FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA PROFESIONAL DE ARQUITECTURA

PREDICCIÓN DEL RENDIMIENTO ACADÉMICO MEDIANTE MINERÍA DE DATOS EN ESTUDIANTES DEL PRIMER CICLO DE LA ESCUELA PROFESIONAL DE INGENIERÍA DE COMPUTACIÓN Y SISTEMAS, UNIVERSIDAD DE SAN MARTÍN DE PORRES, LIMA-PERÚ

PRESENTADA POR

EIRIKU YAMAO

ASESOR

AUGUSTO ERNESTO BERNUY ALVA

TESIS

PARA OPTAR EL GRADO DE MAESTRO EN INGENIERÍA DE COMPUTACIÓN Y SISTEMAS CON MENCIÓN EN GESTIÓN DE TECNOLOGÍAS DE INFORMACIÓN

LIMA – PERÚ

2018

SECCIÓN DE POSGRADO

PREDICCIÓN DEL RENDIMIENTO ACADÉMICO MEDIANTE MINERÍA DE DATOS EN ESTUDIANTES DEL PRIMER CICLO DE LA ESCUELA PROFESIONAL DE INGENIERÍA DE COMPUTACIÓN Y SISTEMAS, UNIVERSIDAD DE SAN MARTÍN DE PORRES, LIMA-PERÚ

TESIS

PARA OPTAR EL GRADO DE MAESTRO EN INGENIERÍA DE COMPUTACIÓN Y SISTEMAS CON MENCIÓN EN GESTIÓN DE TECNOLOGÍAS DE INFORMACIÓN

PRESENTADA POR

YAMAO, EIRIKU

LIMA – PERÚ

2018

DEDICATORIA

A mi esposa Valery y a mis hijos, Eiji, Imari y Akari

AGRADECIMIENTO

A la Facultad de Ingeniería y Arquitectura de la Universidad de San Martín de Porres, en especial a su decano, Ing. Luis Celi Saavedra por su constante apoyo para la realización del presente trabajo de investigación.

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABLAS

RESUMEN

El rendimiento académico es un tema estudiado desde hace mucho tiempo. Los alumnos ingresantes de las universidades son los más vulnerables a enfrentar problemas de rendimiento, resultando en posible deserción. La minería de datos en educación aplica técnicas de minería de datos en la información generada en el sector educación. El presente trabajo consiste en realizar la predicción del rendimiento académico de los alumnos que ingresaron a la Escuela Profesional de Ingeniería de Computación y Sistemas de la Universidad de San Martin de Porres en el primer ciclo utilizando minería de datos. Se extrajeron datos de 1304 ingresantes que fueron clasificados en tres factores: sociales, económicos y académicos. Se realizaron predicciones a través de tres técnicas: regresión lineal, árbol de decisiones y support vector machines, y el mejor resultado de 82.87% se obtuvo utilizando árbol de decisiones. De los diferentes factores, los que más influyeron en el rendimiento académico fueron los siguientes: nota de examen de admisión, género, edad, modalidad de ingreso y distancia desde su casa hasta el centro de estudios. Utilizando minería de datos fue posible realizar predicciones del rendimiento académico de los ingresantes. Esto permitió la detección de ingresantes que podrían enfrentarse a problemas en sus estudios.

Palabras Clave: *rendimiento académico, predicción, educational data mining, educación superior.*

ABSTRACT

Academic performance has been studied for a long time. First year student in higher education are the most likely to face issues concerning their studies possibly resulting in desertion. Educational Data Mining applies data mining methods to data generated in an educational environment. The present document consists in predicting the academic performance of first year students from the School of Computer and Systems Engineering of Universidad de San Martin de Porres – Peru, in the first semester using data mining. Data of 1304 first year students was extracted and classified as social, economic and academic factors. The prediction performed using three data mining techniques: linear regression, decision trees and Support Vector Machines, achieving an 82.87% of precision using C5.0 decision tree algorithm. From the different factors used in this study grade obtained during admission exam, gender, age, admission type, distance and secondary school demonstrated to have an influence in academic performance. The data mining approach was viable to predict the academic performance of the first year students in study. Results obtained in this study will allow the university the early detection of those who might encounter issues during their first semester.

Keywords: *Academic Performance, Prediction, Educational Data Mining, EDM, Higher Education.*

INTRODUCCIÓN

La digitalización de los procesos académicos, en las universidades, permite generar cuantiosos datos electrónicos con relación con los estudiantes, puesto que proporciona información que ayudaron a docentes y gestores. Además, contribuye con el avance sobre la calidad de los procesos educativos proporcionando testimonio oportuno a las diferentes partes interesadas. El propósito de los métodos de minería de datos es extraer conocimientos significativos de los propios datos. La aplicación de los métodos de minería de datos a los datos generados en el sector educación se conoce como Minería de Datos Educativos: EDM (Han, Kamber y Pei, 2012), (Romero y Ventura, 2013).

La minería de datos educativos o Educational Data Mining (EDM) es una disciplina emergente que busca desarrollar métodos para explorar data generada en el sector educación, a fin de lograr una mejor comprensión de las características de los estudiantes y la forma cómo aprenden. Se han propuesto cinco categorías o enfoques principales en la EDM: 1ro) predicción, 2do) agrupación, 3ro) minería de relaciones, 4to) descubrimiento dentro de modelos; y 5to) destilación de datos para el juicio humano. La primera categoría es primordial para predecir los resultados o granularidad de los estudiantes en cualquier escenario de su manifestación. La EDM predice las calificaciones de los estudiantes mediante la integración del tiempo con la

cantidad de asistencia que un determinado estudiante requiere para resolver problemas. (Romero *et al*, 2013), (Chalaris, Gritzalis, Maragoudakis, Sgouropoulou, y Tsolakidis, 2014), (Romero, Ventura, Pechenizkiy, y Baker, 2011).

Strecht, Cruz, Soares, Merdes-Moreria y Abren (2015) refieren que se puede predecir el éxito o fracaso de los estudiantes usando variables sociales como edad, sexo, estado civil, nacionalidad, desplazados (si el estudiante vivió fuera del distrito), becas, necesidades especiales, tipo de admisión, tipo de estudiante (ordinario, empleado, deportista, *etc*.), años de matrícula, cursos diferidos, tipo de dedicación (tiempo completo, tiempo parcial) como posible situación de deuda (ElGamal, 2013).

Un indicador medido en la educación superior está relacionado con el rendimiento académico de los estudiantes, donde el mismo, es el resultado de la interacción sistemática entre las evaluaciones, tareas asignadas y cumplimiento de actividades extra-docentes integrales, entre otros. A partir de la técnica de regresión pueden medirse las calificaciones de desempeño en estudiantes desde su primer año de ingreso hasta la culminación. El desempeño del estudiante, sin duda alguna, puede predecirse aplicando la EDM y ello posibilita evaluar el binomio estudiante-universidad. (Shahiri y Husain, 2015), (Calisir, Basak, y Comertoglu, 2016), (Han *et al*, 2012).

El objeto de investigación son los estudiantes universitarios que ingresan, pues se ha visto que son los más vulnerables por ello, renunciar a los estudios universitarios. La muestra poblacional estuvo construida por los estudiantes ingresantes de la Escuela Profesional de Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres, Lima-Perú.

La estructura metodológica por capítulo de esta investigación es la siguiente: En el capítulo I, describe el problema científico, objetivos y justificación de la investigación. En el segundo, se presenta el marco teórico, formulación de hipótesis, variables de selección e indicadores operacionales. En el tercero, se aborda el marco metodológico donde se describe el método

de investigación, diseño de investigación, población, muestra, técnicas e instrumentos de investigación para la recolección de los datos. En el cuarto, se considera el desarrollo de la metodología referida al proceso estándar de cruzamiento industrial para la minería de datos (por sus siglas en inglés: CRISP-DM); y en el quinto capítulo, se exponen los resultados y su discusión. Finalmente se presentan las conclusiones, recomendaciones, fuentes de información y anexos.

CAPÍTULO I
PLANTEAMIENTO DEL PROBLEMA

1.1 Determinación del problema

Romero *et al* (2013), mencionan que la EDM se encarga de desarrollar métodos para analizar datos sobre un sistema educativo con la finalidad de poder descubrir patrones en grandes conjuntos de datos que, de otro modo, resultarían muy difíciles o imposibles de analizar, ya que existe un gran volumen de datos. Como resultado, la minería de datos se utiliza para decidir sobre cómo mejorar el proceso de enseñanza y aprendizaje, además, de diseñar o rediseñar un entorno de aprendizaje.

El rendimiento de los estudiantes constituye una estructura esencial dentro de la educación superior, pues, el logro del historial académico, es uno de los criterios básicos que toman en cuenta las universidades de alta calidad (Ministery of Education Malaysia: MEM, 2015). Bin Mat, Buniyamin, Arsad, y Kassim (2013) señalaron que el rendimiento de los estudiantes se puede obtener mediante la medición en la evaluación del aprendizaje y co-currículo, pero podría estar dirigida hacia aspectos específicos del periodo de estudio durante el tiempo de permanencia en la institución universitaria (Mohamed,

Husaina & Abdul, 2015). Asimismo, es importante para la institución educativa que el nivel de éxito alcanzado por los estudiantes es un reflejo de la calidad de enseñanza brindada. (York, Gibson y Rankin, 2015).

Existen múltiples estudios que, a través de la aplicación de técnicas de minería de datos, buscan realizar predicciones del rendimiento académico de los estudiantes. Es una tarea difícil de lograr ya que estudios realizados indican que el rendimiento académico depende de múltiples factores tanto de tipo personal como socio-económico, psicológico y otras variables más. Una correcta predicción permitió detectar, de antemano, a aquellos alumnos que tendrían dificultades en superar los cursos, lo que ayudó a tomar acertadas decisiones como brindar apoyo adicional en forma de tutorías o asesorías, cambios o ajustes por parte de los docentes, entre otros. (Ramesh, Parkavi y Ramar, 2013) (Mishra, Kumar y Gupta, 2014).

En el caso de los estudiantes que ingresaron a la universidad, un porcentaje determinado de ellos no continúan sus estudios y es muy difícil, en muchos casos, hacer predicciones sobre el rendimiento académico, pues este tangible, por lo general, trata de pronosticarse mediante indicadores de análisis de progresiones individuales al final de una etapa de estudio, de manera que debería considerarse, de forma anticipada, con datos suficientes, la probabilidad de permanencia universitaria, a través de un soporte cognoscitivo de medición integrador. (Yadav y Pal, 2012) (Sepehrian, 2012) (Cheewaprakobkit, 2013).

1.2 Formulación del problema

Problema general:

¿Cómo puede predecirse el rendimiento académico en estudiantes de primer ciclo mediante una técnica analítica estadística y computacional?

Problemas específicos:

a) ¿Cuáles indicadores de impacto pueden medir el rendimiento académico en estudiantes universitarios de primer ciclo?

b) ¿Qué estadígrafo puede aplicarse para medir el rendimiento académico en estudiantes universitarios de primer ciclo?
c) ¿Cuál técnica estadística y computacional puede medir el rendimiento académico en estudiantes universitarios de primer ciclo?

1.3 Objetivos

Objetivo general:

Predecir el rendimiento académico mediante minería de datos en estudiantes del primer ciclo de la Escuela Profesional de Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres.

Objetivos específicos:

a) Estimar indicadores sociales, económicos y académicos para el rendimiento académico en estudiantes universitarios de primer ciclo de la Escuela Profesional de Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres.
b) Determinar mediante análisis de componentes principales la significación de los indicadores sociales, económicos y académicos en estudiantes universitarios de primer ciclo de la Escuela Profesional de Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres.
c) Aplicar la minería de datos educacional para indicadores sociales, económicos y académicos en estudiantes universitarios de primer ciclo de la Escuela Profesional de Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres.

1.4 Justificación y alcances de la investigación

La aplicación de EDM, en las universidades, permitió crear una cultura de toma de decisiones, basadas en los datos generados en las mismas, involucrando a los departamentos de tecnología de información a crear nuevas políticas en la recolección y uso de datos. Además posibilitó a los

principales stakeholders (estudiantes, padres de familia, docentes, personal administrativo y autoridades) convertirse en consumidores inteligentes de la nueva información que se generan y realizar preguntas críticas sobre las ofertas y demandas que se ofrece en la universidad. (Bienkowski, Feng, y Means, 2012) (Lang, Siemens, Wise y Gasevic, 2017).

Li, Rusk & Song (2013), señalaron que, aproximadamente el 30% de los estudiantes que estudian carreras de Ingeniería, abandonan los estudios durante el primer año de enseñanza universitaria. En múltiples casos, el rendimiento académico es la causa principal, aunque el mismo está influenciado por diferentes motivos.

Ramesh *et al* (2013), por su parte, indican que las limitaciones futuras de ofertas laborales relacionadas con el perfil de estudio, conjuntamente con las acciones viables de adaptación al contexto universitario, influyen tempranamente en el rendimiento académico de manera que, la necesidad de encontrar herramientas predictivas para entender y mantener el nivel de estudiantes universitarios, resulta una necesidad social.

En el Perú, los estudios o investigaciones relacionadas con el análisis del rendimiento académico en las universidades son limitados y más aún, cuando se utiliza la Minería de Datos Educativos. En esta investigación, sobre los estudiantes del primer ciclo académico de la Escuela Profesional de Ingeniería de Computación y Sistemas se evaluó el rendimiento académico aplicando la EDM. Se seleccionaron diversos indicadores, y los que fueron analizados utilizando estadígrafos paramétricos.

1.5 Limitaciones

Algunos indicadores como, por ejemplo, motivación de la carrera universitaria, satisfacción universitaria y expectativas de integración de la carrera universitaria, entre otros, no se incluyeron en la EDM, ya que, en la base de datos de la universidad, no fueron considerados.

CAPÍTULO II
MARCO TEÓRICO

2.1 Antecedentes del problema

La minería de datos es el proceso de descubrir patrones interesantes de gran cantidad de datos. Es considerado como el resultado natural de la evolución de las tecnologías de información y de la industria de base de datos que lograron desarrollar funcionalidades criticas como recolección y creación de datos, administración de datos y técnicas avanzadas de análisis. La gran abundancia de datos generados mediante la evolución de las tecnologías de internet creó la necesidad de contar con bases de datos heterogéneos e interconectados a escala mundial junto con herramientas poderosas para el análisis de datos. Dada la gran velocidad en la evolución de la generación de data, se crea una situación que se considera como rico en datos, pero pobres en información. Se busca acercar esta brecha mediante el desarrollo sistemático de herramientas de minería de datos para poder convertir estos datos en fuentes de información confiables. (Han *et al*, 2012).

Estos cambios también tienen un fuerte impacto en el sector educación, en donde los recursos de e-learning, software de educación instrumental, el uso de Internet y el establecimiento de repositorios de datos

han incrementado considerablemente los repositorios de datos. Toda esta información es muy importante y es necesario explorarla y explotarla para entender mejor el proceso de aprendizaje de los estudiantes. Hoy, uno de los principales retos al que se enfrenta este sector es analizar este crecimiento exponencial de data educacional y usar esta data para mejorar la toma de decisiones. (Romero *et al*, 2013).

La mejora en la toma de decisiones permitirió a las instituciones educativas a incrementar el nivel de calidad que se brinda a través de nuevos conocimientos que pueden ser descubiertos en los datos que se encuentran en las bases de datos y en los resultados de actividades académicas como evaluaciones que podrán ser extraídas utilizando técnicas de minería de datos. (Chalaris *et al*, 2014).

La Minería de Datos para Educación es reconocida, en la actualidad, como una disciplina emergente que se enfoca en el desarrollo dimensional para explorar los datos que se generan en el contexto educativo. Estos datos provienen de diferentes fuentes como ambiente de clases tradicional, software educativo, cursos en línea y evaluaciones. Estas fuentes proveen constantemente de más datos que pueden ser analizados para responder preguntas que no eran posibles anteriormente como diferencias entre la población estudiantil y cambios en el comportamiento de los estudiantes. (Romero et al, 2011).

EDM es un campo interdisciplinario que incluye recopilación de la información, sistema de recomendaciones, análisis de datos, análisis de redes sociales, psicopedagogía, psicología del aprendizaje, entre otros. Es la combinación de tres campos principales: ciencias de la computación, educación y estadística (Figura 1).

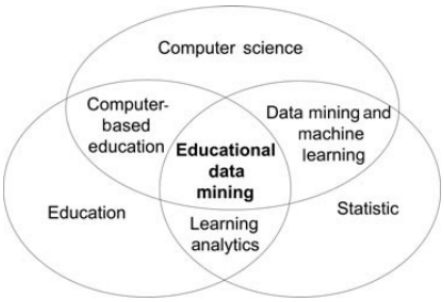


Figura 1. Principales áreas relacionadas con minería de datos para educación

Fuente: Romero et al. (2013)

Una de las principales diferencias que podemos encontrar en comparación con otras áreas en donde se aplicó la minería de datos es en negocios, bioinformática, genética, medicina, etc. es que todos ellos tienen objetivos distintos. Por ejemplo, en el caso de negocio electrónico, el objetivo principal es incrementar las ganancias que se pueden medir en dinero y que ofrecen medidas secundarias muy claras como datos relacionados con clientes como cantidad o fidelización del cliente. En el caso de educación, el objetivo principal es mejorar el proceso de aprendizaje, lo cual es difícil de medir y debe ser estimado en valores como mejoras en rendimiento académico. En general, la EDM permitió mejorar la toma de decisiones basadas en el análisis de data educativo que facilite las prácticas utilizadas en educación y los materiales de aprendizaje. (Romero *et al*, 2013).

Según Liñán y Pérez (2015), la popularidad por parte de los investigadores en los últimos años a la aplicación de minería de datos, en el sector educación, se debe a los siguientes factores:

- Existe un interés en mejorar el proceso de toma de decisiones utilizando un enfoque basado en datos generados, como ocurre en inteligencia de negocios y analítica de datos.
- En la actualidad, se conocen técnicas y métodos basados en estadística, sistemas expertos y minería de datos para buscar patrones en los datos y construir modelos de predicción o reglas que

pueden ser adaptados fácilmente a los datos generados en este sector.

- La generación de los datos por utilizar es relativamente fácil o ya existen, y su almacenamiento y procesamiento es posible con la tecnología de computación actual.
- Dada la fuerte competencia y la crisis financiera, las universidades están bajo presión de reducir costos e incrementar las ganancias mediante la explotación de la creciente demanda en educación de calidad, reducir el abandono académico y mejorar la oferta educativa.

Según Romero et al. (2011), en los últimos años, la EDM se ha dedicado a resolver diferentes tipos de preguntas, entre los que destacan:

- Comunicación con principales interesados. El objetivo es ayudar al personal administrativo y docente a realizar un análisis de las actividades y el uso de información en clase de los estudiantes.
- Mantenimiento y mejora de cursos. El propósito es determinar cómo mejorar los cursos (contenido, actividades, relaciones, etc.), utilizando información sobre los estudiantes y el aprendizaje.
- Generar recomendaciones. La meta es realizar recomendaciones sobre qué temas (o tareas o cursos) son los más adecuados en este momento.
- Predicción de notas y resultados de aprendizaje. El objetivo es predecir la nota final u otros resultados del proceso de aprendizaje como retención o capacidad de aprendizaje a futuro, basado en data de los cursos.
- Creación de modelos. El modelado de usuarios, en el ámbito educativo, tiene ciertas aplicaciones como la detección (normalmente en tiempo real) del estado y característica de los estudiantes como satisfacción, motivación, avances en el aprendizaje o problemas que

pueden impactar negativamente en el proceso de aprendizaje (cometer muchos errores, no utilizar muy seguido las opciones de ayuda, búsquedas ineficientes, etc.), estilos de aprendizaje, preferencias, entre otros.

- Análisis de estructura. El fin es determinar la estructura del dominio de aprendizaje utilizando la habilidad para predecir el rendimiento de los estudiantes y medir la calidad del dominio.

Papamitsiou y Economides (2014), luego de una revisión de la literatura en EDM, identifican las fortalezas, oportunidades, debilidades y amenazas (FODA) de este campo (tabla 1), resaltando que en la actualidad existe una intersección entre los campos de educación, psicología, pedagogía y computación en donde cada *click* en un entorno de aprendizaje electrónico genera información que puede ser analizada, cada acción realizada dentro de un ambiente de aprendizaje puede ser aislado, identificado y clasificado para crear patrones. Cada interacción puede ser clasificada en diferentes tipos de comportamiento y decodificada para mejorar la toma de decisiones.

