tamano<-sample(1:total, size = entreno)
train.arbol <- arbol.D[tamano,]
test.arbol = arbol.D[-tamano,]
train.arbol$MotivoBaja <- as.character(train.arbol$MotivoBaja)
train.arbol$MotivoBaja <- as.factor(train.arbol$MotivoBaja)


#Corro el randomForest para todos los datos
rf <-randomForest(MotivoBaja~.,data=train.arbol, mtry=7, ntree=100)
rf <-randomForest(MotivoBaja~.,data=arbol.D, ntree=500)
print(rf)
plot(rf)
oob.error.data <- data.frame(
  Trees=rep(1:nrow(rf$err.rate), times=3),
  Type=rep(c("OOB", "0", "1"), each=nrow(rf$err.rate)),
  Error=c(rf$err.rate[,"OOB"],
        rf$err.rate[,"0"],
        rf$err.rate[,"1"]))

ggplot(
```

```
  data = oob.error.data,
  aes(x = Trees,
      y = Error,
      fill = Error)
)+
  labs(x = "arboles", title = "# arboles/ OOBs") +
  geom_col() +
  coord_flip() +
  theme_bw() +
  theme(legend.position = "none")

#Evaluate variable importance
importancia_pred <- rf$importance %>%
  enframe(name = "predictor", value = "importancia")


#Mostrar las 25 variables mas importantes
importancia_pred %>%
  dplyr::arrange(desc(predictor)) %>%
  dplyr::top_n(25) %>%
  ggplot(aes(reorder(predictor, importancia), importancia)) +
  geom_col() +
  coord_flip() +
  ggtitle("Top 25 important variables")




#mtry es el numero de variables que considera en cada paso del total
#Se prueba con la raiz cuadrada del total de variables del set de datos
inicio<-floor(sqrt(ncol(arbol.D) - 1))

#Selecciona el valor de mtry con el minimo error of bag(OOB) error.

mtry <- tuneRF(arbol.D[-1],arbol.D$MotivoBaja,mtryStart = inicio, ntreeTry=450,
               stepFactor=1,improve=0.005, trace=FALSE, plot=TRUE, doBest =FALSE )
best.m <- mtry[mtry[, 2] == min(mtry[, 2]), 1]
print(mtry)
print(best.m)



##############################################################################AQUIIIIIIIIIIIIIIIIIII

#Encontrando los mejores parametros con la libreria ranger
#Produce una salida ordenada (variando la primera columna como la más lenta, en lugar de la más rápida).
param_grid = expand_grid(
  'num_trees' = c(50, 100, 300, 500, 1000),
  'mtry'      = c(3, 5, 7, ncol(train.arbol)-1),
  'max_depth' = c(1, 3, 10, 20))

#Loop para ajustar un modelo con cada combinación de hiperparámetros
```

```
      #
================================================================================
      oob_error<-matrix(nrow=nrow(param_grid), ncol=1)
      menorError<-9999
      for(i in 1:nrow(param_grid)){

        modelo <- ranger(
          formula   = MotivoBaja ~ .,
          data      = train.arbol,
          num.trees = param_grid$num_trees[i],
          mtry      = param_grid$mtry[i],
          max.depth = param_grid$max_depth[i],
          importance = "impurity",
          seed      = 123
        )
        modelo

        if (menorError>modelo$prediction.error )
        {
          menorError<- modelo$prediction.error
          mejormodelo<-modelo
          posicion<-i
          arboles<-modelo$num.trees
          maxvariables<-modelo$mtry
          nodos<-modelo$num.independent.variables
          importantes<-modelo$variable.importance
        }
        oob_error[i] <- modelo$prediction.error

      }
      hist(oob_error, breaks = 20)
      modelo
      mejormodelo
      oob_error
      menorError
      posicion
      arboles
      maxvariables
      nodos
      ########data = importantes + reorder(importantes[1,],importantes[,2])

      resultados <- param_grid
      resultados$oob_error <- oob_error
      resultados <- resultados %>% arrange(oob_error)
      head(resultados,1)

      maxvariables=as.integer(resultados[1,2])
      arboles<-as.integer(resultados[1,1])
      depth<-as.integer(resultados[1,3])

      #Se construye el modelo con los mejores parametros
      model <-randomForest(MotivoBaja~.,data=train.arbol, mtry = maxvariables,importance=TRUE,
                           ntree=arboles, max_nodes =depth )

      model
      importance(model)
```

```
varImpPlot(model, sort = TRUE, n.var = 5, main = "Importancia de los predictores")