De las diferentes aplicaciones o tareas que existen en el ámbito de educación que se pueden resolver mediante minería de datos (Tabla 2), uno de los más antiguos y el más popular es la predicción de rendimiento académico de los estudiantes (Romero et al, 2013).

El rendimiento académico es considerado una parte esencial en instituciones de educación universitaria y es utilizado como uno de los criterios para medir el nivel de las universidades. Según algunas investigaciones, el rendimiento académico puede ser medido de acuerdo con las notas obtenidas en las diferentes evaluaciones y en cursos relacionados. Otros estudios evalúan el rendimiento académico como la capacidad de los estudiantes para culminar satisfactoriamente sus estudios y egresar de la universidad (Shahiri et al, 2015).

Tabla 1
Matriz FODA EDM

Fortalezas	**Debilidades**
• Gran volumen de datos disponibles. • Uso de algoritmos poderosos y validados ya existentes. • Múltiples formas de visualizar la data. • Modelos más precisos de los usuarios para la mejora y personalización de los sistemas. • Encontrar momentos críticos y patrones de aprendizaje. • Obtener una visión de las estrategias de aprendizaje y sus resultados.	• Errores en la interpretación de los resultados debido a factores humanos. • Fuentes de datos heterogéneos. No existe todavía un estándar para los datos. • Los resultados en su mayoría son cuantitativos. Los métodos cualitativos no han brindado resultados significantes. • Sobrecarga de información. Sistemas complejos. • Incertidumbre debido a que solo los docentes o instructores con cierto nivel de habilidades pueden interpretar correctamente los resultados.
Oportunidades	**Amenazas**
• Estandarización de la data y mejora de compatibilidad entre las diferentes aplicaciones y herramientas. • Aprendizaje multimodal y afectiva. • Capacidad de auto-aprendizaje en sistemas inteligentes y autónomos. • Integración de los resultados obtenidos con otros sistemas de toma de decisiones. • Modelo de aceptación, describiendo usabilidad, expectativas, confiabilidad, entre otros.	• Aspectos éticos como privacidad de los datos. • Sobre-análisis. • Posibilidad de errores en la clasificación de patrones. • Confiabilidad: resultados contradictorios durante la implementación de modelos ya establecidos.

Fuente: Papamitsiou y Economides (2014)

Tabla 2
Aplicación de Minería de Datos

Tarea	Objetivo/Descripción	Aplicaciones claves
Predicción	Inferir la variable objetivo desde la combinación de otras variables. Se aplican técnicas de clasificación, regresión y estimación de densidad.	Predicción de rendimiento académico y detectar comportamiento de estudiantes.
Agrupamiento	Identificar grupos con características similares.	Crear grupos de estudiantes con características similares en su patrón de interacción y aprendizaje.
Minería de relaciones	Estudio de relaciones entre variables y descubrimiento de reglas.	Identificar relaciones en patrones de aprendizaje y diagnosticar dificultades en aprendizaje.
Destilación de datos	Representación de datos de manera más inteligible mediante resumen, visualización e interfaces interactivos.	Brindar herramientas de apoyo a los instructores para visualizar y analizar actividades relacionadas con sus estudiantes.
Descubrimiento con modelos	Emplear modelos validados previamente a nuevos datos.	Identificación de relación entre el comportamiento y las características de los estudiantes o de variables contextuales.
Detección de anomalías	Detección de individuos con diferencias significativas.	Detección de estudiantes con dificultades o con aprendizaje irregular
Minería de texto	Extracción de información de calidad desde los datos.	Análisis de contenido de foros, chats, páginas web y documentos.
Seguimiento de conocimiento	Estimar el dominio de habilidades de estudiantes, utilizando modelos cognitivos y los registros de la evaluación como evidencia.	Monitorear el nivel alcanzado por los estudiantes en el tiempo.
Factorización de matriz no negativa	Definición de matrices en base a resultados de evaluación que se descomponen en otras matrices que representan habilidades obtenidas.	Evaluación de habilidades.

Fuente: Romero y Ventura (2013), Shahiri y Husain (2015).

Estudios previos de factores que afectan el rendimiento académico

Se han realizado estudios de los factores que tienen alguna relación con el rendimiento académico de los estudiantes, los cuales se detallan a continuación:

- **Promedio general y semestral**

El promedio general y semestral es utilizado como atributo principal para realizar predicciones. La importancia de este atributo es que tiene un valor tangible que puede ser medido y utilizado, en muchos casos, como un indicador del potencial académico alcanzado (Shahiri, et al., 2015)

Los promedios son muy importantes desde el punto de vista de los alumnos. Por ejemplo, para conseguir o mantener una beca es necesario generalmente tener un promedio de 3.0 de 4.0 en el sistema norteamericano, ó 15 de 20 convirtiéndolo al sistema peruano y también al momento de buscar trabajo, en donde se encontró que el 77% de las empresas evalúan a los postulantes según su promedio (Honken y Ralston, 2013).

Estos promedios también pueden llegar a afectar otros aspectos como la motivación de los estudiantes y su percepción sobre la carrera tanto positiva como negativamente. Si un estudiante ve o siente que sus esfuerzos no se reflejan en sus notas podría incrementar el riesgo de que el estudiante abandone la carrera. Para evitar los efectos negativos de un promedio bajo en el primer año, algunas universidades están adaptando un sistema de aprobado/desaprobado que da la flexibilidad de mejorar su promedio si es que no le va muy bien al inicio, ya que el promedio recién se genera en el segundo año, cuando ya llevan cursos de carrera y están más involucrados con su futura profesión (Siller y Paguyo, 2012).

- **Rendimiento en el examen de Admisión**

Uno de los aspectos importantes por considerar, en una universidad, es el proceso de admisión. El objetivo principal de este proceso es determinar cuáles son los candidatos que tendrán un buen rendimiento luego de ser admitido. La calidad de los estudiantes admitidos tendrá una gran influencia

en los procesos educativos posteriores a admisión. Una correcta predicción del rendimiento académico de los postulantes ayudaría en el proceso de selección de las oficinas de admisión a detectar aquellos que podrían encontrar problemas para cumplir con las tareas académicas requeridas una vez admitido en la universidad. (Chen y Do, 2014).

En muchos estudios consideran el rendimiento del estudiante en el examen de admisión como una base para saber las habilidades y conocimientos adquiridos en las etapas anteriores a la universidad. Se ha encontrado que en los estudiantes de Ingeniería la calificación en matemáticas y ciencias es superior a estudiantes de otras carreras. En otros países, donde existen exámenes de admisión estandarizados como el SAT o ACT (en EE.UU.) es común utilizarlos como medida de las capacidades de los estudiantes y también para predecir el rendimiento de los ingresantes, saber si tienen las capacidades para cursar satisfactoriamente los primeros años de educación superior. También se ha hallado que un estudiante con buenas calificaciones, en estos exámenes, tiene seis veces mayor posibilidad de graduarse en los 4 años que dura la carrera que su contraparte (Brown, 2012).

- **Educación Secundaria**

El nivel de estudios alcanzados en secundaria es también un factor importante a considerar para medir el grado de conocimientos y la preparación del estudiante para enfrentarse a los retos de la universidad, en especial las calificaciones obtenidas en las clases de matemáticas, lenguaje y ciencias (Jin, Imbrie, Lin y Chen, 2011). Se ha encontrado que un estudiante que tiene un promedio general en secundaria de “A” (3.5 a 4.0 o 17.5 a 20) posee siete veces más posibilidad de graduarse en los cuatro años que dura la carrera en EE.UU. (Brown, 2012).

Para muchos los estudios que hayan realizado antes de la universidad son vitales para la preparación hacia los estudios de nivel superior. Los cursos que se han llevado y su rendimiento también son factores importantes por considerar ya que, en algunos casos, los estudiantes de Ingeniería han

sentido que quizás hayan podido aprovechar mejor su tiempo en el colegio y adquirido mayor conocimiento y mejores hábitos de estudio (Ecklund, 2013).

Específicamente en las carreras de Ingeniería y Tecnología, los cursos de matemática y ciencias de los primeros años sirven como bases para cursar satisfactoriamente los cursos de mitad y final de carrera. Aquellos estudiantes que tienen rendimientos bajos, en estos cursos, en muchos casos, han tenido problemas o simplemente no lograron culminar su carrera, por lo que el rendimiento en dichos cursos antes de ingresar a la universidad impactaría considerablemente una vez dentro de la universidad. (Li et al, 2013).

- **Auto-control**

Las investigaciones han encontrado relaciones importantes entre el auto-control y factores como salud y el éxito personal, pero también un efecto sobre el rendimiento académico. Un estudio realizado por Honken *et al* (2013), busca evaluar la relación entre las habilidades académicas y el auto-control, que los autores definen como la capacidad de controlarse en servicio de sus metas y estándares personales del estudiante. Esta investigación fue realizada a diferencia de otros estudios similares a estudiantes que han tenido muy buenos resultados académicos antes de ingresar a la universidad, pero que no logran cumplir con las expectativas dentro de ella. Los autores deciden analizar la parte de auto-control según su comportamiento en el pasado en donde demostraron una falta de auto-control con comportamientos irresponsables como: fumar, beber, quedarse dormido en clase y no entregar las tareas a tiempo.

Dentro de los resultados se encontró que los que respondieron que nunca han realizado estas acciones negativas tienen un mejor promedio que aquellos que dijeron que sí lo hicieron ocasional o frecuentemente. En donde más se nota es entre los que no entregaron las tareas a tiempo, la diferencia en el promedio entre los que siempre entregaron a tiempo y los que frecuentemente no entregaban sus tareas es de más de un punto en promedio (3.23 y 2.16, de un máximo de 4.0).

También se confirmaron los resultados de otros estudios que indican que no existe relación entre las notas del examen de admisión y la falta de auto-control. Es posible que aquellos que no tienen buenos hábitos de estudio, no entregan tareas a tiempo y no son buenos administrando su tiempo de estudio, pero logran obtener buenas notas en el examen de admisión e ingresar a universidades exigentes.

- **Confianza en sí mismo**

La confianza en sí mismo y en sus habilidades académicas también se les considera para evaluar las capacidades de los estudiantes. La confianza en sí mismo está muy relacionada con la motivación del estudiante a continuar con la carrera y seguir participando vivamente en las actividades de la universidad. Para las escuelas de Ingeniería se ha encontrado, en general, que los estudiantes tienen una buena percepción de sus habilidades en los cursos de matemáticas, ciencias y tecnologías, pero también se ha hallado que en algunos casos la percepción de sí mismo no encaja con la realidad y en temas que pensaban que dominaban llegan a tener problemas. En estos casos, la brecha entre la realidad y su percepción personal afecta negativamente tanto a la motivación como a la confianza del estudiante para continuar con sus estudios. (Jin *et al*, 2011).

Viéndolo desde otro punto de vista, se han encontrado diferencias de personalidad entre los estudiantes que son exitosos y los que tienen problemas durante sus estudios. Los estudiantes exitosos aceptan como responsabilidad individual el hecho de que sus acciones, comportamiento y creencias son las causas del porqué ocurren las cosas y también aceptan el hecho de que a veces necesitan ayuda y no tienen miedo o problema de solicitarlo. Los estudiantes con problemas en cambio creen que sus problemas son el resultado de fuerzas externas que no dependen de él y se sienten como víctimas y también tienen problemas para solicitar ayuda, creen que lo pueden hacer todo ellos mismos y ven como una debilidad el hecho de solicitar apoyo a alguien más. (Shull, Ford, y Carrier, 2013).

- **Expectativas**

Las expectativas que posean los estudiantes con respecto a su experiencia universitaria puede tener un impacto positivo si la universidad ofrece lo que espera y mucho más, o negativo si es que no se cumplen con las expectativas, debido a que existe mucha competencia, sus habilidades no son suficientes, entre otros. También se podría considerar la expectativa como el propósito por la que uno atiende a estudios superiores. Se ha encontrado, por ejemplo, que un estudiante que atiende la universidad en busca de más conocimientos y habilidades tiene mayor posibilidad de persistir con sus estudios que otro que solo busca adquirir habilidades para ejercer una profesión. (Brown, 2012).

- **Género**

Con respecto al género, existe cada vez menos brecha entre los hombres y mujeres con respecto al rendimiento académico y sus habilidades. También se ha encontrado que las mujeres han superado a los hombres en lo que respecta a la obtención de grados académicos en estudios superiores. (Park, Behrman & Choi, 2013).

Es uno de los factores más utilizados dentro de la predicción de rendimiento académico ya que está demostrado que entre los estudiantes del género masculino y femenino existen diferencias en el estilo de aprendizaje y en el comportamiento. Las mujeres son más disciplinadas, pueden permanecer concentradas por mayor tiempo, organizarse y mantener un nivel de motivación adecuado para enfrentarse a los retos de estudios universitarios (Shahiri *et al*., 2015).

- **Inteligencia emocional**

La inteligencia emocional puede ser utilizada como un factor a través del correcto procesamiento de datos que genera una carga emocional en los estudiantes. Sentimientos como empatía y el utilizar de una manera positiva estos sentimientos en las relaciones con otras personas tiene un impacto en la salud emocional y como consecuencia, en el rendimiento. Del estudio

realizado se ha encontrado que la capacidad de resolver problemas, realizar pruebas en la realidad, la responsabilidad, el asertividad y las relaciones interpersonales son importantes en la predicción del rendimiento académico. (Sepehrian, 2012)

Otro aspecto importante de la parte emocional es que en la universidad, además de adquirir todos los conocimientos, se espera que al final de la carrera logren conseguir un empleo. Para los evaluadores es importante además de un buen rendimiento académico una buena actitud hacia el aprendizaje, capacidad de adaptarse a nuevas situaciones, auto eficiencia, compromiso hacia sus labores, entre otros. La actitud hacia los estudios y la habilidad de poder adquirir nuevos conocimientos de manera independiente son factores importantes en relación con el rendimiento académico (Kumar y Vijayalakshmi, 2012).

De los distintos factores de personalidad, el más asociado con el éxito académico es el factor responsabilidad. A nivel comportamental, los estudiantes responsables tienen más probabilidades de poseer buenas habilidades para organizar su trabajo, proponerse metas y monitorear sus progresos. En este sentido, los educadores podrían fomentar estas características en los estudiantes enseñando estrategias cognitivas comportamentales asociadas con estos rasgos. Por ejemplo, los maestros pueden reforzar el comportamiento responsable de los estudiantes, premiando con créditos extras a aquellos estudiantes organizados y auto disciplinados que entregan sus deberes en tiempo y forma. (Cupani, Garrido y Tavera, 2013)

- **Motivación**

Las actividades de enseñanza y aprendizaje de una entidad educativa se producen en un contexto social concreto, que están definidos por ciertas reglas sociales que regulan el aspecto social de un salón de clases y pueden existir en los estudiantes diferentes tipos de metas, tanto académicas como sociales que no son mutuamente excluyentes. La motivación por asistir a la universidad puede darse desde un punto de vista social (interactuar con sus

amigos), académico (aprender cada vez más nuevos temas) y psicológico (querer destacar entre sus compañeros, obtener aprobación de alguna persona). Esta motivación o metas que desean alcanzar es medible a través de cuestionarios y pueden ser utilizados como factores en la predicción del rendimiento académico. (Martínez, Llorca, Tello y Mira, 2016).

- **Influencia de sus pares**

Los estudiantes pueden ser influenciados por las decisiones que tomen sus compañeros, mientras existan más alumnos que quieran estudiar ingeniería en su salón o colegio, son más propensos a estudiar la misma carrera. Asimismo, los compañeros de estudios sufren un impacto en el rendimiento de los estudiantes. En este estudio, se analizaron las relaciones de amistad e interacción entre los estudiantes mediante redes sociales, correo electrónico y otros medios electrónicos. Como resultado, se han encontrado que las habilidades de comunicación y las interacciones entre los estudiantes tienen un impacto sobre el rendimiento. Un caso interesante hallado en este estudio es el de un grupo de estudiantes que no tenían buen rendimiento ha logrado mejorar y graduarse satisfactoriamente debido a que tenían en común amistades que eran académicamente sobresalientes, por lo que los autores asumen que el hecho de estar rodeados de buenos estudiantes y estudiar con ellos o simplemente tener amigos que pueden resolver alguna duda que posean influye positivamente. (Bayer, Bydzovska, Geryk, Obsıvac y Popelinsky, 2012).

Los compañeros de la universidad posibilitan al estudiante disfrutar de las actividades y la conexión que realiza con otras personas con la que comparten intereses y formas de pensar. Estar en este tipo de comunidad permite al estudiante pensar y repensar sobre lo aprendido en sus clases y reforzar lo estudiado mediante el intercambio de ideas con sus compañeros. También sirven como apoyo emocional cuando algo no sale bien o en situaciones de estrés (cuando les iba mal en un examen), los compañeros frueon el primer nivel de soporte. Es por ello que las instituciones están cada vez más interesadas en formar comunidades y grupos de estudios para crear

lazos entre los estudiantes y dar un sentido de pertenencia a la universidad. (Budny, Paul y Newborg, 2013).

Técnicas de minería de datos utilizadas en la predicción de rendimiento académico

El objetivo de una predicción es inferir el valor de un atributo o algún aspecto de los datos desde una combinación de otros aspectos dentro de los datos. En EDM, la predicción normalmente se realiza a través de la creación de modelos que se realizan mediante técnicas como clasificación, regresión y categorización. La técnica más utilizada, en la tarea de predicción, es la clasificación. De los algoritmos que se pueden utilizar en clasificación se encuentran: árbol de decisiones, redes neuronales, bayesiano, k-nearest neighbor y Suppor Vector Machine. (Romero *et al*., 2013; Shahiri *et al*., 2015).

A continuación, se detallan las investigaciones encontradas en la revisión de la literatura que corresponden a la técnica utilizada:

- **Árbol de decisiones**

Es una de las técnicas más populares en predicción por su simplicidad y facilidad de interpretación dada su estructura y que puede ser convertida directamente en un conjunto de reglas SI-ENTONCES. (Shahiri et al., 2015).

En el estudio de Mishra *et al* (2014), se realizó la predicción de rendimiento académico de estudiantes en el tercer semestre de una carrera de computación. La predicción se efectuó a través de dos algoritmos de árbol de decisiones: J48 y Random Tree. Se utilizaron como atributos para realizar la predicción datos sobre integración social, integración académica y habilidades emocionales. Como resultado de esta investigación se ha encontrado que el resultado obtenido en el semestre anterior es el más importante al momento de realizar la predicción. Asimismo, un buen rendimiento constante en semestres anteriores y también en otros cursos es un buen indicador de resultados positivos a futuro. De los atributos emocionales, el liderazgo y la motivación tienen un impacto en el rendimiento

y las condiciones socio económicas no influyen mucho. Con respecto a los algoritmos aplicados tanto J48 como Random Tree han obtenido resultados satisfactorios logrando una exactitud de 88.372% (J48) y 94.418% (Random Tree).

En otro estudio, realizado por Elakia y Aarthi (2014), aplica los algoritmos ID3, C4.5 y CHAID para analizar las características de estudiantes de nivel secundaria que están por postular a una institución de educación superior. La predicción se realizó para predecir a base de las características de los estudiantes que carreras serían las que tendrían un mejor rendimiento después del ingreso a la universidad. Asimismo, se analizó el comportamiento durante el colegio para predecir factores relacionados con la disciplina de los estudiantes.

Con respecto a los algoritmos aplicados a los resultados se observan en la Figura 2. El algoritmo ID3 obtuvo mejores resultados en comparación con los otros dos algoritmos utilizados

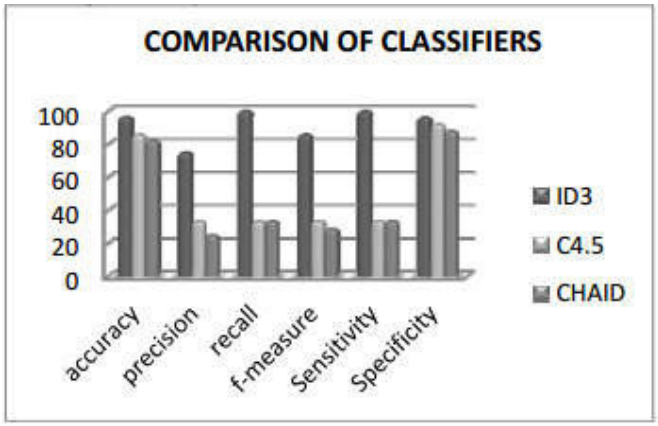


Figura 2. Comparación de Algoritmos de clasificación ID3, C4.5 y CHAID

Fuente: Elakia y Aarthi (2014)

Yadav *et al.* (2012) también realiza una comparación de múltiples algoritmos de árbol de decisiones (ID3, C4.5, CART y ADT) para predecir el bajo rendimiento de los estudiantes y la probabilidad de que un estudiante abandone sus estudios. Los resultados obtenidos se muestran en la tabla 3, en donde se observan resultados parecidos al estudio anterior Elakia *et al.*

(2014) en que el algoritmo ID3 obtuvo una mayor exactitud en la predicción con 85.7% de los casos detectó correctamente la posibilidad de abandono por parte de los estudiantes.

Ramesh *et al.* (2013) utilizaron diferentes algoritmos para predecir el rendimiento en las evaluaciones finales de diferentes cursos como un mecanismo para identificar a aquellos que podrían tener problemas para aprobar los cursos. Como resultados encontraron que el tipo de colegio que asistió a nivel secundario no influye en el rendimiento y la ocupación de los padres tiene un gran impacto en la predicción.

Tabla 3
Comparación de Resultados de algoritmos de clasificación

Algoritmo	Clasificados correctamente	Clasificados incorrectamente
ID3	90.9091%	3.6364%
C4.5	89.0909%	10.9091%
CART	86.0606%	13.9394%
ADT	87.2727%	12.7273%

Fuente: Yadav y Pal (2012)

Singh y Kaur (2016) realizan una comparación de técnicas de clasificación para predecir el rendimiento académico de estudiantes de la carrera de ingeniería de computación de la Universidad Punjabi. Como variables se utilizaron las condiciones sociales como tipo de vivienda, nivel de educación alcanzado por los padres, etc. y rendimiento académico en cursos llevados previamente, tanto en colegios como en universidades. Se utilizaron algoritmos J48 y REP Tree para realizar la predicción y obtuvieron una exactitud de 67.3729% en J48 y 56.7797% en REP Tree.

Gray, McGuinness y Owende (2014) evaluaron la capacidad de predicción de diferentes técnicas de clasificación para predecir el rendimiento académico en el primer año de estudios, basado en datos obtenidos antes o durante sus estudios. Como variables se utilizaron información obtenida durante el proceso de admisión como género, edad, rendimiento durante secundaria y elementos de aspecto emocional obtenidos mediante test de personalidad, motivación y estilo de aprendizaje. La exactitud de la predicción,

en este estudio, tuvo resultados variados de acuerdo con la edad de los estudiantes. Para el caso de los estudiantes más jóvenes, menores a 21 años, se obtuvo mayor exactitud en la predicción que los mayores de 21 años. Asimismo, los clasificadores bayesianos y regresión logística tuvieron resultados pobres en comparación con otras técnicas. Dado que estas técnicas son lineales, se puede asumir que el modelo es complejo y difícil de representar con una función lineal.

- **Redes neuronales**

Las redes neuronales constituyen otra técnica popular utilizada para la predicción. Esta tiene como ventaja la gran capacidad de predicción que presenta cuando es entrenada correctamente y puede ajustarse a gran cantidad de datos. La desventaja es que el modelo que se obtiene es de caja negra, es decir, es muy difícil de interpretar para un ser humano. (Lantz, 2013).

De la literatura destaca el estudio realizado por Arsad, Buniyamin y Ab Manan (2013) que aplicaron las redes neuronales para predecir el rendimiento de estudiantes de octavo semestre de una carrera de Ingeniería, basado en las notas obtenidas en los cursos base llevados en los primeros semestres. Se obtuvo una exactitud mayor a 90% al realizar la predicción. Se puede rescatar de los resultados de este estudio que en las carreras de Ingeniería las asignaturas llevadas en los primeros semestres tienen un impacto en las asignaturas de fin de carrera, ya que comúnmente en estas carreras se van profundizando hacia temas cada vez más avanzados que requieren combinar las diferentes habilidades, obtenidas en los cursos base. Un bajo promedio en estas asignaturas sirve como un indicador de la falta de dominio de los temas y da como resultado un bajo rendimiento en cursos posteriores.

El estudio de Kumar *et al*. (2012) aplica el algoritmo de perceptron multicapa para la predicción del rendimiento académico de un programa de maestría. Este algoritmo perceptron consiste en una serie de capas (entrada, oculta y salida) que en su conjunto forman un modelo de predicción. La capa de entrada alimenta al modelo en los múltiples nodos de la capa oculta que

procesa toda la información y presenta los resultados en la capa de salida. Los resultados de este estudio muestran una exactitud de 96.4% en la predicción a través de redes neuronales.

El estudio de Gray *et al*. (2014) también aplicó las redes neuronales para realizar la predicción de rendimiento académico de estudiantes del año 2012 a base de los datos de los años 2010 y 2011. Para este estudio, las redes neuronales no lograron distinguir correctamente los estudiantes con buen rendimiento y bajo rendimiento académico.

Chen *et al.* (2014) desarrollaron múltiples algoritmos basados en sistemas de inferencia neuro difuso adaptativo jerárquicos (ANFIS) cambiándolos con algoritmos genéticos y *cuckoo search.* ANFIS tiene una arquitectura de cinco capas con un aprendizaje híbrido que puede ser entrenado eficientemente. Como resultado en todas las versiones de los algoritmos aplicados se obtuvo que el porcentaje de los ítems en donde la predicción fue correcta se encuentra entre 89.27% a 81.99%, lo cual demuestra una gran capacidad en la predicción del rendimiento en estos algoritmos más avanzados.

- **Support Vector Machines (SVM)**

Es un método de aprendizaje supervisado que puede ser utilizado para tareas de clasificación y predicción (Lantz, 2013). El estudio de Sembiring, Zarlis, Hartama, Ramliana, y Wani (2011) utiliza algoritmos de SVM y clustering para encontrar relación entre el rendimiento académico y factores relacionados con aspectos psicométricos como intereses, creencias, soporte familiar, actitud y tiempo dedicado hacia los estudios, entre otros. La data se obtuvo de una encuesta tomada a los estudiantes en donde debían responder preguntas relacionadas con los factores antes mencionados y la nota que obtuvieron al finalizar el curso que se dividió en cinco clases: Excelente, muy bueno, bueno, regular y pobre. La predicción a través de SVM logró obtener una exactitud de 93.67% para identificar a los estudiantes con nota “pobre” como la mejor y en el otro extremo, 61% para los estudiantes “regular”. Como conclusión, los autores mencionan que el SVM tiene una

buena capacidad para generalizar y es más veloz que las otras técnicas de predicción

En la comparación de técnicas de Gray *et al*. (2014) también se aplicaron técnicas de SVM y fue la que obtuvo los mejores resultados al momento de predecir a aquellos estudiantes en peligro de desaprobar. Sin embargo, en estos algoritmos de aprendizaje automatizado los autores recomiendan tener cuidado con el sobreajuste o *overfitting.*

Sobreajuste es un problema que suele aparecer en este tipo de algoritmos en donde los datos que presentan ruido, es decir, que no se ajustan al modelo o no ayudan al momento de entrenar el modelo influyen en los resultados. Como consecuencia del sobreajuste, en algunos casos, también llamado sobre entrenamiento, es que el modelo comienza a perder su validez ya que se termina ajustando más al ruido que a los datos reales. (Lantz, 2013).

2.2 Bases teóricas

Rendimiento académico

El rendimiento académico puede ser definido o explicado desde diferentes puntos de vista, para unos es el nivel de conocimientos demostrado por una persona en una área o materia comparado con la norma de edad y nivel académico. También puede ser entendido como la capacidad de un estudiante a responder a las exigencias de la currícula de una universidad, considerado por algunos como simplemente cumplimiento de las obligaciones de una universidad o institución educativa. (Brito-Jiménez y Palacio-Sañudo, 2016) (Rodríguez, y Arenas, 2016).

Múltiples estudios también consideran al rendimiento académico como un equivalente a los promedios y notas obtenidas en los cursos. La capacidad de un estudiante de cumplir con las exigencias establecidas para los cursos dictados en una universidad y lograr superar dichos retos se verá reflejado en sus notas y se le podrá considerar como un alumno con un buen rendimiento

académico. Por otro lado, aquellos que no demuestran los conocimientos o capacidades y habilidades para aprobar dichos requerimientos presentarán un bajo rendimiento académico, que podrían favorecer el abandono o deserción. (Calisir, Basak, y Comertoglu, 2016) (Rodríguez, y Arenas, 2016).

Hay diversos aspectos que pueden influir en el éxito de los estudiantes en cuanto a su rendimiento académico. Así, por ejemplo, Rasberry, Lee, Robin, Laris, Russell, Coyle y Nihiser (2011) reagrupan el rendimiento académico en tres apartados que pueden ser: (1) las actitudes y habilidades cognitivas, que incluyen, entre otros aspectos, la atención, memoria, comprensión verbal, procesamiento de información, motivación, autoconcepto y satisfacción; (2) los comportamientos académicos, que abarcan, entre otros, a la organización, planificación y asistencia; y (3) los resultados académicos, que incluyen las puntuaciones en las materias de enseñanza.

El rendimiento académico es utilizado también como una medida para evaluar las habilidades de los ingresantes a una universidad. El rendimiento durante sus estudios en secundaria es un factor clave para la selección de nuevos ingresantes. (Aluko, Adenuga, Kukoyi, Soyingbe y Oyedeji, 2016).

Desde un punto de vista práctico, el rendimiento académico es el resultado de múltiples factores y no solamente académicos, que se reflejan en la nota o promedio obtenido en los cursos. Sin embargo, es importante considerar que los resultados pueden ser en forma inmediata y diferida. Los primeros se refieren a las calificaciones que obtienen los alumnos durante sus estudios hasta la consecución del título correspondiente y se definen en términos de éxito o fracaso en relación con un período temporal determinado. En cambio, los resultados diferidos hacen referencia a la conexión con el mundo laboral y se precisan en términos de eficacia y productividad vinculados, sobre todo, con criterios de calidad de la institución educativa. (López Bonilla, López Bonilla, Serra, y Ribeiro, 2015).

Tanto los resultados inmediatos como los diferidos pueden ser usados como criterios para evaluar si un estudiante es exitoso. El éxito académico es un aspecto importante tanto desde el punto de vista de los estudiantes como de

la institución educativa. Es un aspecto principal de la reputación de una universidad en que sus estudiantes sean exitosos tanto durante y después de sus estudios universitarios (Calisir *et al.*, 2016).

York *et al.* (2015) define al éxito académico como una combinación de múltiples factores: rendimiento académico, logro de objetivos de aprendizaje, adquisición de habilidades y competencias, satisfacción, persistencia y rendimiento post-universitario. El rendimiento académico se considera un factor clave para la medición del éxito de un estudiante ya que representa, en un valor medible (en forma de nota o promedio), su desempeño en las clases para alcanzar los criterios establecidos de aprendizaje y adquisición de competencias.

El modelo completo establecido por York *et al.* (2015) se encuentra en la Figura 3. Dos de cada tres artículos relacionados con el éxito académico utilizan el rendimiento académico, medido a través de notas y promedios, como la representación de un estudiante exitoso. El rendimiento académico y la retención son datos que se encuentran normalmente disponibles en cualquier universidad por lo que explica su popularidad como medida del éxito alcanzado.

Figura 3. Modelo detallado de Éxito Académico

Fuente: York, Gibson y Rankin (2015)

Minería de Datos

La minería de datos es un sistema de información basado en computación que explora grandes repositorios de datos para generar información y descubrir conocimiento. La palabra se origina en la minería tradicional; sin embargo, el objetivo es buscar conocimiento que permita descubrir elementos como: patrones interesantes, relación entre los datos, definir reglas, predecir valores desconocidos, agrupar objetos homogéneos y otros aspectos que son difíciles de descubrir en sistemas de información tradicionales (Peña-Ayala, 2014).

Es el proceso de descubrir nuevos patrones y conocimientos a partir de gran cantidad de datos. Las fuentes para estos datos pueden ser: bases de datos, data warehouse, internet, otros tipos de repositorios o datos dinámicos que ingresan directamente al sistema. La minería de datos aparece como una forma de cubrir la necesidad de analizar, de una manera automatizada e inteligente la gran cantidad de datos que se generan día a día. Terabytes o petabytes de data son generados por los diferentes tipos de redes y transacciones que se realizan desde los negocios, las personas, ciencia, ingeniería, medicina y otros aspectos de la vida diaria. (Han et al., 2012)

Esta gran cantidad de datos es demasiado extensa como para que pueda ser analizado por una persona, ya que estos repositorios muchas veces se convierten en “tumbas” que contienen información, pero son raramente utilizados. Consecuentemente, las grandes decisiones muchas veces son tomadas basándose más en la intuición de los encargados de la toma de decisiones, sin tener información que pueda sustentarla. El desarrollo de nuevas técnicas y herramientas de minería de datos logran que la brecha entre datos e información sean cada vez menores y permitan convertir estos datos en conocimiento (Han et al., 2012).

En el fondo, el objetivo principal es encontrar algún significado de datos complejos. La minería de datos ha sido aplicada, por ejemplo para:

- Predecir el resultado de elecciones.
- Identificar y filtrar correos electrónicos no deseados.

- Rastrear actividad criminal.
- Automatizar señales de tránsito según condiciones de tráfico.
- Realizar estimaciones de tormentas y desastres naturales.
- Examinar el comportamiento de clientes.
- Identificar posibles participantes a eventos.
- Seleccionar publicidad según el tipo de cliente (Lantz, 2013).

Según Han *et al*. (2012), la minería de datos es considerada como un sinónimo de descubrimiento de conocimiento en bases de datos (knowledge discovery from data o KDD) que es una secuencia interactiva e iterativa basados en los siguientes pasos:

- Comprender el dominio de aplicación: este paso incluye el conocimiento relevante previo y las metas de la aplicación.
- Extraer la base de datos objetivo: recogida de los datos, evaluar la calidad de ellos y utilizar análisis exploratorio de los datos para familiarizarse con ellos.
- Preparar los datos: incluye limpieza, transformación, integración y reducción de datos. Se intenta mejorar la calidad de los mismos a la vez que disminuir el tiempo requerido por el algoritmo de aprendizaje aplicado, posteriormente.
- Minería de datos: como se ha señalado anteriormente, esta es la fase fundamental del proceso. Está constituido por una o más de las siguientes funciones: clasificación, regresión, clustering, resumen, recuperación de imágenes, extracción de reglas, etc.
- Interpretación: explicar los patrones descubiertos, así como la posibilidad de visualizarlos.
- Utilizar el conocimiento descubierto: hacer uso del modelo creado.

Técnicas de Minería de datos

Existen diferentes tipos de conocimiento o patrones que pueden ser descubiertos a través de la minería de datos. Para ello, se aplican las

diferentes técnicas o métodos que se clasifican en dos categorías: descriptivo y predictivo. Los métodos descriptivos permiten encontrar características de los datos y los métodos predictivos facilitan realizar inducción en los datos actuales para realizar predicción de valores futuros (Han *et al*., 2012).

A continuación, se describen las diferentes técnicas que se aplican en los datos para el descubrimiento de patrones:

- **Clasificación**: clasifica un dato dentro de una de las clases categóricas predefinidas. Responde a preguntas tales como: ¿Cuál es el riesgo de conceder un crédito a este cliente?, dado este nuevo paciente ¿qué estado de la enfermedad indican sus análisis?
- **Regresión**: el propósito de este modelo es hacer corresponder un dato con un valor real de una variable. Responde a cuestiones como ¿cuál es la previsión de ventas para el mes que viene?, ¿de qué depende?
- **Clustering**: se refiere a la agrupación de registros, observaciones, o casos en clases de objetos similares. Un clúster es una colección de registros que son similares entre sí, pero distintos a los registros de otro clúster. ¿Cuántos tipos de clientes vienen a mi negocio?, ¿qué perfiles de necesidades se dan en un cierto grupo de pacientes?
- **Generación de reglas**: aquí se extraen o generan reglas de los datos. Estas reglas hacen referencia al descubrimiento de relaciones de asociación y dependencias funcionales entre los diferentes atributos. ¿Cuánto debe valer este indicador en sangre para que un paciente se considere grave?, ¿si un cliente de un supermercado compra pañales también compra cerveza?
- **Resumen o sumarización**: estos modelos proporcionan una descripción compacta de un subconjunto de datos. ¿Cuáles son las principales características de mis clientes?
- **Análisis de secuencias**: se modelan patrones secuenciales como análisis de series temporales, secuencias de genes, etc. El objetivo es modelar los estados del proceso, o extraer e informar de la desviación y tendencias en el tiempo. ¿El consumo de energía

eléctrica de este mes es similar al del año pasado?, dados los niveles de contaminación atmosférica de la última semana cuál es la previsión para las próximas 24 horas (Han et al. 2012) (Lantz, 2013) (Sin y Muthu, 2015).

CRISP-DM

Cross-Industry Standard Process for Data Mining o CRISP-DM es una metodología y un modelo de proceso que define un ciclo de seis tareas principales para establecer un marco de trabajo del ciclo de vida de minería de datos (Wirth y Hipp, 2000). El modelo se muestra en la Figura 4.

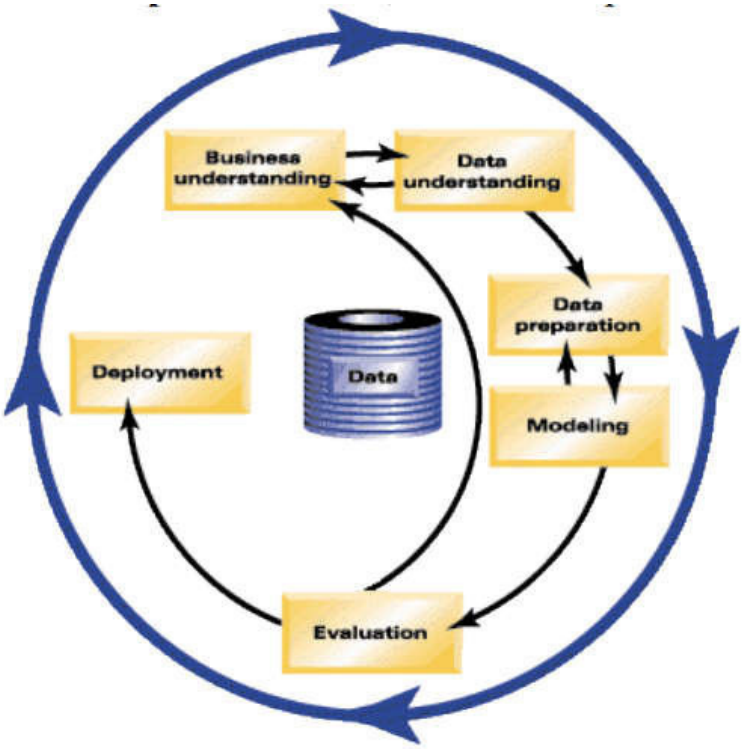


Figura 4. Fases del modelo CRISP-DM

Fuente: Wirth y Hipp (2000)

Las flechas, en el modelo, indican las dependencias más importantes y frecuentes entre las fases. El modelo presenta un proceso iterativo por lo que la secuencia de las fases no es estricta y es posible avanzar y retroceder entre las fases tantas veces como sea necesario. Se busca que el modelo sea

flexible y pueda ser adaptado en cualquier situación independiente de la industria, la herramienta y la aplicación de minería de datos (Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer, y Wirth, 2000).

Una guía visual detallando todos los pasos incluidos en cada uno de los procesos con sus subprocesos y outputs se encuentra en la figura 5.

Según la guía de CRISP-DM de Chapman et al. (2000) los detalles de cada proceso y subproceso del modelo son las siguientes:

- **Comprensión del negocio**

  Esta fase inicial se enfoca a la comprensión de los objetivos del proyecto y los requerimientos desde el punto de vista del negocio, con el objeto de convertir esta información en la definición del problema de minería de datos y un plan de proyecto preliminar para alcanzar dichos objetivos. Incluye los subprocesos de: determinar los objetivos del negocio, evaluar la situación, establecer las metas de minería de datos y producir un plan de proyecto.