### TRATANDO DE VISUALIZAR LOS
MODELOS#############################################################################

### FUNCION PARA VISUALIZAR ARBOLES

#Thomas Lin Pedersen's
library(dplyr)
library(ggraph)
library(igraph)

options(ggrepel.max.overlaps = Inf)

tree_func <- function(final_model, tree_num) {

#get tree by index
tree<-getTree(final_model,tree_num,labelVar = TRUE) %>%

  tibble::rownames_to_column() %>%
# make leaf split points to NA, so the 0s won't get plotted
  mutate(`split point` = ifelse(is.na(prediction), `split point`, NA))


##https://shiring.github.io/machine_learning/2017/03/16/rf_plot_ggraph

# prepare data frame for graph
graph_frame <- data.frame(from = rep(tree$rowname, 2),
                          to = c(tree$`left daughter`, tree$`right daughter`))

# convert to graph and delete the last node that we don't want to plot
graph <- graph_from_data_frame(graph_frame) %>% delete_vertices("0")

# set node labels
V(graph)$node_label <- gsub("_", " ", as.character(tree$`split var`))
V(graph)$leaf_label <- as.character(tree$prediction)
V(graph)$split <- as.character(round(tree$`split point`, digits = 2))

# plot
plot <- ggraph(graph, 'dendrogram') +
  theme_bw() +
  geom_edge_link() +
  geom_node_point() +
  geom_node_text(aes(label = node_label), na.rm = TRUE, repel = TRUE) +
  geom_node_label(aes(label = split), vjust = 2.5, na.rm = TRUE, fill = "white") +
  geom_node_label(aes(label = leaf_label, fill = leaf_label), na.rm = TRUE,
                  repel = TRUE, colour = "white", fontface = "bold", show.legend = FALSE) +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major = element_blank(),
        panel.background = element_blank(),
        plot.background = element_rect(fill = "white"),
        panel.border = element_blank(),
        axis.line = element_blank(),
        axis.text.x = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks = element_blank(),
```

```
      axis.title.x = element_blank(),
      axis.title.y = element_blank(),
      plot.title = element_text(size = 18))


print(plot)

}

  # llamando a la función para imprimir el arbol con menos nodos
  tree_num <- which(model$forest$ndbigtree == min(model$forest$ndbigtree))
  tree_func(model, tree_num)

#Or we can plot the tree with the biggest number of nodes:
tree_num <- which(model$forest$ndbigtree == max(model$forest$ndbigtree))
tree_func(model, tree_num)


tree_func(model, 150)
model$votes
model$y

#Evaluate variable importance
importancia_pred <- model$importance %>%
  enframe(name = "predictor", value = "importancia")
reorder(importancia_pred$predictor,importancia_pred$importancia[,3])
predictoras <-importancia_pred$predictor

importanciamodelo<-matrix(nrow = 25, ncol = 4)
importanciamodelo<-importancia_pred$importancia
predictores<-matrix(1:25, nrow = 25, ncol =5)
for (i in 1:25) {

  predictores[i,1]<-predictoras[i]
  predictores[i,2]<-importanciamodelo[i,1]
  predictores[i,3]<-importanciamodelo[i,2]
  predictores[i,4]<-importanciamodelo[i,3]
  predictores[i,5]<-importanciamodelo[i,4]

}

predictorestotales<-as.data.frame(predictores)
predictorestotales$V2<-as.double(predictorestotales$V2)
predictorestotales$V3<-as.double(predictorestotales$V3)
predictorestotales$V4<-as.double(predictorestotales$V4)
predictorestotales$V5<-as.double(predictorestotales$V5)

predictorestotales %>%
  dplyr::arrange(desc(V1)) %>%
  dplyr::top_n(25) %>%
  ggplot(aes(reorder(V1, V5), V5)) +
  geom_col() +
  coord_flip() +
  ggtitle("Top 25 important variables")
```

```
# MeanDecreaseGinin
ggplot(
  data = predictorestotales,
  aes(x = reorder(V1,V5),
      y = V5,
      fill = V5)
) +
  labs(x = "predictor", title = "Importancia predictores (pureza de nodos)") +
  geom_col() +
  coord_flip() +
  theme_bw() +
  theme(legend.position = "none")

#Analisis de correlación
arbol.D %>%tidyselect:::select(admision, avance, redaccion, matematico, verbal, becaingreso, edad, piafi,
promedio_ingreso,
                    onal, pafeni, piaf, egre, ping) %>%
    cor() %>%
    corrplot()

#MeanDecreaseAccuracy
ggplot(
  data = predictorestotales,
  aes(x = reorder(V1,V4),
      y = V4,
      fill = V4)
) +
  labs(x = "predictor", title = "Importancia predictores (Mean Decrease Accuracy)") +
  geom_col() +
  scale_fill_viridis_c() +
  coord_flip() +
  theme_bw() +
  theme(legend.position = "none")