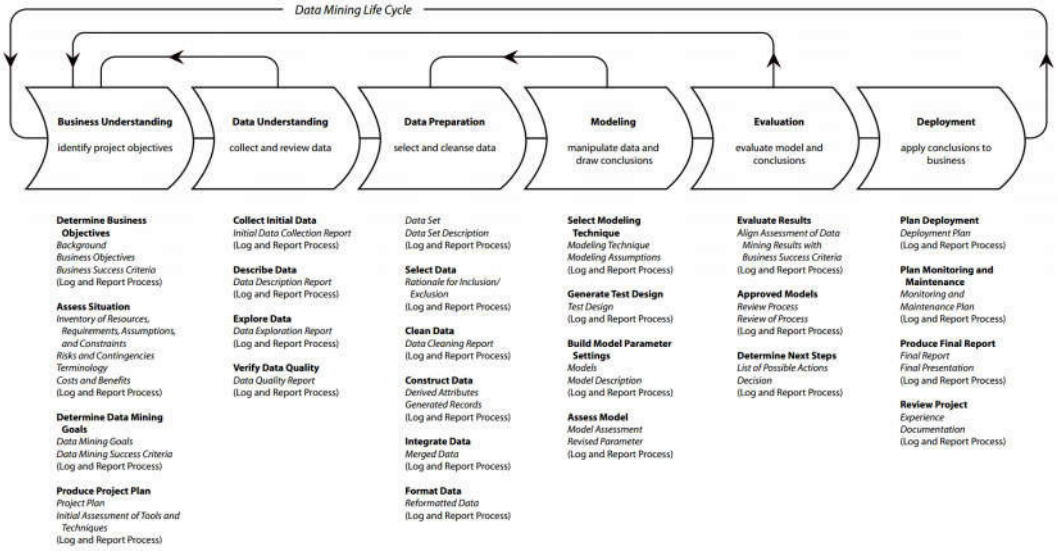


Figura 5. Modelo detallado de las fases de la modelo CRISP-DM

Fuente: Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer, y Wirth (2000)

- **Comprensión de los datos**

  Esta fase se inicia con la recolección de datos inicial y continúa con las tareas a fin de familiarizarse con los datos, identificar problemas de calidad, establecer una visión general y detectar alguna información oculta. Existe una relación cercana entre la comprensión del negocio y el entendimiento de los datos. Para poder formular el problema de minería de datos y el plan del proyecto, es necesario tener algún conocimiento de los datos disponibles.

- **Preparación de los datos**

  Cubre todas las actividades relacionadas con la construcción del conjunto de datos a utilizarse para la minería desde los datos iniciales. En muchos casos, estas tareas se realizan múltiples veces y no necesariamente en algún orden específico. Los subprocesos incluyen: selección, limpieza, construcción, integración y dar formato a los datos.

- **Modelamiento**

  En esta fase, múltiples técnicas son seleccionadas y aplicadas, calibrando los parámetros para optimizar los resultados. Típicamente, existen múltiples técnicas para un tipo de problema de minería de datos. Algunas requieren un formato específico de los datos. La fase de preparación y modelamiento están relacionados. La preparación de los datos genera ideas de las técnicas que se podrían aplicar para resolver los problemas frecuentemente.

- **Evaluación**

  Para esta etapa, ya deberían de existir uno o más modelos que parecen cumplir aspectos de calidad desde el punto de vista del análisis de datos. Antes de proceder con la fase final de despliegue es importante realizar una evaluación más profunda del modelo y revisar los pasos utilizados en la construcción para asegurarse de que

cumplen con los objetivos establecidos. Un punto importante es determinar si es que existe algún problema del negocio que no se haya considerado lo suficiente. Al finalizar esta fase, la decisión de utilizar los resultados de la minería de datos debe establecerse.

- **Despliegue**
  Los proyectos generalmente no finalizan con la creación de los modelos. Usualmente, el conocimiento obtenido debe ser organizado y presentado para que el cliente pueda utilizarlo. Dependiendo de los requerimientos la fase de despliegue puede ser tan simple como generación de reportes o tan compleja como la creación de un proceso repetible. En muchos casos, el usuario o el cliente será el encargado de definir y realizar las tareas de esta fase.

Educational Data Mining

Dentro de los campos en los que se aplica el DM, existe uno en particular que utiliza estas técnicas con el tipo de datos que se encuentran en el campo de la educación, conocida como Educational Data Mining (EDM).

Es una disciplina emergente que busca, con la participación de investigadores de distintas áreas (computación, estadística, minería de datos y educación), investigar cómo la minería de datos puede mejorar el campo de la educación y facilitar las investigaciones que se realizan. EDM se enfoca en el desarrollo de métodos que permitan explorar los tipos de datos únicos que se presentan en este rubro.

Estos datos pueden provenir desde diferentes fuentes, incluyendo información de ambientes tradicionales de salón de clases, software educativo, cursos en línea, evaluaciones sumativas tipo examen de admisión, etc. Estas fuentes de datos proveen cada vez más información que pueden ser analizadas para responder preguntas que no eran factibles anteriormente como buscar diferencias entre la población de los estudiantes o comportamientos inesperados por parte de un grupo de estudiantes (Romero *et al.*, 2011).

Según Romero *et al.* (2013), estas contribuciones en educación se apoyan en los avances realizados en el campo de la minería de datos aplicados en el pasado en otros rubros como comercio y biología. Aunque existen similitudes en las metodologías utilizadas la aplicación de la minería de datos para el campo de la educación presenta las siguientes diferencias:

- **Dominio**. En el comercio, la minería de datos se utiliza generalmente como una forma de influenciar en las decisiones de los consumidores, mientras que el propósito de los sistemas educativos es guiar los estudiantes en el aprendizaje.

- **Datos**. En el comercio electrónico, normalmente se limitan a las acciones que realiza el cliente dentro de su sistema o página web. En cambio, en EDM existe gran cantidad de información disponible de los estudiantes provenientes de repositorios tradicionales de datos, observaciones en campo, encuestas sobre motivación, entre otros. Se generan también diferentes tipos de datos según el lugar o ambiente en donde se realizan las clases (salón de clases tradicional, clases en laboratorio de cómputo o clases virtuales) y el sistema de información que soporta el desarrollo de las clases. Se están recopilando diferentes tipos de datos como perfil de estudiantes, actividad o interacción con los sistemas de información, tareas realizadas, objetivos alcanzados y muchos más.

- **Objetivo.** En el comercio, el objetivo principal es incrementar las utilidades, que es una medida tangible. EDM tiene como objetivo principal mejorar la educación y el proceso de aprendizaje, cuya medición es compleja por ser una medida no tangible y, por lo tanto, es necesario realizar estimaciones.

- **Técnicas**. Se han logrado aplicar exitosamente las técnicas tradicionales de minería de datos en el rubro de la educación, pero es necesario tener en consideración que existen jerarquías y dependencias entre los datos, ya que los estudiantes generan los datos mientras avanzan en su carrera y son influenciados por sus compañeros y profesores. Como consecuencia existen técnicas que pueden ser adaptados a ser usados y otros, no.

La aplicación de técnicas de minería de datos presenta información útil que puede ser utilizada por las instituciones educativas como base pedagógica en la toma de decisiones sobre modificaciones a realizarse en el ambiente o enfoque de la enseñanza. La aplicación de minería de datos en educación es un proceso iterativo de establecimiento de hipótesis, pruebas y mejoras. Los conocimientos extraídos vuelven a entrar al ciclo para guiar, facilitar y mejorar el proceso de aprendizaje en su totalidad (Romero *et al*., 2013) (Romero *et al*., 2011).

El objetivo principal del EDM es mejorar el aprendizaje a través de la toma de decisiones, basados en información generada del conjunto de datos que existen en el ámbito de la educación. Sin embargo, existen varios objetivos específicos que varían según el usuario final y el problema que quiere resolver. Algunos ejemplos son:

- ¿Cómo organizar clases o tareas? o ¿qué materiales entregar, basados en el uso y rendimiento?
- ¿Cómo identificar a aquellos alumnos que se beneficiarían de algún consejo sobre sus estudios u otro tipo de ayuda?
- Decidir qué tipo de ayuda sería más efectiva para los estudiantes con algún tipo de problema.
- ¿Cómo ayudar a encontrar material de estudio útil para sus estudios?, ya sea de manera individual o colectiva.

La data que se obtiene, en el ámbito de educación, es específica y diferente de lo que se obtiene en otras áreas. Es importante tomar en cuenta los aspectos pedagógicos y la interacción entre los estudiantes y docentes ya sea

de manera directa en un ambiente tradicional o de manera virtual (Romero *et al.*, 2013).

Existen varios ejemplos de aplicación o tareas en el ámbito de la educación que se pueden solucionar con el DM, que se mencionan a continuación:

- **Predecir el rendimiento del estudiante**. Estimar un valor desconocido de un estudiante, ya sea rendimiento, nivel de conocimiento o nota.
- **Preguntas científicas.** Desarrollar y probar teorías científicas sobre aprendizaje apoyado por la tecnología, formular nuevas hipótesis, etc.
- **Proveer una retroalimentación a los docentes.** Proveer retroalimentación sobre cómo mejorar la experiencia de aprendizaje y tomar acciones proactivas/reactivas para cada caso.
- **Personalización de información a los estudiantes.** Adaptar una automatización en aprendizaje, navegación, contenido, presentación, etc. para cada estudiante de forma personalizada.
- **Recomendaciones para los estudiantes.** Dar recomendaciones automatizadas sobre actividades o tareas, enlaces a visitar, problemas o cursos a llevar, etc.
- **Crear alertas para Stakeholders.** Monitorear el proceso de aprendizaje y detectar comportamientos no deseados de estudiantes en tiempo real, tales como baja motivación, distracciones, mal uso, trampa, etc.
- **Modelo de usuario/estudiante.** Desarrollar y afinar modelos cognitivos de estudiantes que representen sus habilidades y conocimientos.
- **Modelado de dominio.** Describir el dominio de los cursos en cuanto a conceptos, habilidades, material de aprendizaje y la relación entre ellos.
- **Agrupar/crear perfiles de estudiantes.** Crear grupos de estudiantes con características similares.

- **Desarrollo de material de clase.** Apoyar a los docentes en el proceso de desarrollo de material de clase y contenido de aprendizaje automáticamente.
- **Planificación.** Planificar cursos a futuro, establecer horarios, planificar recursos a asignar, procesos de admisión y consejería, desarrollo de currícula, etc.
- **Estimación de parámetros.** Obtener parámetros de modelos probabilísticos para predecir la probabilidad de una ocurrencia de interés.

Entre ellos, destaca predecir el desempeño de los estudiantes, que es el más antiguo y más popular entre las aplicaciones del EDM (Romero *et al.*, 2011).

Algoritmos de clasificación

Clasificación es una forma de análisis de datos que permite extraer modelos que describen las clases importantes de los datos. Estos modelos permiten predecir valores categóricos utilizando un proceso de dos pasos: de aprendizaje (en donde se construye el modelo) y de clasificación (en donde se utiliza el modelo para predecir los valores de otros datos) (Han *et al.*, 2012).

A continuación, se hace una breve descripción de los principales algoritmos de clasificación según lo mencionado por Lantz (2013) y Han *et al.* (2012):

Árbol de decisiones

Representa un conjunto de reglas de clasificación en forma de un árbol. Los primeros árboles de decisiones eran construidos por humanos expertos, pero en la actualidad es más común encontrar árboles construidos en base a un algoritmo. Entre los más conocidos están ID3 y C4.5. La idea básica detrás del algoritmo de árbol de decisiones es que, a partir de los atributos de cada clase alcanzará un punto final de una ruta.

Los árboles de decisiones presentan muchas ventajas: son simples y fáciles de entender, pueden trabajar con variables numéricas y categóricas, clasificar

nuevos datos de manera rápida, manejar ruido y valores faltantes de una clase con relativa facilidad y además, son flexibles. La desventaja, si la cantidad de datos de entrenamiento es poca, podría generar un modelo que fuese muy específico y difícil de generalizar.

La principal restricción que se encuentra en un árbol de decisiones es que se asume que es posible clasificar de manera determinística a todos los datos hacia una clase específica. Como resultado de todas las inconsistencias se consideran como errores, por lo que no es recomendado para clasificar, por ejemplo, rendimiento de los estudiantes en una clase, donde se pueden encontrar muchas inconsistencias en la data.

Clasificadores bayesianos

En una red bayesiana, las dependencias estadísticas se representan visualmente en una estructura gráfica. Se toma en cuenta que toda la información tiene independencia condicional y representan atributos con una estructura de dependencia mínima. Cada vértice del gráfico corresponde a un atributo y los bordes representan el grupo de atributos de la que dependen. La fuerza de la dependencia es representada por una probabilidad condicional. Para aplicar las redes bayesianas, a fin de realizar predicciones, es necesario primero entrenar a la red para determinar la estructura de dependencia de cada atributo y generar un modelo base sobre el cual se realizarán las predicciones de datos futuros.

En la práctica, se presenta como problema la gran cantidad de probabilidades que debemos estimar para establecer correctamente las probabilidades que se presentan en cada atributo. Para resolver este problema, se puede usar un modelo bayesiano ingenuo en donde se asume que todos los atributos son condicionalmente independientes y se reduce la cantidad de estimaciones que se tendría que realizar. En la práctica, el modelo bayesiano ingenuo ha mostrado buenos resultados en cuanto a la predicción, pero como consecuencia de la presunción que se realiza para aplicarla, el poder de

representación del modelo generado por este modelo es inferior al de, por ejemplo, árbol de decisiones.

Sin embargo, existen muchas ventajas en el modelo bayesiano ingenuo: es simple, eficiente, robusto al ruido y fácil de interpretar. Se adapta con gran facilidad a muestras pequeñas de datos, ya que combina una complejidad pequeña con un modelo probabilístico flexible. El modelo básico se adapta solo a datos discretos por lo que los datos numéricos deben ser transformados.

Redes neuronales

Son muy populares en cuanto a reconocimiento de patrones, pero en su aplicación en la educación pueden ser problemáticos, a menos que existan gran cantidad de datos numéricos y se cuente con los conocimientos de un experto para poder entrenar el modelo.

Las Feed-Forward neural network (FFNN) es el tipo de red neuronal más utilizado. Su arquitectura consiste de múltiples capas de nodos: una de entrada, otra de salida y una o más capas intermedias ocultas. Cada capa oculta está conectada a los nodos de la capa anterior y posterior y se asignan pesos a cada uno. En teoría es posible representar cualquier función en una red de tres capas, pero en la práctica entrenar una red presenta gran dificultad.

Como ventajas, las redes neuronales pueden representar modelos no lineales y en teoría, pueden representar cualquier tipo de clasificador. Si los datos originales no son discriminatorios, la FFNN lo puede transformar de manera implícita. Además, son robustos al ruido y actualizables con una nueva data.

La principal desventaja es la necesidad de gran cantidad de datos, mucho más de los que comúnmente se obtienen en un entorno educativo. Los datos deben ser numéricos y los valores categóricos tienen que ser cuantificados, de alguna manera, antes de utilizarse. Esto incrementa la complejidad del modelo y los resultados son sensibles al tipo de cuantificación que se haya utilizado.

El modelo de redes neuronales es una caja negra y por ello, también se presenta la dificultad de poder explicar de una manera simple los resultados obtenidos. Adicionalmente, tienden a ser inestables y solo presentan buenos resultados en las manos correctas. El entrenamiento del modelo puede ser un proceso que consume tiempo, especialmente si se quiere crear un modelo generalizado.

Clasificadores k-nearest neighbor

Representan un enfoque distinto a la forma de realizar una clasificación. No construyen un modelo generalizado, sino lo aproximan de manera local e implícita. La idea de fondo es clasificar a los nuevos objetos examinando los k datos más similares existentes en la muestra actual. La clase seleccionada puede ser la más común entre los datos cercanos o una distribución entre los valores más cercanos.

Las únicas tareas para realizar el aprendizaje del modelo es definir dos parámetros: la cantidad de vecinos o valores cercanos a evaluar k y la distancia d. El valor adecuado de k para cada modelo se puede deducir mediante pruebas de diferentes valores para k y su posterior validación en datos de prueba. Existen diferentes puntos por considerar en la selección de un valor para k adecuado para cada modelo ya que, si k es muy pequeño, entonces la muestra se realiza solamente con pocos datos, lo que resulta en un modelo inestable. Por otra parte, si k es muy grande, es posible que se desvíe mucho del valor que realmente debería representar.

La definición de la distancia *d* es una tarea mucho más crítica. Normalmente, las métricas toman en cuenta todos los atributos presentes, incluyendo las que se podrían considerar como irrelevantes, lo que podría resultar en una representación que se aleja de la realidad y un objeto puede terminar en un vecindario incorrecto. Para solucionar este problema se ha sugerido asignar pesos a cada atributo y también realizar una pre-selección de los atributos más relevantes.

Como ventaja se puede mencionar que solo existen dos parámetros que entrenar; muestran gran precisión para algunos datos y son robustas al ruido y valores faltantes. Tienen gran poder de representación ya que pueden adaptarse a cualquier modelo, dado que existen suficientes datos.

La dificultad aparece en la definición del valor para *d*. Los datos educacionales presentan tanto valores numéricos como categóricos y pueden existir diferentes escalas para los valores numéricos, por lo que es necesario contar con una gran cantidad de datos para definir correctamente el valor para d. Existen también muchos datos irrelevantes que tendrían que ser eliminados previamente.

La falta de un modelo explicito puede ser tomado como una ventaja o desventaja, dado que, si el modelo es muy complejo, es más fácil aproximarlo de manera local y representarlo de esa manera. También no es necesario actualizar el modelo cuando se adicionan nuevos datos, pero es necesario mencionar que estos modelos tienden a ser mucho más lentos que los enfoques basados en modelos generalizados.

Support Vector Machines

La máquina de vector de soporte es un algoritmo que construye un modelo que representa a los puntos de muestra, en una dimensión superior, en donde define un hiperplano que realiza una separación óptima de las dos clases que busca clasificar. Para la definición del hiperplano, el algoritmo utiliza los vectores de soporte que con el mapeo de los datos, en una dimensión lo suficientemente alta, permita realizar la clasificación.

Los SVM son ideales cuando los límites entre las clases no son lineales y no existe suficiente data para crear un modelo complejo. La idea de fondo es que, si los datos son mapeados hacia una dimensión superior, las clases se vuelven linealmente separables.

Se enfocan solamente en los límites o fronteras que pueden existir dentro de la muestra de datos. La meta es encontrar el plano con el mayor margen para

poder separar con una línea a los datos presentes. Se alcanzan mejores resultados cuando se establecen límites “flexibles” que permiten que existan algunos datos mal clasificados.

La principal ventaja es que siempre encontrará un óptimo global. La precisión de la clasificación no depende de la cantidad de datos y producen modelos simples a diferencias de otros paradigmas de clasificación. Sin embargo, presentan las mismas restricciones que las redes neuronales: los datos deben ser continuos numéricos, el modelo es difícil de interpretar y es difícil de definir los parámetros correctos para una representación adecuada.

2.3 Definición de términos básicos

- Árbol de decisiones
  Representa un conjunto de reglas de clasificación en forma de un árbol que, a partir de los atributos de cada clase alcanzan un punto final de una ruta.
- Características cognitivas
  Capacidad de las personas para poder pensar, razonar y resolver problemas basados en conocimientos y experiencias anteriores.
- Características afectivas
  Relacionadas con el aspecto emocional del estudiante como la motivación, autocontrol, perseverancia, etc.
- Clasificación
  Forma de análisis de datos que permiten extraer modelos que describen las clases importantes de los datos.
- Cross-Industry Standard Process for Data Mining (CRISP-DM)
  Es una metodología y un modelo de proceso que define un ciclo de seis tareas principales para establecer un marco de trabajo del ciclo de vida de desarrollo en minería de datos.
- Minería de datos
  Sistema de información basado en computación que explora grandes repositorios de datos para generar información y descubrir conocimiento.

- Minería de datos para educación
  Campo de aplicación de Minería de datos orientado hacia el sector de educación.
- Ingresante
  Estudiante que ha logrado ingresar a la universidad y está cursando el primer ciclo de estudios de su carrera.
- Redes neuronales
  Modelo de clasificación que consiste en múltiples capas de nodos que simulan el funcionamiento del cerebro humano. Son populares por su gran capacidad de predicción. Son modelos de caja negra por lo que su interpretación puede ser dificultosa.
- Rendimiento académico
  El grado de conocimiento reconocido por el sistema educativo que posee un estudiante y es expresado por el promedio de calificaciones asignadas por el profesor.
- Situación Socio-Económica (SSE)
  Es la medida relativa económica y social de una persona o familia. Para este estudio, se utilizó como medida el tipo de colegio y la mensualidad, asumiendo que el estudiante no recibe algún tipo de beca o apoyo externo.
- Support Vector Machines (SVM)
  Algoritmo de clasificación que define hiperplanos basados en los vectores de soporte para crear modelos que permitan establecer la mejor separación entre los puntos por clasificar.

2.4 Hipótesis y variables

2.4.1 Hipótesis

Hipótesis general:

El rendimiento académico puede predecirse mediante minería de datos para estudiantes del primer ciclo de la Escuela Profesional de Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres.

Hipótesis específicas:

H1: El rendimiento académico en estudiantes universitarios de primer ciclo de la Escuela Profesional de Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres puede estimarse mediante indicadores sociales, económicos y académicos.

H2: La significación de los indicadores sociales, económicos y académicos en estudiantes universitarios de primer ciclo de la Escuela Profesional de Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres puede determinarse mediante el análisis de componentes principales.

H3: La minería de datos educacional puede aplicarse para indicadores sociales, económicos y académicos en estudiantes universitarios de primer ciclo de la Escuela Profesional de Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres.

2.4.2 Variables

Variable independiente

Es aquella que sirve como entrada o predictor del modelo para la predicción de la variable dependiente.

- Plan de estudios

  Para el periodo seleccionado, existen dos planes de estudios del 2010 y 2014

- Cursos

  Los cursos llevados durante el primer ciclo en la universidad. Según el plan de estudio, existen algunas materias que son diferentes.

- Puntaje de examen de admisión

El puntaje obtenido en el examen de admisión. Dependiendo de la modalidad, el puntaje máximo varía, por lo que se optó por expresarlo en porcentajes.

- Modalidad de ingreso

  La modalidad por la que logró ingresar a la universidad (Ordinario, primera alternativa, convenio excelencia, centro pre universitario, tercio superior, primeros puestos y traslado).

- Colegio de procedencia

  El nombre del colegio en donde ha realizado sus estudios el estudiante en el nivel secundario.

- Tipo de colegio

  El tipo de colegio en donde realizó sus estudios secundarios (estatal, privado, religioso, militar, etc.)

- Sexo

  Es su género, masculino o femenino, según lo registrado en el momento de inscribirse en el examen de admisión.

- Edad

  La edad del ingresante al momento de haber iniciado las clases en el primer ciclo

- Distancia

  Es la distancia en kilómetros calculada desde la dirección de domicilio hasta el lugar en donde se realizan las clases (Facultad de Ingeniería y Arquitectura – Universidad de San Martín de Porres).

- Escala de pensión

La escala de pensión asignada por la Oficina de Categorización de la Universidad.

- Provincia

  Para indicar si el estudiante viene de provincia

Variable dependiente

Es el valor que queremos predecir utilizando las variables independientes. En este caso, solamente existe una variable dependiente:

- Rendimiento académico, expresado como condición de aprobado, promedio calculado de las notas obtenidas en el primer ciclo de estudios.

2.4.3 Matriz de Consistencia

PROBLEMAS	OBJETIVOS	HIPÓTESIS	VARIABLES	INDICADORES
Problema general: ¿Cómo puede predecirse el rendimiento académico en estudiantes de primer ciclo mediante una técnica analítica estadística y computacional? **Problemas específicos:** ¿Cuáles indicadores de impactos pueden medir el rendimiento académico en estudiantes universitarios de primer ciclo? ¿Qué estadígrafo puede aplicarse para medir el rendimiento académico en estudiantes universitarios de primer ciclo? ¿Cuál técnica estadística y computacional puede medir el rendimiento académico en estudiantes universitarios de primer ciclo?	**Objetivo general:** Predecir el rendimiento académico mediante minería de datos en estudiantes del primer ciclo de la Escuela Profesional de Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres **Objetivos específicos:** • Estimar indicadores sociales, económicos y académicos para el rendimiento académico en estudiantes universitarios de primer ciclo de la Escuela Profesional de Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres • Determinar mediante análisis de componentes principales la significación de los indicadores sociales, económicos y académicos en estudiantes	**Hipótesis general:** El rendimiento académico puede predecirse mediante minería de datos para estudiantes del primer ciclo de la Escuela Profesional de Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres. **Hipótesis específicas:** H1: El rendimiento académico en estudiantes universitarios de primer ciclo de la Escuela Profesional de Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres puede estimarse mediante indicadores sociales, económicos y académicos. H2: La significación de los indicadores sociales, económicos y académicos en estudiantes universitarios de primer ciclo de la Escuela Profesional de	**Independiente:** Indicadores sociales Indicadores económicos Indicadores académicos **Dependiente:** Rendimiento Académico	• Sexo • Edad • Distancia • Provincia • Tipo de colegio • Escala de pensión • Puntaje de examen de admisión • Plan de Estudios • Cursos • Modalidad de ingreso • Colegio de procedencia • Condición de aprobado

	universitarios de primer ciclo de la Escuela Profesional de Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres • Aplicar la minería de datos educacional para indicadores sociales, económicos y académicos en estudiantes universitarios de primer ciclo de la Escuela Profesional de Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres	Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres puede determinarse mediante el análisis de componentes principales. H3: La minería de datos educacionales puede aplicarse para indicadores sociales, económicos y académicos en estudiantes universitarios de primer ciclo de la Escuela Profesional de Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres.		

CAPÍTULO III
METODOLOGÍA

3.1 Tipo de investigación

3.1.1 Tipo de investigación o procedimiento

El enfoque de investigación fue cuantitativo. Se gestionó la búsqueda de información científica con relación a las variables para demostrar la hipótesis, luego se operacionalizaron las mismas y finalmente, se estableció la relación teórica entre las variables de selección (Sampieri, Collado y Lucio, 2014).

El tipo de investigación, según el estado actual del conocimiento con enfoque cuantitativo, fue explicativo y correlacional.

3.1.2 Métodos de la investigación

Los métodos cuantitativos de la investigación correspondieron a:

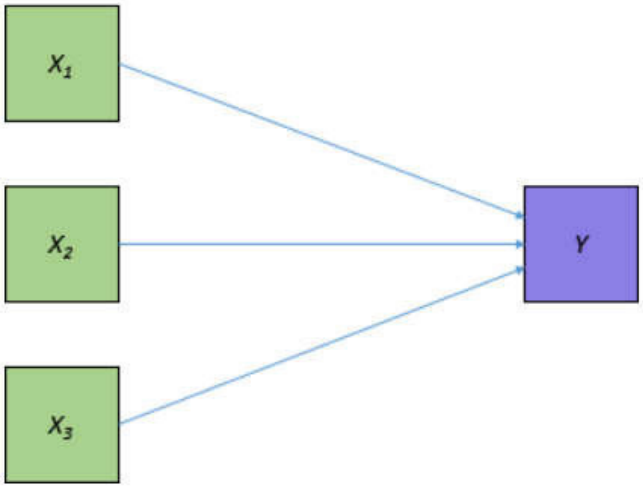
- Teórico-analítico-sintético (caracterización de niveles con relación a indicadores sociales, económicos y académicos)

- Teórico lógico-histórico (análisis de la evolución histórica sobre la no continuidad en los estudiantes universitarios del primer ciclo académico)
- Teóricos por modelación (descomposición operacional sobre asignación de cualidades con relación a indicadores sociales, económicos y académicos)
- Empírico por medición (comprobación de indicadores sociales, económicos y académicos).

3.2 Diseño de la investigación

El diseño de la investigación es transeccional del tipo correlacional-causal. Se busca describir la relación que existe entre el rendimiento académico y los factores social, económico y académico de los ingresantes que se utilizaron para probar la validez de la predicción del rendimiento académico.

El diseño de la investigación es el siguiente:



En donde:

X1: Variable indicadores sociales
X2: Variable indicadores económicos
X3: Variables indicadores académicos
Y: Rendimiento académico de los ingresantes

3.3 Población y muestra

3.3.1 Población

La población del presente estudio son los estudiantes ingresantes a la carrera de Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres.

3.3.2 Muestra

Se utilizó una muestra probabilística en donde se realizaron estimaciones de variables en la población. Definiendo un margen de error menor al 25%, con un intervalo de confianza de 95% y dado el análisis previo en donde existe un 30% de los estudiantes que logran aprobar, satisfactoriamente, el primer ciclo de estudio tenemos:

$$ME = z\sqrt{\frac{p(1 \quad p)}{n}}$$

En donde:

- ME es el margen de error deseado.
- *z* es el z-score, que para un intervalo de confianza de 95% es 1.96.
- *p* es el valor esperado de la variable por analizar.
- *n* es el tamaño de muestra (por calcular).

$$0.025 = 1.96\sqrt{\frac{0.3 \quad 0.7}{n}} \qquad n = \frac{0.3 \quad 0.7}{0.0001627} \qquad n = 1291$$

Por lo que se va a requerir una muestra de aproximadamente 1300 estudiantes. Realizando un corte en los semestres académicos disponibles en la data se ha seleccionado como muestra para la siguiente investigación los estudiantes ingresantes de los periodos 2010-I a 2015-II a la carrera de Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres, que hacen un total de 1304 ingresantes admitidos.

3.4 Técnicas de recolección de datos

3.4.1 Descripción de los instrumentos

Se utilizó información histórica extraída de las bases de datos de la Oficina de Admisión y de la Facultad de Ingeniería y Arquitectura de la Universidad de San Martín de Porres relacionado con los estudiantes y su rendimiento académico tales como: promedio general, modalidad de ingreso, distrito de procedencia, situación familiar y socio-económica, entre otros.

Se aplicaron técnicas de extracción, transformación y carga de datos para convertir la información extraída de una base de datos transaccional hacia un formato apropiado para la aplicación de las técnicas de minería de datos.

3.5 Técnicas para el procesamiento de la información

3.5.1 Presentación, análisis e interpretación de los datos (tratamiento estadístico)

Para la presentación y análisis de los datos, se utilizaron las herramientas Microsoft Excel y R Studio. Se elaboraron tablas e histogramas para analizar la relación entre la variable dependiente rendimiento académico y las variables independientes.

3.5.2 Prueba de hipótesis

- Estimar indicadores sociales, económicos y académicos para el rendimiento académico en estudiantes universitarios de primer ciclo
- Determinar mediante el análisis de la varianza, componentes principales y técnicas de regresión de la significación de los indicadores sociales, económicos y académicos en estudiantes universitarios de primer ciclo
- Aplicar la minería de datos educacional para indicadores sociales, económicos y académicos mediante el análisis de la varianza, componentes principales y regresión múltiple en estudiantes universitarios de primer ciclo

3.6 Aspectos éticos

Los principales aspectos éticos relacionados con este tipo de investigación son las de privacidad respecto a la información personal tanto de los alumnos como docentes. El uso de información que permita identificar, individualmente, a las personas requiere autorización y consentimiento previo al uso y publicación de la información.

Para asegurar la privacidad de los datos se realizó un proceso para anonimizar la información, asignando un identificador propio para cada registro en los datos y eliminar los atributos que permitieron identificar a las personas de los diferentes registros como nombres, apellidos, dirección de domicilio, entre otros. De esta manera, se aseguron la validez de los resultados obtenidos en las pruebas sin exponer la información personal de los ingresantes.

CAPÍTULO IV
DESARROLLO DEL PROYECTO

4.1 Captura de información

4.1.1 Recolección de data inicial

Se recolectó información de las bases de datos de la Oficina de Admisión y de la Facultad de Ingeniería y Arquitectura de la Universidad de San Martín de Porres.

De la Oficina de Admisión se extrajeron los siguientes datos:

- Semestre de ingreso
- Código del alumno
- Apellido paterno
- Apellido materno
- Nombres
- Fecha de nacimiento
- Sexo
- Modalidad de ingreso
- Nombre del colegio de procedencia

- Departamento, provincia, distrito del colegio
- Tipo de colegio
- Puntaje obtenido en el examen de Admisión

De la Facultad de Ingeniería y Arquitectura:

- Código del alumno
- Apellido paterno
- Apellido materno
- Nombres
- Semestre de ingreso
- Estado de matrícula
- Escala de pensión
- Dirección de domicilio
- Cursos llevados
- Sección
- Nota

4.1.2 Exploración y verificación de la data inicial

La exploración inicial se efectuó tanto para los datos de la Oficina de Admisión y de la Facultad de Ingeniería y Arquitectura. Se ha logrado contrastar utilizando el código de alumno y los datos personales de la información existente de los ingresantes y de sus notas en la Facultad.

Se halló, en la información de la que han sido extraídos, de la base de datos, hacia las hojas de cálculo que, en casos de información tipo texto, los programas tipo Microsoft Excel ignoran los espacios en blanco al momento de mostrar la información, lo que no ocurre en las bases de datos y en programas de análisis como R, por lo que fue necesario, en la siguiente tarea, realizar la eliminación de los espacios y reemplazar los caracteres que se consideren erróneos.

Asimismo, debido a los cambios en los planes de estudio realizado durante el periodo del que se extrajo la data, se dieron tres variantes

en los cursos llevados durante el primer ciclo de estudios. El primer plan fue utilizado entre los años 2010 y 2013. En el año 2014, se realizaron cambios, se incorporaron cursos y otros se retiraron. En el año 2015, debido a los cambios normados por la Ley Universitaria N° 30220, el curso de inglés fue obligatorio desde el primer ciclo.

Debido a que, en el último cambio del Plan de estudios del año 2017, los cursos del primer ciclo son las mismas en el plan utilizado durante el periodo 2010-2013, con excepción del curso de inglés. Se priorizó el análisis de los cursos existentes en dicho plan de estudios, sin eliminar características o patrones que se han podido encontrar en los cursos que pasaron a ciclos superiores.

4.2 Preparación de la data

4.2.1 Selección de data

Una vez realizada la verificación de la data, en la fase anterior, se integró la información en un solo conjunto de datos utilizando el código de alumno. Se descarta información incompleta, faltante o incongruente para evitar que se analicen datos que puedan crear errores al momento de ejecutar los algoritmos de minería de datos.

Se generó un ID único para cada fila, en la tabla, a fin de diferenciar los datos y también eliminar la dependencia de otros identificadores que pudieran existir en la data, en este caso, el código del alumno.

4.2.2 Limpieza de datos

En esta etapa, se aplicaron las funciones de EXCEL: TRIM(), CLEAN(), SUBSTITUTE() para eliminar los espacios en blanco existentes en los datos de tipo texto y se aplicaron las funciones de redondeo para los valores numéricos con decimales.

Se eliminaron las informaciones personales que pudieran identificar, individualmente, a cada alumno como código del alumno, apellido paterno, apellido materno, nombre y dirección.

En una primera revisión de la data, durante esta fase, se encontró que existen algunos alumnos que sí estaban matriculados, pero aparentemente nunca asistieron a clases, ya que en todos los cursos figuraba la nota cero. La data de estos estudiantes fueron eliminados del conjunto de datos a fin de evitar posibles errores.

4.2.3 Construcción de la data

Esta fase se inició con la construcción de nuevas características para enriquecer la data con la finalidad de adaptarla a que sea apta para la aplicación de minería de datos. Entre las características agregadas se hallan:

- EDAD. Es la edad considerada al momento de haber iniciado las clases, que se calcula restando de la fecha del primer día de clases y la fecha de nacimiento.

  Por ejemplo, de un alumno ingresante del ciclo 2010-I con fecha de nacimiento 01 de enero de 1992 se calcula:

  Primer día de clases – fecha de nacimiento = tiempo en años (redondeado hacia abajo)

  01/03/2010 – 01/01/1992 = 18.1853425 -> 18 años

- SEXO. Para facilitar la ejecución de algoritmos se requiere que todos los datos sean numéricos se transformó a binario asignado 1 para masculino y 0 para femenino
- PROVINCIA. Para identificar si el estudiante ha asistido al colegio en provincia o en Lima, se utiliza un formato binario
- TIPOCOLEGIO. Se asignan valores numéricos según el tipo de colegio.
- COLEEXC. A base del listado de Colegio de Excelencia Académica, de los 500 colegios que han tenido mejor rendimiento en otra universidad, se indicó si el colegio de procedencia de los alumnos pertenece a dicho listado.

- COLPRE- Revisión inicial de la data se encontró que hay una cantidad considerable de alumnos que proceden de colegios pre-universitarios (Trilce, Saco Oliveros, etc.), se agregó como una de las variables.
- EXADMISIÓN. Se aplica una fórmula de normalización para convertir en valores entre 1 y 0. Para examen de admisión ordinario que tiene como puntaje máximo 2000 puntos, se dividió el puntaje de cada alumno entre 2000.
- DISTANCIA. Utilizando la API de Google Maps (Figura 6) se calculó la distancia entre la dirección de domicilio y la dirección de la Facultad de Ingeniería y Arquitectura de la Universidad de San Martín de Porres en donde se dictaron las clases (Av. La Fontana 1250 Urb. Santa Patricia 2da. etapa, La Molina, Lima, Perú). Para casos en donde no está registrada la dirección exacta se utilizó una distancia promedio al distrito indicado.
- APROBADO. La condición de aprobado para los cursos expresado en binario, 1 para aprobado (nota >= 11) y 0 para desaprobado (nota <11). Se crearon variantes para cada curso que llevaron durante el primer ciclo

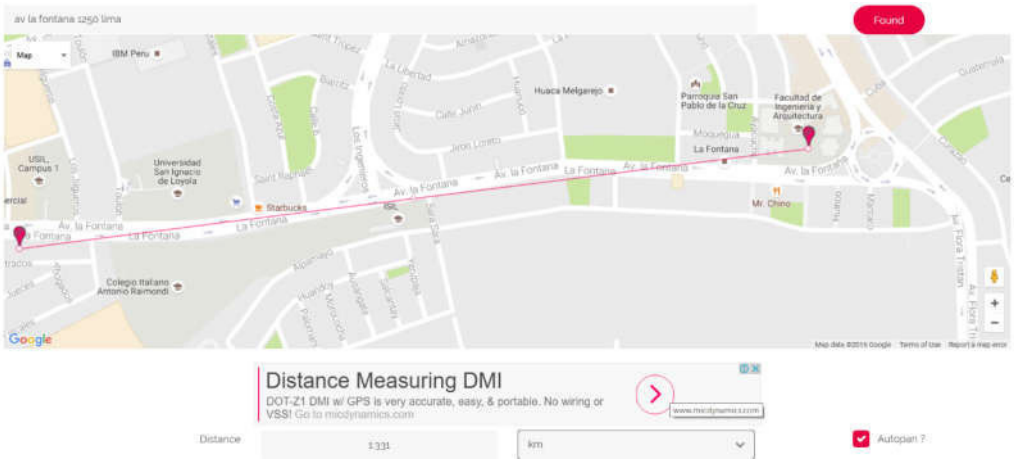


Figura 6. Aplicación para calcular distancias con Google Maps

Elaboración: *el autor*

4.3 Minería de datos

4.3.1 Análisis exploratorio

En una primera iteración para detectar posibles errores en la data, se realizó el estudio con los ingresantes de un semestre, 2010-1. Visualizando gráficamente la data, se encontró una cantidad elevada de alumnos procedentes de colegios estatales (*Figura 7*).

Haciendo una revisión del conjunto de datos se descubrió que muchos de los colegios estaban mal clasificados como ESTATAL. Se efectuó una revisión y corrección en el resto de la data (Figura 8) y se observó como se esperaba una predominancia de ingresantes que proceden de colegios privados, resaltando también los colegios pre-universitarios y en una medida menor, pero igualmente significativa los colegios estatales.