#Predictora Deserción
ggplot(
  data = predictorestotales,
  aes(x = reorder(V1,V3),
      y = V3,
      fill = V3)
) +
  labs(x = "predictor", title = "Importancia predictores (Clase:Desertores)") +
  geom_col() +
  coord_flip() +
  theme_bw() +
  theme(legend.position = "none")


#Predictora No desercion
ggplot(
  data = predictorestotales,
  aes(x = reorder(V1,V2),
      y = V2,
      fill = V2)
```

```
  ) +
    labs(x = "predictor", title = "Importancia predictores (Clase:No Desertores)") +
    geom_col() +
    coord_flip() +
    theme_bw() +
    theme(legend.position = "none")

  #performance del modelo en el set de entrenamiento
  pred1=predict(model,type = "prob")

  perf = prediction(pred1[,2], train.arbol$MotivoBaja)
  # 1. Area under curve
  auc = performance(perf, "auc")
  auc
  # 2. True Positive and Negative Rate
  pred3 = performance(perf, "tpr","fpr")
  # 3. Plot the ROC curve
  plot(pred3,main="Curva ROC para el modelo Random Forest",col=2,lwd=2, legend= "En terminos de
probabilidad")
  abline(a=0,b=1,lwd=2,lty=2,col="gray")

  #predicciones en el set de pruebas
  predicciones.0 = predict(model,test.arbol)
  model$err.rate
  mc.0<-with(test.arbol, table(predicciones.0,MotivoBaja))
  mc.0
  100 * sum(diag(mc.0)) / sum(mc.0)

  #LOS ERRORES Y LAS CLASES
  oob.error.data <- data.frame(
    Trees=rep(1:nrow(model$err.rate), times=3),
    Type=rep(c("OOB", "0", "1"), each=nrow(model$err.rate)),
    Error=c(model$err.rate[,"OOB"],
          model$err.rate[,"0"],
          model$err.rate[,"1"]))

  #GRAFICA DE LOS ERRORES Y LAS CLASES CON # ARBOLES
  ggplot(
    data = oob.error.data,
    aes(x = Trees,
        y = Error,
        fill = Error)
  )+
    labs(x = "arboles", title = "# arboles/ OOBS") +
    geom_col() +
    coord_flip() +
    theme_bw() +
    theme(legend.position = "none")


  ggplot(data = oob.error.data, aes(x=Trees, y=Error)) +
    geom_line(aes(color=Type))
  # EVALUACIÓN DEL MODELO RANDOM FOREST Y MEDICIÓN DEL DESEMEPEÑO

  library(rattle)
  library(ROCR)
```

```
library(pROC)
library("party")
library(Epi)

#EVALUACION DEL MODELO FINAL
modelo.arbol.final = model
print(modelo.arbol.final)


predicciones.f = predict(modelo.arbol.final,test.arbol, type = "class")
modelo.arbol.final$err.rate
mc<-with(test.arbol, table(predicciones.f,MotivoBaja))
mc
100 * sum(diag(mc)) / sum(mc)


#Predicciones en terminos de clasificacion TPR FPR
pred1=predict(modelo.arbol.final,type = "class")

perf = prediction(pred1[,2], test.arbol$MotivoBaja)
# 1. Area under curve
auc = performance(perf, "auc")
auc
# 2. True Positive and Negative Rate
pred3 = performance(perf, "tpr","fpr")
# 3. Plot the ROC curve
plot(pred3,main="Curva ROC en términos de clasificación",col=2,lwd=2)
abline(a=0,b=1,lwd=2,lty=2,col="gray")
#Grafica
plot(pred3, print.cutoffs.at=seq(0,1,0.1), text.adj = c(-0.2,1.7))


# prediccion del modelo en terminos de probabilidad
pred1<-modelo.arbol.final%>%
  predict(new_data=test.arbol,type="prob")
# El valor de cada columna se corresponde con la probabilidad de que la observación pertenezca a la clase

head(pred1)
ggplot(data=oob.error.data, aes(x=Trees, y=Error)) +
  geom_line(aes(color=Type))

pred1=predict(modelo.arbol.final,type = "prob")
perf = prediction(pred1[,2], arbol.D$MotivoBaja)
# 1. Area under curve
auc = performance(perf, "auc")
auc
# 2. True Positive and Negative Rate
pred3 = performance(perf, "tpr","fpr")
# 3. Plot the ROC curve
plot(pred3,main="Curva ROC en términos de probabilidad ",col=2,lwd=2)
abline(a=0,b=1,lwd=2,lty=2,col="gray")
#Grafica
plot(pred3, print.cutoffs.at=seq(0,1,0.1), text.adj = c(-0.2,1.7), main="Area bajo la curva en términos de
probabilidad")
```