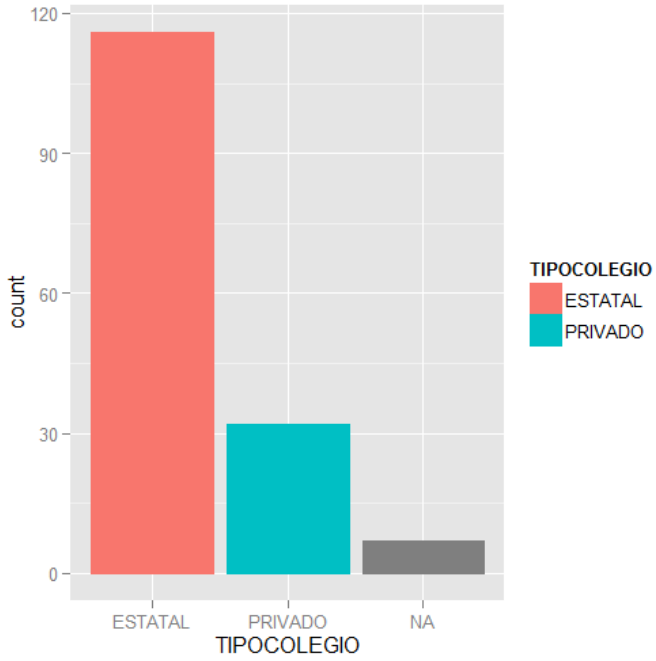


Figura 7. *Ingresantes 2010-1 agrupados por tipo de colegio*

Elaboración: *el autor*

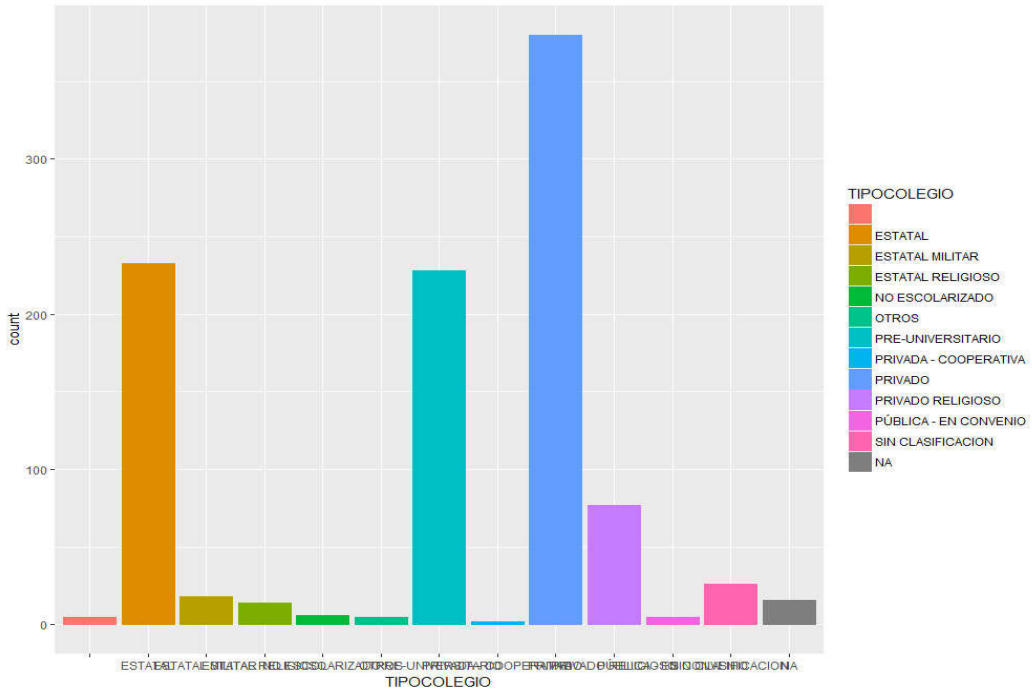


Figura 8. *Ingresantes por tipo de colegio versión final*

Elaboración: *el autor*

Con respecto a la variable APROBADO, es decir, si ha logrado pasar de manera invicta sin desaprobar ningún curso en su primer ciclo, se ha encontrado que solamente 22.6% del total de ingresantes cumplió con los retos propuestos por la universidad. Esta cifra se torna más grave si consideramos que el Plan de estudios del 2010, es muy similar al actual, pero tiene un menor porcentaje (19.86%) de invictos que en el 2014 (30%).

Visualizando, gráficamente la data, para ver si es que existe alguna relación entre la variable APROBADO y las demás, no se ha podido encontrar una relación resaltante con variables como edad, sexo, tipo de colegio, si viene de provincia, nota de examen de admisión, etc.; solamente con la variable MODALIDAD (*Figura 9*), más específicamente para los que han ingresado por la modalidad de TERCIO SUPERIOR y EXONERADOS PRIMER O SEGUNDO PUESTO, quienes en su mayoría han logrado aprobar todos los cursos.

Analizando relaciones entre los cursos se encontró que los cursos de Matemática Discreta y Geometría Analítica son los que mayor

dificultad generó a los alumnos. Como se observa en la *Figura 10,* casi el 90% de los alumnos que aprobaron dichos cursos también lograron superar, satisfactoriamente, el primer ciclo.

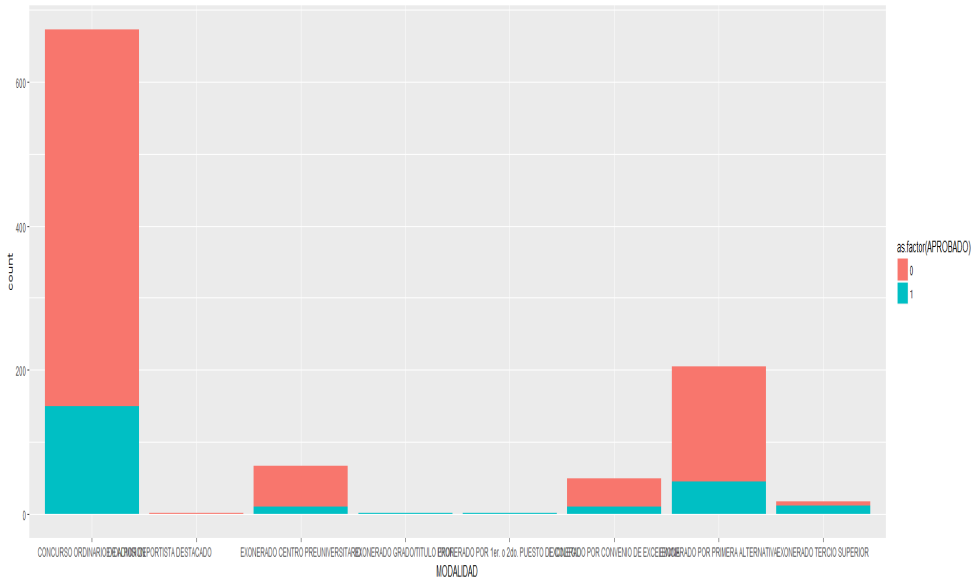


Figura 9. Ingresantes que aprobaron el primer ciclo por modalidad

Elaboración: *el autor*

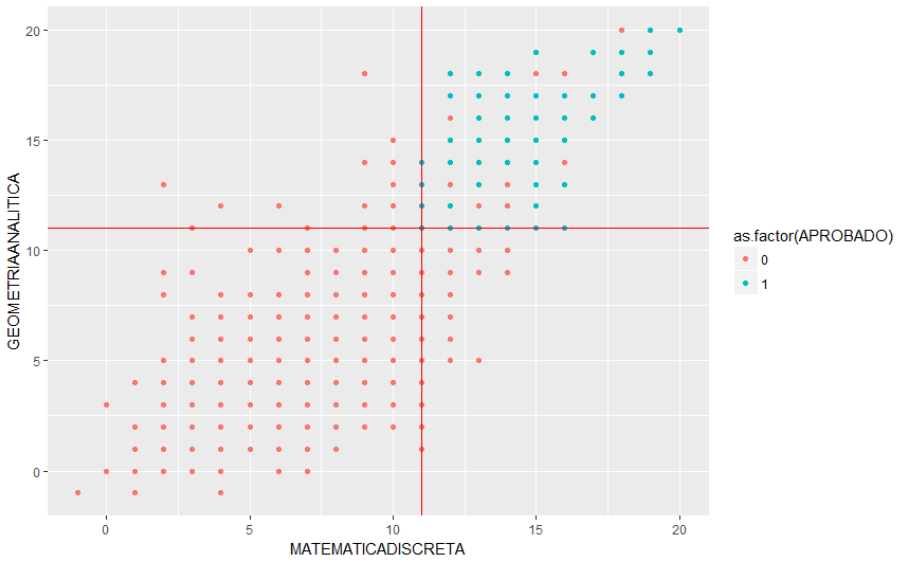


Figura 10. Condición de aprobado agrupado según nota obtenida en los cursos de Geometría Analítica y Matemática Discreta

Elaboración: *el autor*

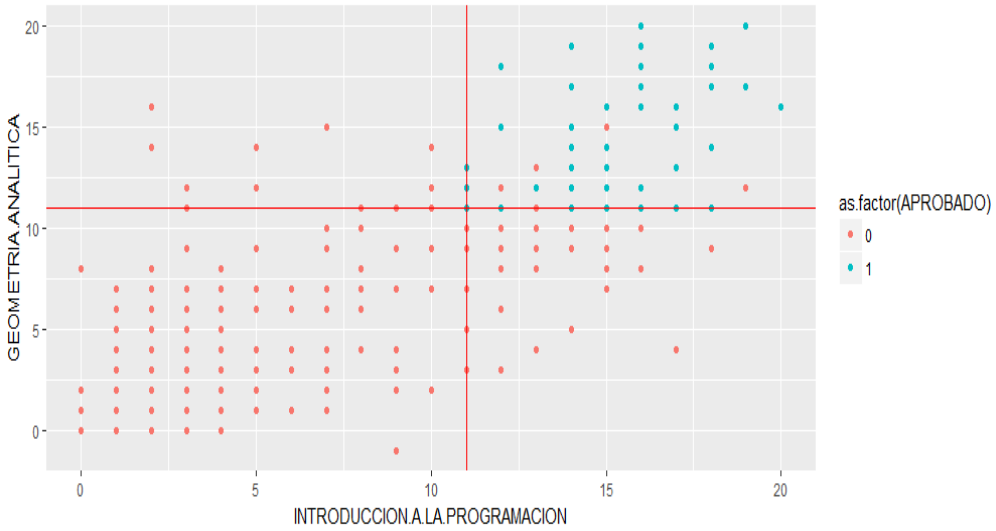


Figura 11. Condición de aprobado agrupado según nota obtenida en los cursos de Geometría Analítica e Introducción a la Programación

Elaboración: *el autor*

En el caso del Plan de estudios de 2014-2016 (*Figura 11*), se observa un fenómeno similar, en este caso con el curso de Introducción a la Programación que reemplaza a Matemática Discreta. En este caso, también existe una muy alta posibilidad de que si tienen una nota igual o mayor a 11, en ambos cursos, terminan aprobando el resto de ellos.

Analizando los cursos de Geometría Analítica y Matemática Discreta en comparación con el resto de las asignaturas tienen, en promedio, notas inferiores en 3 o 4 puntos: 6.829 para Geometría Analítica y 8.068 para Matemática Discreta, mientras que el resto tiene promedio de 11 o 12. Para el Plan de estudios de 2014-2016, también se observa un fenómeno similar, con promedio de 7.929 para Geometría Analítica y 8.739 en Introducción a la Programación repitiéndose para el resto de los cursos, promedios alrededor de 11 o 12.

Se puede concluir entonces que uno de los puntos clave para mejorar el rendimiento académico de los alumnos es realizar cambios y mejoras en los cursos de Geometría Analítica y Matemática Discreta. También sería importante evaluar el curso de Introducción a la Programación que genera dificultades para los estudiantes.

La tabla 4 muestra el número de alumnos matriculados de forma comparada con el número de alumnos aprobados en los semestres correspondientes a un periodo de cinco años. El número de matriculados correspondió a 876, mientras que los que concluyeron satisfactoriamente fueron 202 (23%), lo cual indica que aproximadamente 3/4 de la población matriculada, no culmina el semestre académico.

Tabla 4
Categorías por indicadores sociales, económicos y académicos

Identificación	Semestre					
	2010	2011	2012	2013	2014	2015
	I	II	III	IV	V	VI
Alumnos matriculados	214	196	174	141	150	125
Alumnos Aprobados	42	37	27	42	55	32

Elaboración: *el autor*

Las tablas 5 y 6 muestran el ANOVA realizado con relación al número de alumnos matriculados y aprobados en los semestres analizados y dado que el valor-P de la prueba-F fue menor que 0.05, existió una diferencia, estadísticamente significativa, entre los valores promedios con un nivel del 95.0% de confianza.

Tabla 5
Análisis de la varianza con relación a los alumnos matriculados / semestre

Fuente	Suma de Cuadrados	Grados de libertad	Cuadrado Medio	Coeficiente - F	Valor - P
Entre grupos	11468.4	4	2867.1	2867.10	0.0000
Intra grupos	10.0	10	1.0		
Total (Corr.)	11478.4	14			

Elaboración: *el autor*

Tabla 6
Análisis de la varianza con relación a los alumnos aprobados / semestre

Fuente	Suma de Cuadrados	Grados de libertad	Cuadrado Medio	Coeficiente - F	Valor - P
Entre grupos	1221.6	4	305.4	305.40	0.0000
Intra grupos	10.0	10	1.0		
Total (Corr.)	1231.6	14			

Elaboración: *el autor*

Las tablas 7 y 8 muestran la prueba de contraste múltiple de rangos para conocer las diferencias entre los valores promedios de los alumnos matriculados en los semestres seleccionados. El método empleado actualmente para discriminar entre las medias fue el procedimiento de diferencia mínima significativa (LSD) de Fisher. Con este método hay un riesgo del 5.0% al decir que cada par de medias es, significativamente diferente, cuando la diferencia real es igual a 0.

Se observó que, el número de alumnos matriculados, por lo general, disminuyó de un semestre a otro, correspondiéndose este comportamiento durante los tres primeros semestres.

Tabla 7

Prueba de contraste múltiple de rangos / alumnos matriculados

Semestre	Media	Grupos Homogéneos
AM 2015	125.0	X
AM 2013	141.0	X
AM 2014	150.0	X
AM 2012	174.0	X
AM 2011	196.0	X
AM 2010	215.0	X

Elaboración: *el autor*

Tabla 8
Prueba de contraste múltiple de rangos / alumnos culminados

Semestre	Media	Grupos Homogéneos
AC 2012	27.0	X
AC 2015	32.0	X
AC 2011	37.0	X
AC 2010	41.0	X
AC 2013	42.0	X
AC 2014	55.0	X

Elaboración: *el autor*

La tabla 9 muestra el número de estudiantes que no lograron aprobar todos los cursos según su rendimiento de acuerdo con la categoría analítica. Al realizarse una comparación entre las categorías base (social, económico y académicos) se encontró que el carácter social tuvo mayor número de estudiantes como razón de no culminación, siendo de 116, mientras que lo económico y académico presentaron 114 estudiantes, respectivamente. Existieron diferencias estadísticamente significativas ($p \leq 0,05$), entre el número de estudiantes por categoría (tabla 10 y 11).

Tabla 9
Prueba de contraste múltiple de rangos / alumnos culminados

Categoría	**Rendimiento**		
	bajo	medio	alto
S	43	63	10
E	3	20	91
A	37	70	7
S-E	2	42	77
E-A	3	33	119
S-E-A	3	33	18

Elaboración: *el autor*

Tabla 10
Análisis de la varianza con relación a los alumnos no culminados / categoría base

Fuente	Suma de Cuadrados	Grados de libertad	Cuadrado Medio	Coeficiente - F	Valor - P
Entre grupos	8.0	2	4.0	4.00	0.0787
Intra grupos	6.0	6	1.0		
Total (Corr.)	14.0	8			

Elaboración: *el autor*

Tabla 11
Prueba de contraste múltiple de rangos en alumnos no culminados / categoría base

Categoría	Media	Grupos Homogéneos
E	114,0	X
A	114,0	X
S	116,0	X

Elaboración: *el autor*

Al compararse las interacciones entre las categorías, se observó que lo económico y académico representó el mayor número de estudiantes seguidamente de lo social con lo económico, aunque lo social-económico y académico representó más del 50% de las interacciones entre las categorías anteriores, correspondiendo a 185, 121 y 84, respectivamente (*tabla 12*). Según los resultados de la tabla 11, tanto lo económico como lo académico al no tener diferencias significativas entre ellas, resultaron con un valor predominante. La categoría social no influye en las categorías económicas y académicas, pues se les permite a los estudiantes un posible atraso en la regulación de los pagos por derecho de enseñanza.

Tabla 12
Número de alumnos con cursos desaprobados según interacción entre categorías

Interacción entre categoría	Alumnos desaprobados
S-E	121
E-A	185
S-E-A	84

Elaboración: *el autor*

Tabla 13
Número de alumnos no culminados según clasificación por rendimiento

Rendimiento	Alumnos no culminados
bajo	91
medio	261
alto	322

Elaboración: *el autor*

Tabla 14
Análisis de componentes principales según el peso de cada categoría base y por interacción

Componente	Eigenvalor	Porcentaje de	
Número		Varianza	Acumulado
R B S	13.5826	75.459	75.459
R M S	4.41742	24.541	**100.000**
R A S	1.17266E-15	0.000	100.000
R B E	5.69011E-16	0.000	100.000
R M E	3.20812E-16	0.000	100.000
R A E	2.31751E-16	0.000	100.000
R B PL	2.04724E-16	0.000	100.000
R M PL	1.60915E-16	0.000	100.000
R A PL	1.48804E-16	0.000	100.000
R B SE	6.45904E-17	0.000	100.000
S M SE	4.26175E-17	0.000	100.000
R A SE	0.0	0.000	100.000
R B EP	0.0	0.000	100.000
R M EP	0.0	0.000	100.000
R A EP	0.0	0.000	100.000
R B SEPL	0.0	0.000	100.000
R M SEPL	0.0	0.000	100.000
R A SEPL	0.0	0.000	100.000

Elaboración: *el autor*

La tabla 13 muestra el número de estudiantes que no finalizaron, satisfactoriamente sus estudios, según la clasificación de rendimiento. Puede valorarse que este número fue incrementándose desde el nivel bajo, medio hasta el nivel alto. Este comportamiento era el esperado, por lo cual, hubo un efecto acumulativo influyente o determinante por la categoría referida a lo social según el análisis de componentes principales (tabla 14).

Luego se aplicaron distintas técnicas de minería de datos para descubrir nuevos conocimientos de la data.

4.3.2 Técnicas de regresión lineal

Las técnicas de regresión lineal buscan encajar hacia un modelo lineal en donde se establece la relación entre la variable dependiente (y ó f(x)) y la variable independiente (x), asumiendo que la relación entre las variables sigue una línea recta. En el caso específico de esta investigación, se utilizó la regresión logística para crear un modelo para la condición de APROBADO (dado que es una variable cualitativa) y la regresión lineal múltiple para el caso del modelo para las notas obtenidas en cada curso.

Para la selección de las variables que tienen mayor influencia en la variable dependiente, se utilizó el método de regresión paso a paso del tipo selección hacia atrás (Backward selection) que se inicia con un modelo general que incluye todas las variables que se irán eliminando del modelo uno a uno hasta cumplir con el criterio especificado, en este caso las variables que tenga un valor de *p* menor a 0.05, valor normalmente utilizado para pruebas de hipótesis.

Tabla 15
Resultados de regresión logística ingresantes Plan 2010

Coeficientes:					
	Estimado	Error Estándar	Valor Z	Pr(>|z|)	Significancia
(Intercept)	-3.37717	0.87575	-3.856	0.000115	***
EXADMISION	2.70438	0.47570	5.685	1.31e-08	***
EDAD	0.10228	0.04379	2.335	0.019525	*
SEXO	-0.55567	0.24382	-2.279	0.022663	*
DISTANCIA	-0.03678	0.01665	-2.209	0.027194	*

Elaboración: *el autor*

Para el caso del modelo de regresión logística como se muestra en la tabla 15, se observa que las variables que tienen mayor relación con la condición de aprobado son: Nota de examen de admisión, edad, sexo y la distancia. Como era de esperarse la nota de examen de admisión es la que tiene mayor influencia y también la distancia dado que tiene un coeficiente negativo, lo que significa que a menor distancia existe mayor probabilidad de que el ingresante apruebe los cursos. Se observa, adicionalmente, que la variable sexo tiene un coeficiente negativo con un valor elevado, es decir, el sexo femenino tiene mayor posibilidad de aprobar el primer ciclo.

No se ha logrado demostrar que las variables relacionadas con el colegio de procedencia tengan alguna influencia sobre la condición de aprobado de los ingresantes. La única que estuvo cerca es la variable colegio de excelencia, que tuvo un valor de *p* de 0.083, la cual está cerca al 0.05 necesario para demostrar su validez estadístico, pero para este modelo no se consideró incluirlo.

Para realizar la predicción utilizando este modelo, se dividió la data en 75% para entrenar el modelo y 25% como datos de prueba. La matriz de confusión obtenida con los datos de prueba utilizando un punto de corte de 0.2 se muestra en la *tabla 16.*

Tabla 16
Resultado de regresión logística Plan 2010 con punto de corte de 0.2 para el grupo de prueba

	FALSO	VERDADERO
0	100	44
1	15	22

Elaboración: *el autor*

La predicción realizada por el modelo de regresión logística ha obtenido una exactitud de 67.4%, con una sensibilidad de 69.44%, una especificidad de 59.45% y el área bajo la curva ROC de 68.82%, como se muestra en la *Tabla 17*, la curva ROC se muestra en la *Figura 12*.

Realizando pruebas de validación sobre los resultados se ha encontrado que existe mucha variación en los resultados según la selección aleatoria que se realizó al momento de dividir la data. Analizando más a profundidad se ha encontrado que como se había observado en la Figura 13 la cantidad de aprobados varía considerablemente según la modalidad de ingreso que tuvieron.

Tabla 17
Resultados de predicción Regresión logística Plan 2010

Exactitud	Sensibilidad	Especificidad	AUC
67.4%	69.44%	59.45%	68.82%

Elaboración: *el autor*

Considerando el punto anterior, se volvieron a realizar las pruebas para crear el modelo y ejecutar la predicción separando la data según la modalidad de ingreso y se lograron resultados con mejoras importantes, especialmente con los ingresantes por examen ordinario de admisión.

El mejor modelo encontrado con solo la data de ingresantes por examen ordinario de admisión se encuentra en la Tabla 18. A diferencia del modelo anterior, las variables edad y distancia no cumplen con el valor esperado de *p* y ha sido reemplazado por la variable colegio de excelencia.

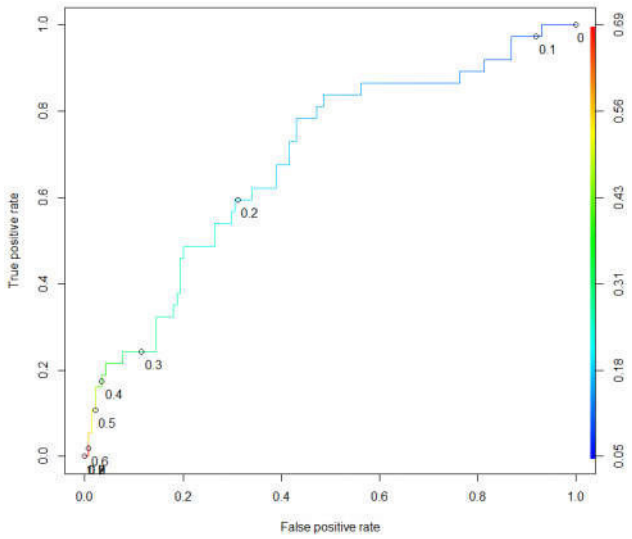


Figura 12: Curva ROC del modelo de regresión logística del Plan 2010

Elaboración: *el autor*

Los resultados obtenidos de la predicción se muestran en la Tabla 19, en donde se observan mejoras cercanas a 10%, obteniéndose también una curva ROC con un AUC que puede representar el 82.12% de los datos presentados para realizar la predicción en el modelo.

Para el caso del Plan de estudios 2014 – 2016 (Tabla 20), la variable que cumple con el punto de corte para ser considerado como importante es solo la nota de examen de admisión. Sin embargo, aparecen variables como colegio de excelencia y la escala como variables que podrían tener algún tipo de relación con la condición de aprobado.

Tabla 18
Resultados regresión logística modalidad ordinario Plan 2010

Coeficientes:					
	Estimado	Error Estándar	Valor Z	Pr(>|z|)	Significancia
(Intercept)	-4.2345	0.5179	-8.176	2.96e-16	***
EXADMISION	10.9480	1.3225	8.278	< 2e-16	***
SEXO	-0.8137	0.3398	-2.395	0.0166	*
COLEEXC	-1.1600	0.5839	-1.987	0.0470	*

Elaboración: *el autor*

Tabla 19
Resultados de predicción Regresión logística modalidad ordinario Plan 2010

Exactitud	Sensibilidad	Especificidad	AUC
74.4%	74%	76%	82.12%

Elaboración: *el autor*

Tabla 20
Resultados regresión logística Plan 2014

Coeficientes:					
	Estimado	Error Estándar	Valor Z	Pr(>|z|)	Significancia
(Intercept)	-6.4831	2.3331	-2.779	0.00546	**
EXADMISION	4.6074	1.2429	3.707	0.00021	***
COLEEXC	-1.0447	0.5736	-1.821	0.06859	.
ESCALA	0.2887	0.1585	1.822	0.06845	.

Elaboración: *el autor*

Al realizar la predicción bajo este modelo se obtuvieron valores similares a los del plan de estudios antiguo con una exactitud de 71.01% y un área bajo la curva de 70.69%. Los resultados se muestran en la Tabla 21 y la curva ROC en la Figura 13.

Se ha creado para este grupo un modelo de regresión sobre la data de ingresantes por modalidad de examen ordinario de admisión, utilizando de igual manera 75% de la muestra para el entrenamiento del modelo y 25% para la prueba. Para este caso solamente se ha logrado demostrar que la nota del examen es la que ayuda en el modelo con un valor de *p* de 1.97e-05.

Tabla 21
Resultados de predicción Regresión logística Plan 2014

Exactitud	Sensibilidad	Especificidad	AUC
71.01%	95.74%	18.18%	70.69%

Elaboración: *el autor*

Tabla 22
Resultados de predicción Regresión logística modalidad ordinario Plan 2014

Exactitud	Sensibilidad	Especificidad	AUC
75%	92.59%	38.46%	70.08%

Elaboración: *el autor*

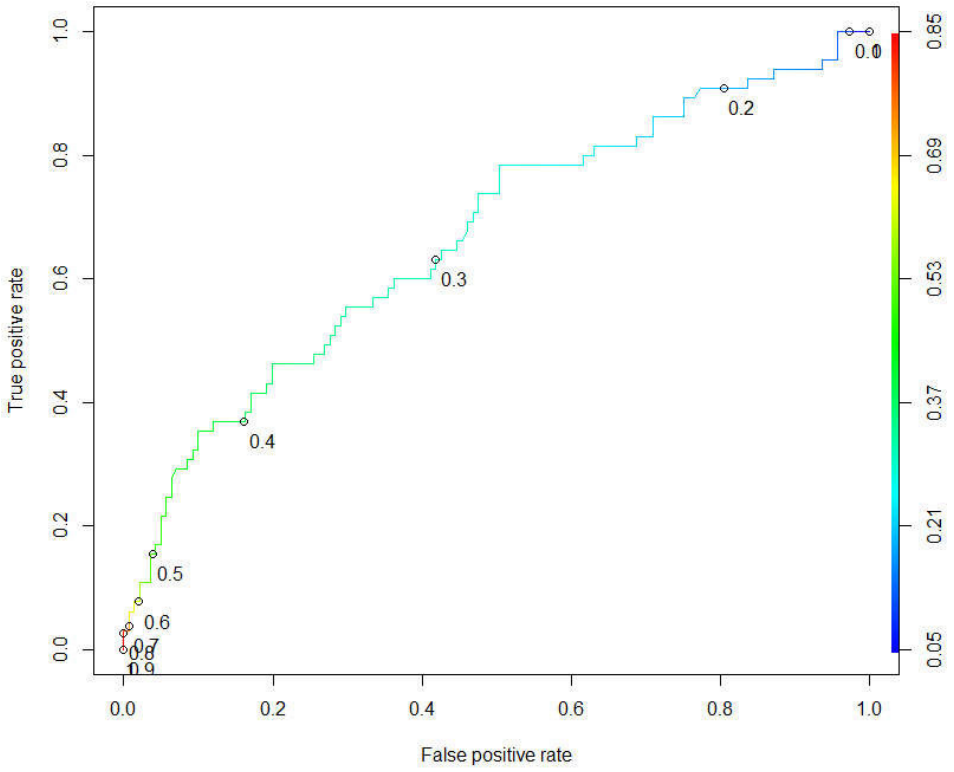


Figura 13. Curva ROC del modelo de regresión logística del Plan 2014

Elaboración: *el autor*

Los resultados de la predicción bajo este modelo se muestran en la Tabla 22 y no muestran cambios como ocurría con la data del Plan de estudios 2010.

4.3.3 Árbol de decisiones

Se utilizó el algoritmo C5.0 para crear modelos basados en la técnica de árbol de decisiones que permitieron establecer un conjunto de reglas para poder realizar la predicción y analizar la importancia de cada variable independiente utilizada. Asimismo, se utilizaron técnicas de *boosting* para realizar de manera automatizada múltiples iteraciones del algoritmo lo

que permitió obtener mejores modelos y también realizar la validación de los resultados obtenidos.

Estableciendo como parámetro para que se realicen 10 iteraciones de la ejecución del algoritmo para la data del periodo 2010 – 2013, el mejor modelo obtenido es el de la Figura 14.

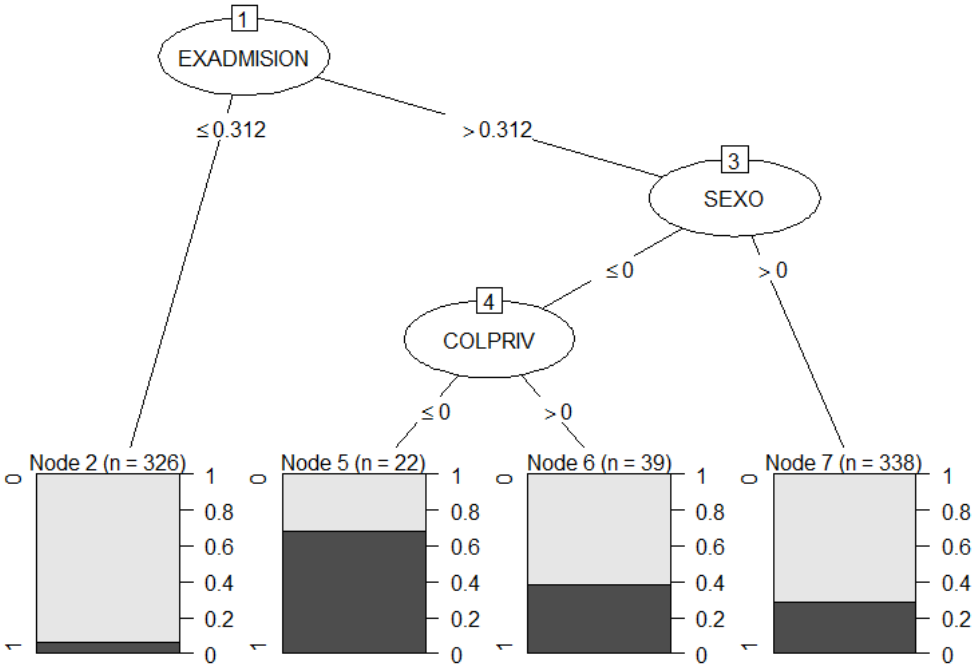


Figura 14. Modelo de árbol de decisiones para Plan 2010

Elaboración: *el autor*

Las variables que mayor uso tuvieron fueron el de examen de admisión, sexo y los relacionados con el colegio de procedencia, en específico si proviene de un colegio privado en el modelo obtenido. Al igual que en los modelos de regresión lineal se observa que hay mayor posibilidad de aprobar todos los cursos si se tiene una nota alta en el examen de admisión y también si pertenece al sexo femenino. El punto adicional que hay que resaltar es que según el modelo hay mayor probabilidad de aprobar si es que no provienen de un colegio privado, es decir, son alumnos de colegios estatales o algún otro lo cual resulta interesante y sería importante realizar un análisis más profundo.

La matriz de confusión obtenida para este modelo (Tabla 23) muestra una exactitud de 81.93%, pero presenta muchos falsos positivos, lo cual hay que tener en consideración cuando se compare este modelo con los demás.

Realizando la predicción a base de este modelo se ha obtenido la siguiente matriz de confusión (Tabla 24) con una exactitud de 82.87%.

Tabla 23
Matriz de confusión para árbol de decisiones con Plan 2010.

(a)	(b)	
565	12	(a): Clase 0
119	29	(b): Clase 1

Elaboración: *el autor*

Tabla 24
Matriz de confusión resultante de la predicción con árbol de decisiones Plan 2010

	PREDICCION		
REAL	0	1	Total
0	133 0.735%	11 0.061%	144
1	20 0.110%	17 0.094%	37
Total	153	28	181

Elaboración: *el autor*

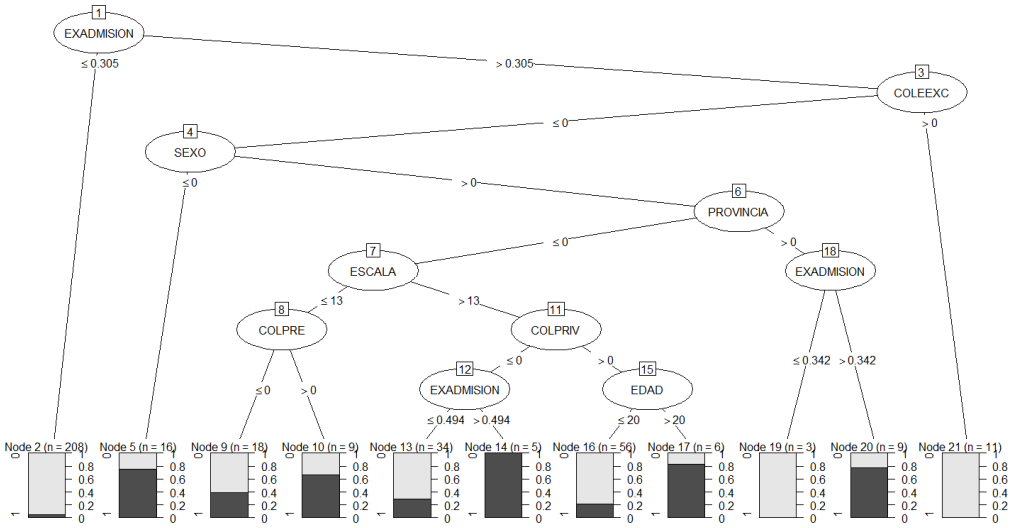


Figura 15. Modelo generado por árbol de decisiones para el plan 2010 filtrado por modalidad ordinario

Elaboración: *el autor*

Asimismo, según lo aprendido en los modelos de regresión lineal se aplicaron técnicas de árbol de decisión solo con los estudiantes que ingresaron por modalidad de examen ordinario. El modelo obtenido se muestra en la *Figura 15*.

Como se puede observar, es un modelo mucho más complejo con mayor uso de variables que anteriormente no se utilizaban para la creación de los modelos de predicción como escala o si el alumno viene de provincia. El resultado obtenido con este modelo sobre el dato de entrenamiento se muestra en la *Tabla 25*.

El modelo logra representar mucho mejor la data con una exactitud de 90.4%, con una sensibilidad de 62.66% y una especificidad de 97.33%.

Tabla 25
Matriz de confusión para modelo con árbol de decisiones Plan 2010 filtrado por modalidad ordinario

(a)	(b)	
292	8	(a): Clase 0
28	47	(b): Clase 1

Elaboración: *el autor*

El modelo logró predecir correctamente el 82.87% de los casos presentados en la data de prueba. Se obtuvo resultados similares con lo realizado en el modelo con todas las modalidades.

Utilizando la data del Plan de estudios más reciente se ha logrado obtener un modelo con una exactitud de 84% sobre los datos de prueba. Con este modelo para realizar predicciones se obtuvo la matriz de confusión que se muestra en la *Tabla 26* con una exactitud de casi 60%.

Utilizando la data filtrando solamente a los ingresantes se ha obtenido el modelo que se muestra en la *Figura 16,* en donde se destaca al igual que en los otros modelos la nota obtenida en el examen de admisión y variables relacionadas con el colegio de procedencia.

Tabla 26
Resultado de predicción modelo de árbol de decisiones Plan 2014

	PREDICCION		
REAL	0	1	Total
0	38 0.551%	9 0.130%	47
1	19 0.275%	3 0.043%	22
Total	57	12	69

Elaboración: *el autor*

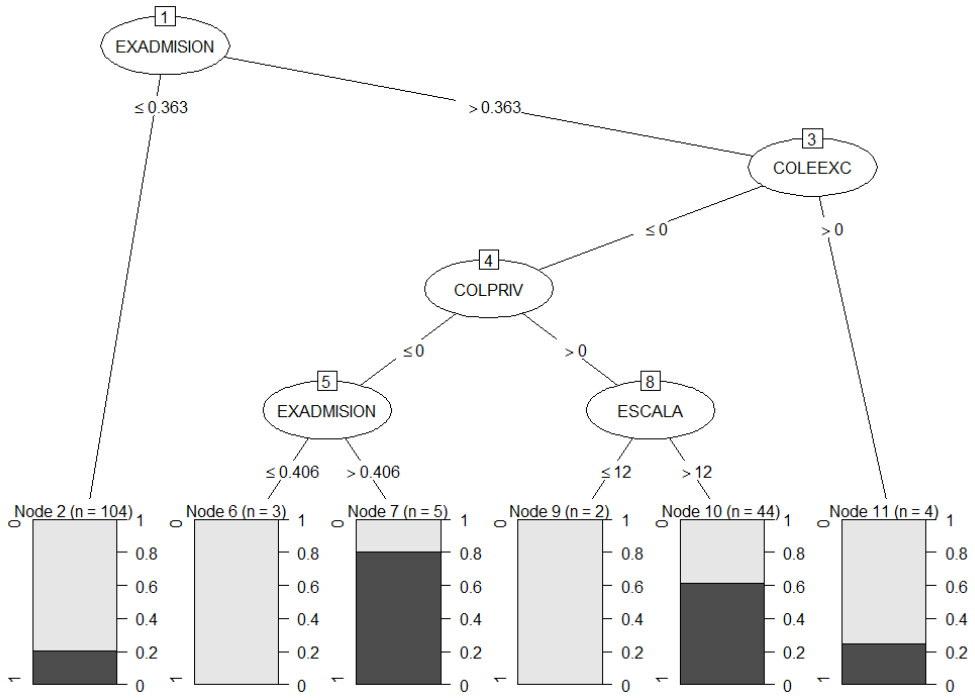


Figura 16. Modelo de árbol de decisiones Plan 2014 modalidad ordinario

Elaboración: *el autor*

Este modelo logra representar el 82.09% de los datos existentes y en la predicción logra una exactitud de 76.32% como se muestra en la Tabla 27.

Tabla 27
Resultados de predicción modelo de árbol de decisiones plan 2014 modalidad ordinario

	PREDICCIÓN		
REAL	0	1	Total
0	22 0.735%	4 0.061%	26
1	5 0.110%	7 0.094%	12
Total	27	11	38

Elaboración: *el autor*

4.3.4 Support Vector Machines

Por último, se aplicaron técnicas de Support Vector Machines (SVM) para crear modelos basados en vector de soporte para la creación de hiperplanos que corresponden a un punto de corte para realizar la clasificación. Un punto importante para SVM es la selección del tipo de kernel o núcleo con la que va a generar la división. Las funciones lineales y polinomiales no lograron encontrar ninguna manera de clasificar la data ya que como se muestra en la F*igura 17,* por ejemplo, que muestra la relación entre la nota de examen de admisión y distancia el algoritmo no logran encontrar maneras de separar linealmente la muestra.

Sí se lograron crear modelos para lograr la clasificación es con el kernel radial gaussiana y el kernel sigmoid. El mejor resultado fue el de radial gaussiana con una exactitud de 75.2%.

Finalmente, para el Plan de estudios 2014 – 2016, de igual manera, los modelos con kernel lineal, polinomial y radial se obtienen resultados similares con una exactitud de 68.11%, que puede ser resultado de falta de muestras para poder realizar una predicción adecuada.

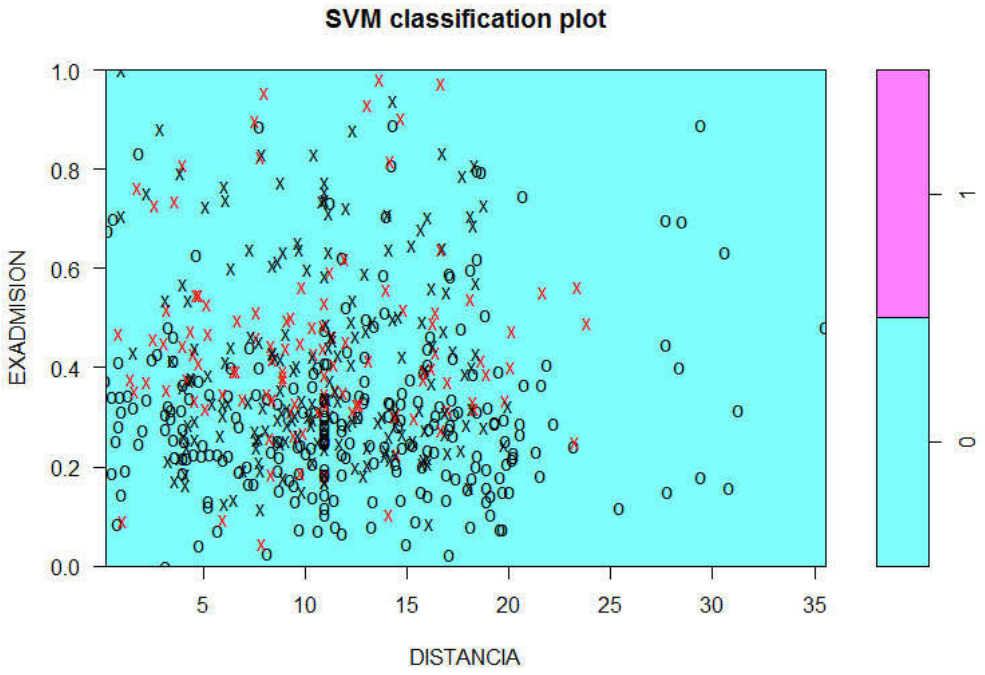


Figura 17. Modelo SVM de condición de aprobado con vectores nota de examen de admisión y distancia

Elaboración: *el autor*

CAPÍTULO V
RESULTADOS Y DISCUSIÓN

5.1 Evaluación de resultados

Se presentan los resultados obtenidos de la aplicación de minería de datos a la información obtenida de los estudiantes ingresantes según las hipótesis planteadas en la presente investigación

- Hg: El rendimiento académico puede predecirse mediante minería de datos para estudiantes del primer ciclo de la Escuela Profesional de Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres.