```
#Tabla de confusion de las predicciones
confusion<-mc
confusion

fourfoldplot(mc, color = c("red", "green"),
          conf.level = 0, margin = 1, main = "Matriz de confusión")
oob.error.data <- data.frame(
  Trees=rep(1:nrow(modelo.arbol.final$err.rate), times=3),
  Type=rep(c("OOB", "0", "1"), each=nrow(modelo.arbol.final$err.rate)),
  Error=c(modelo.arbol.final$err.rate[,"OOB"],
        modelo.arbol.final$err.rate[,"0"],
        modelo.arbol.final$err.rate[,"1"]))

n<-sum(confusion) # numero de instancias
nc<-sum(nrow(confusion)) # numero de clases
diag <- sum(diag(confusion)) # numero de instancias clasificadas correctamente
rowsums <- (apply(confusion, 1, sum)) #numero de instancias por clase
colsums <- (apply(confusion, 2, sum)) # numero de predicciones por clase
p <- rowsums / n # distribucion de instancias entre las clases
q <- colsums / n # distribucion de instancias entre las clases predichas
p
q

oneVsAll = lapply(1 : nc,
                function(i){
                  v = c(confusion[i,i],
                       rowsums[i] - confusion[i,i],
                       colsums[i] - confusion[i,i],
                       n-rowsums[i] - colsums[i] + confusion[i,i]);
                  return(matrix(v, nrow = 2, byrow = T))})
oneVsAll
s = matrix(0, nrow = 2, ncol = 2)
for(i in 1 : nc){s = s + oneVsAll[[i]]}
s
avgAccuracy = sum(diag(s)) / sum(s)
avgAccuracy

micro_prf = (diag(s) / apply(s,1, sum))[1];
micro_prf

#ROC
tot<-colSums(confusion)# Number salidas w/ each test result
truepos<-unname(rev(cumsum(rev(confusion[2,2])))) # Number of true positives
falsepos<-unname(rev(cumsum(rev(confusion[2,1])))) # Number of false positives
falseneg<- unname(rev(cumsum(rev(confusion[1,2])))) # Number of false negativos
totneg<-sum(confusion[1,1]+confusion[1,2])  # The total number of negatives (one number)
trueneg<- unname(rev(cumsum(rev(confusion[1,1])))) # Number of true negativos
totpos<-sum(confusion[2,1]+confusion[2,2])  # The total number of positives (one number)
sens=truepos/totpos  # Sensitivity (fraction true positives)
sens
omspec=trueneg/totneg # specificity (false positives)
sens=c(sens,0); omspec=c(omspec,0) # Numbers when we classify all as normal
```

```
sens
plot(omspec, sens, type="b", xlim=c(0,1), ylim=c(0,1), lwd=2,
    xlab="Especificidad", ylab="Sensibilidad") # perhaps with xaxs="i"
grid()
abline(0,1, col="red", lty=2)

height = (sens[-1]+sens[-length(sens)])/2
width = -diff(omspec) # = diff(rev(omspec))
sum(height*width)

auc<-sum(height*width)
auc
ROC(test.arbol$MotivoBaja,predicciones.f ,plot = "ROC")



#Accuracy
#Exactitud (acurracy) se refiere a cuán cerca del valor real se encuentra el valor medido.
# (VP+VN)/(VP+FP+FN+VN)
exactitud <- sum(diag)/ n
exactitud
(truepos+trueneg)/n

#Precision
#precisión (precision) se refiere a la dispersión del conjunto de valores obtenidos
#de mediciones repetidas de una magnitud.  Cuanto menor es la dispersión mayor la precisión
#(VP)/(VP+FP)
precision<- diag/colsums
precision
precision<-truepos/(truepos+falsepos)
precision

#Sensibilidad
#Sensibilidad (recall) se refiere a la respuesta que el instrumento de medición
#tenga para medir una variable y que tan rápida sea este para estabilizar su medida.
#VP/(VP+FN)
sensibilidad <- diag/ rowsums
sensibilidad
sensibilidad<-truepos/(truepos+falseneg)
sensibilidad

#Especificidad
#También conocida como la Tasa de Verdaderos Negativos, ("true negative rate") o TN. Se
#trata de los casos negativos que el algoritmo ha clasificado correctamente.  Expresa cuan bien puede el
modelo detectar esa clase.
#Se calcula:  VN/(VN+FP)
especificidad<-trueneg/(trueneg+falsepos)
especificidad



#F-valor
#F-Valor (F-measure) medida de precisión que tiene una prueba.
#F1 =2*(precision*recall)/(precision+recall)=tp/(tp+1/2(fp+fn) ),  siendo un modelo perfecto cuando F1=1.
#2*(precision*sensibilidad)/(precision+sensibilidad)
```

```
f1 <- (2*(precision*sensibilidad) / (precision+sensibilidad))
f1
f1<- truepos/(truepos+(1/(2*(falsepos+falseneg))))
f1
f1<- 2*(sensibilidad*precision)/(sensibilidad+precision)
f1
```