La predicción realizada por las distintas técnicas de minería de datos ha brindado los siguientes niveles de exactitud que se muestran en la *Tabla 28*. Los resultados expresan que existen suficientes evidencias para indicar que si es posible realizar predicciones a futuro para clasificar a los nuevos ingresantes si son capaces de aprobar todos los cursos en su primer ciclo. Se alcanzó valores máximos de 90.4% en el mejor modelo encontrado con una exactitud en la predicción de 82.87%

Tabla 28
Resultados obtenido de la predicción con las diferentes técnicas de minería de datos

	Plan 2010	Plan 2014
Regresión logística	74.4%	75%
C5.0	82.87%	76.32%
SVM	75.2%	68.11%

Elaboración: *el autor*

- H1: El rendimiento académico en estudiantes universitarios de primer ciclo puede estimarse mediante indicadores sociales, económicos y académicos.

Los diferentes datos encontrados en la información extraída han sido agrupados como indicadores sociales, económicos y académicos de selección relacionados con los aspectos que podrían influir en el rendimiento académico de los estudiantes ingresantes.

De los resultados obtenidos del análisis estadístico ANOVA se observa que existe una diferencia, estadísticamente significativa, entre los valores promedios con un nivel de 95.0% de confianza dado que se obtuvo un valor-P de la prueba-F menor a 0.05. Asimismo, de las técnicas de minería de datos se observa en múltiples variables que la misma prueba también se obtuvo un valor-P menor a 0.05, como se observa en la *tabla 29*.

- H2: La significación de los indicadores sociales, económicos y académicos en estudiantes universitarios de primer ciclo puede determinarse mediante el análisis de componentes principales.

En el análisis de la relación entre las categorías base (social, económico y académico) con el rendimiento se encontró que el carácter social tuvo mayor número de estudiantes como razón de no culminación con 116 con respecto a lo económico y académico presentaron que 114. Se encontraron también diferencias, estadísticamente significativas ($p \leq 0,05$), entre el número de estudiantes por categoría, a través de análisis de varianza y prueba de contraste múltiple.

La interacción entre las categorías se observó que lo económico y lo académico, al no tener diferencias significativas entre ellas, resultó con un valor predominante. También se descubrió que la categoría social no influye en las otras categorías, reflejando la situación particular de la universidad en estudio en donde se permite continuar con sus estudios, aunque no estén al día en sus pagos por derechos de enseñanza.

Finalmente, el análisis de componentes principales (*Tabla 29*), muestran un efecto acumulativo influyente o determinante por la categoría referida a lo social, reflejando los resultados de las relaciones entre las diferentes categorías.

Tabla 29
Análisis de componentes principales según el peso de cada categoría base y por interacción

Componente	Eigenvalor	Porcentaje de	
Número		Varianza	Acumulado
R B S	13.5826	75.459	75.459
R M S	4.41742	24.541	**100.000**
R A S	1.17266E-15	0.000	100.000
R B E	5.69011E-16	0.000	100.000
R M E	3.20812E-16	0.000	100.000
R A E	2.31751E-16	0.000	100.000
R B PL	2.04724E-16	0.000	100.000
R M PL	1.60915E-16	0.000	100.000
R A PL	1.48804E-16	0.000	100.000
R B SE	6.45904E-17	0.000	100.000
S M SE	4.26175E-17	0.000	100.000
R A SE	0.0	0.000	100.000
R B EP	0.0	0.000	100.000
R M EP	0.0	0.000	100.000
R A EP	0.0	0.000	100.000
R B SEPL	0.0	0.000	100.000
R M SEPL	0.0	0.000	100.000
R A SEPL	0.0	0.000	100.000

Elaboración: *el autor*

La aplicación de técnicas de regresión ha encontrado la existencia de varios factores o variables que influyen en la condición de aprobado. Las variables que según ese modelo son importantes, es decir, tienen un p <=

0.05, que permiten demostrar la validez de esta hipótesis son las que se muestran en la *Tabla 30*.

Tabla 30
Variables independientes relacionados con la variable dependiente rendimiento académico

VARIABLE	VALOR P	PENDIENTE
Nota de examen de admisión	0.0000000131	Positiva
Edad	0.0166	Positiva
Sexo	0.022663	Negativa
Distancia	0.027194	Negativa
Colegio de Excelencia	0.0470	Negativa
Escala	0.06845	Positiva

Elaboración: *el autor*

De las variables encontradas como influyentes, en muchos casos, concuerdan con lo encontrado en la revisión de la literatura en donde el rendimiento académico anterior a la universidad, que se refleja en la nota obtenida en el examen de admisión, es el que causa mayor impacto en el rendimiento académico. Tanto la edad como la nota en examen de admisión tienen una pendiente positiva, que significa que a mayor nota y edad tiene mayor probabilidad de obtener un buen rendimiento. Asimismo, la pendiente negativa en la variable sexo y distancia, indica que el sexo femenino y la menor distancia del lugar de estudios favorecen a la variable dependiente.

Un resultado un poco inesperado es con respecto a la variable que representa a los que provienen de colegios considerados de excelencia. Se esperaba que alumnos que vienen de los mejores colegios obtengan mejores resultados en la universidad; sin embargo, como se refleja en los resultados, esta variable influye negativamente lo que significa que, al menos para la población en estudio, haber estudiado la secundaria en un buen colegio no significa que obtengan buen rendimiento académico como ingresantes de la universidad.

Asimismo, las técnicas de árbol de decisiones han logrado confirmar que las variables son importantes y tienen una influencia sobre los ingresantes, lo que permite crear modelos basados en ellos para realizar predicciones. Adicionalmente, se ha encontrado que las variables

relacionadas con el colegio de procedencia también son influyentes según los resultados de la aplicación de esta técnica por lo que se puede afirmar que si bien es cierto que no existe seguramente alguna relación lineal entre estas variables con la variable dependiente si son importantes de ser considerados para la realización de predicciones.

En cuanto a otras variables analizadas, como la escala de pensión o la procedencia del ingresante de un colegio de provincia no se han logrado demostrar que influyen en la variable dependiente, por lo cual podemos concluir que no son necesarios de ser considerados para estudios futuros.

- H3: La minería de datos educacional puede aplicarse para indicadores sociales, económicos y académicos en estudiantes universitarios de primer ciclo.

Según los resultados, existen diferencias marcadas entre las tres técnicas de minería de datos utilizadas. El mejor resultado obtenido fue el de algoritmo C5.0 de árbol de decisiones, con mejoras de hasta 15% en la predicción del rendimiento académico de los ingresantes.

Esto se debe a que las características de la data de los ingresantes según los resultados de esta tesis no es posible realizar una representación muy exacta mediante un modelo lineal y la definición de un conjunto de reglas como el árbol de decisiones se adapta mejor.

Para el caso de SVM, que según los textos (Lantz, 2013) es posible crear modelos que permiten representar diferentes tipos de data sin ser influenciados por problemas de data con ruido. Sin embargo, en este caso, no se lograron resultados muy diferentes con el de modelo más simple que es el lineal a pesar de ser un algoritmo mucho más complejo. Se sospecha como causa de esta situación que el formato y la calidad de la data utilizada no están adecuadas para este tipo de algoritmo avanzado, como ocurre normalmente en algoritmos de redes neuronales, que requieren que los datos cumplan con

ciertas características específicas, además de requerir una mayor cantidad de muestras para obtener mejores resultados

Los resultados encontrados sobre los factores que influyen en el rendimiento académico, como se indica en el punto anterior sí permiten agrupar dichos factores y establecer un perfil para los ingresantes que tendrían mayores probabilidades de cursar el primer ciclo de estudios aprobando todos los cursos y también los que poseerían dificultades o algún apoyo adicional para aprobar el primer ciclo.

Dado que se han encontrado diferencias marcadas entre los ingresantes varones y damas, se ha establecido un perfil para cada uno de ellos. Los perfiles establecidos se encuentran en las Figuras 18 y 19. En cuanto al perfil de ingresantes mujeres, es relativamente simple, si no proviene de colegio de excelencia y tiene una nota en examen de admisión mayor al 30.3% del total, se considera que no se encontrará con problemas durante su primer ciclo. El perfil de ingresantes varones, de igual manera, presenta la condición de que no debe provenir de colegios considerados de excelencia. Si ha realizado sus estudios secundarios en provincia debe tener nota en examen de admisión mayor a 34.2%. En caso de haber realizado sus estudios en colegio privado de Lima, debe tener nota en examen de admisión mayor a 30.3% y tener edad mayor a 20 años y si lo realizó en cualquier otro tipo de colegio en Lima una nota mayor a 49.4% del total. Aquellos que no cumplan con los perfiles definidos deberán ser considerados como ingresantes que podrían enfrentarse con problemas académicos y requieren algún tipo de ayuda adicional.

Como perfil general debemos considerar que los ingresantes que obtuvieron mejores resultados en el primer ciclo también tienen notas relativamente altas en el examen de admisión, sin depender de la modalidad de ingreso que pueden representar diferentes niveles de dificultad para alcanzar un cupo e ingresar a la universidad.

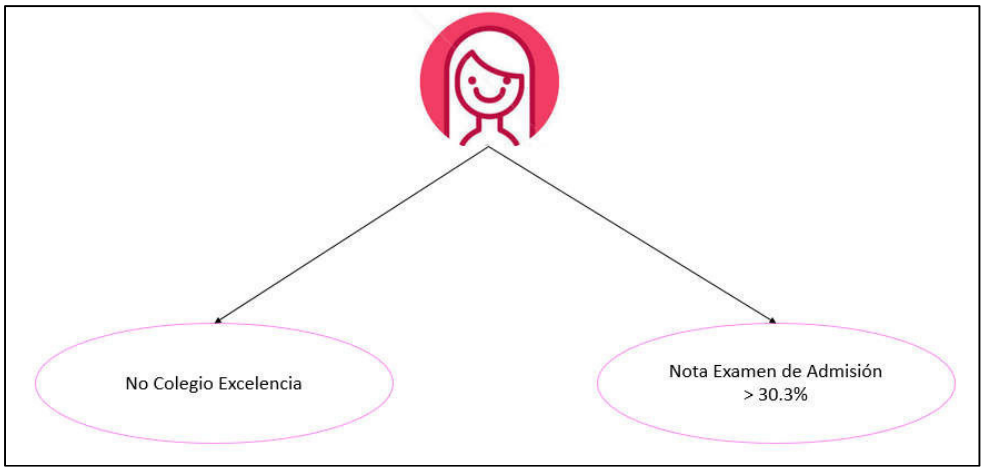


Figura 18. Perfil ingresantes sexo femenino

Elaboración: *el autor*

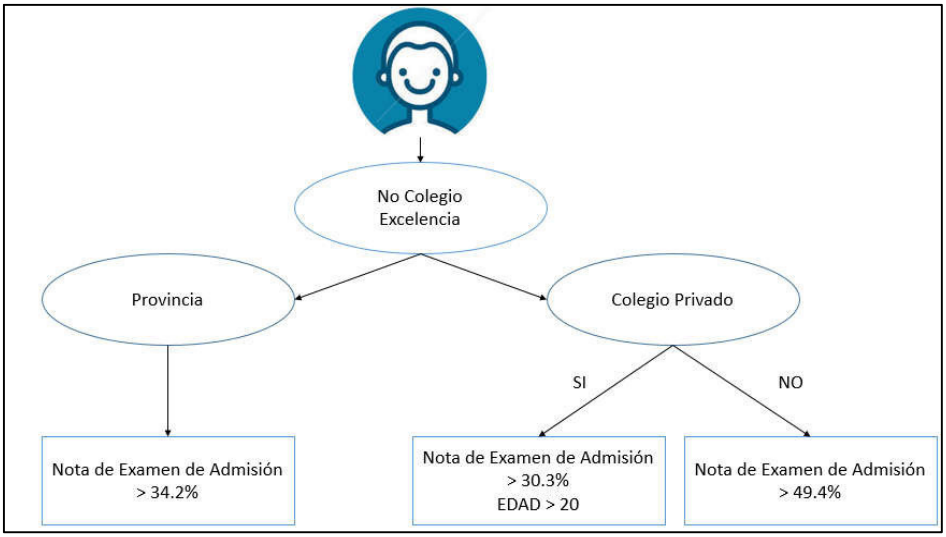


Figura 19. Perfil ingresantes sexo masculino

Elaboración: *el autor*

También es importante mencionar que dado el caso en la que todas las características sean similares, los ingresantes de sexo femenino tendrán mayores posibilidades de cursar sin problemas el primer ciclo.

El colegio de procedencia es también un factor importe a considerar según los resultados de la aplicación de minería de datos, pero no por las razones que, normalmente se sospecharía, ya que se ha encontrado que proceder de un colegio grande o importante no garantiza que el ingresante

tenga mejores resultados. Se podría considerar que los problemas a los que se enfrentan estos alumnos de colegios importantes no son académicos sino son más de adaptación o motivación para realizar sus estudios. Asimismo, no se ha logrado demostrar que ingresantes que han cursado estudios secundarios en un colegio estatal, privado o pre-universitario tenga un impacto en el rendimiento académico en el primer ciclo de universidad.

Vivir a distancias relativamente cercanas es otro factor importante, obteniéndose en muchos casos que, según el modelo de árbol de decisiones, los ingresantes obtienen mejores resultados en sus estudios. Sin embargo, en esta tesis no se ha considerado el medio de transporte que los ingresantes utilizaron para asistir a su casa de estudios, dado que se ha trabajado con data histórica y esa información no se encontraba en las bases de datos.

5.2 Discusión

La presente investigación tuvo como objetivo principal desarrollar mediante la aplicación de técnicas de minería de datos modelos que permitan analizar y predecir el rendimiento académico de los futuros ingresantes. La minería de datos permite automatizar a través de la computación los diferentes procedimientos estadísticos necesarios para crear y validar los distintos modelos de predicción

Uno de los aspectos importantes para este tipo de investigación es la calidad de la data utilizada para realizar las predicciones. Fueron necesarias al igual que en otros estudios similares varias revisiones y ajustes menores a la data para las diferentes técnicas aplicadas. También es importante considerar la influencia de la cantidad de muestras obtenidas para crear los modelos y realizar la predicción. La cantidad de data extraída de las bases de datos muestra diferencias de aproximadamente 20% por ciclo de ingresantes y como se observa, en los resultados, con el plan antiguo se obtuvieron más muestras y modelos más exactos.

Las predicciones realizadas no lograron el 95% de exactitud necesarios para negar complemente la hipótesis nula. Los mejores resultados de la predicción se obtuvieron con el algoritmo C5.0 de árbol de decisiones.

Los resultados obtenidos con técnicas de SVM no lograron cumplir con las expectativas de crear modelos con mayor capacidad de predicción. Se presume que se requiere de mayor cantidad de análisis y ajuste a la data utilizada para obtener mejores resultados.

Para investigaciones que requieren realizar un análisis de los efectos de las variables independientes a la dependiente es mejor utilizar técnicas como regresión y árbol de decisiones que permiten observar cómo se obtuvo el modelo y qué variables son las que más influyen. Las técnicas como SVM y de redes neuronales si bien es cierto que, son más poderosas, tienen como característica generar modelos complejos de caja negra, lo que dificulta interpretar el modelo.

De las variables analizadas se puede concluir que una mayor nota en el examen de admisión resulta, en la mayoría de los casos, un mejor rendimiento en el primer ciclo. Asimismo, los ingresantes de sexo femenino tienen mayores probabilidades de aprobar los cursos del primer ciclo, lo cual sería interesante para un estudio posterior si existen aspectos emocionales o de madurez que influyen en el rendimiento académico.

El colegio de procedencia presentó resultados un poco inesperados ya que es común asumir que haber estudiado en un colegio considerado como “bueno” tienen estudiantes con mayores capacidades académicas; sin embargo, se encontró que no es el caso al menos para la muestra utilizada ya que se hallaron varios casos en la que estudiantes de colegios estatales obtuvieron mejor rendimiento. Se presume que existen aspectos emocionales relacionados con sus estudios como motivación, prioridades en su vida, adaptación del colegio a la vida universitaria, entre otros.

Sobre la distancia se ha podido encontrar evidencias que sí son influyentes en el rendimiento académico, pero para el caso de esta investigación se ha utilizado información histórica obtenida de las bases de datos por lo que no se han tomado en consideración aspectos como medio de transporte utilizado o más, específicamente, el tiempo promedio que demoran

en trasladarse desde sus casas hacia la universidad, aspectos que se podrían considerar para estudios posteriores.

En general, se ha logrado demostrar la importancia de las técnicas de minería de datos para realizar análisis estadístico y predicción a base de la data relacionada con los ingresantes. No se han aplicado para este estudio técnicas de inteligencia artificial como redes neuronales que podrían ser consideradas como una continuación de este estudio para ver la posibilidad de alcanzar modelos con mayor capacidad de predicción.

CONCLUSIONES

1. Se logró predecir a través de técnicas estadísticos y minería de datos el rendimiento académico de los estudiantes ingresantes a la Escuela Profesional de Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres.

2. Se identificaron los principales indicadores que permiten predecir el rendimiento académico de los estudiantes ingresantes y aplicaron pruebas que permitieron validar los resultados obtenidos como estadísticamente significantes.

3. Se logró demostrar la influencia que tienen los diferentes factores de los ingresantes en el rendimiento académico en su primer ciclo de estudios. Las variables creadas a partir de la información extraída de la base de datos de la universidad permitieron crear perfiles que ayudaron a la identificación temprana de estudiantes que podrían encontrarse con dificultades en sus estudios.

4. Se aplicaron tres técnicas de minería de datos para realizar la predicción de rendimiento académico. El algoritmo C5.0 de árbol de decisiones es el que obtuvo los mejores resultados con una exactitud de predicción de 82.87%. Asimismo, la facilidad de interpretación de los resultados de la técnica de árbol de decisiones lo convierten en la mejor técnica a utilizar para este tipo de estudio.

RECOMENDACIONES

1. La calidad de la data es muy importante para este tipo de investigación. La data obtenida de las bases de datos de la universidad provenía de sistemas transaccionales por la que en muchos casos el formato y el tipo de datos no eran los ideales y fue necesario realizar varios procesos de limpieza y transformación de datos.

2. Actualmente, no existen mecanismos de almacenamiento de data histórica como Data Warehouse o Data Mart en la universidad. Dado el crecimiento de la importancia y la cantidad de datos generados en los últimos años, se debería almacenar la data histórica que se genera examinarla en un futuro.

3. No se consideraron aspectos emocionales en esta investigación por haberse trabajado con data histórica. Se encontraron algunas evidencias de que elementos como la madurez y la motivación son importantes en

el rendimiento de los alumnos, por lo que sería interesante realizar estudios futuros integrando estos aspectos.

4. No se obtuvo resultados esperados con Support Vector Machines que es considerada una técnica más avanzada y que, en otros estudios obtuvieron mejores resultados. Se sospecha que se debe a que la data no se ajustaba a lo requerido para aplicar estas técnicas de caja negra por lo que, si en estudios futuros se busca aplicar SVM, redes neuronales u otro algoritmo genético, en este tipo de data se recomienda realizar mayores estudios para validar su correcto funcionamiento.

FUENTES DE INFORMACIÓN

Bibliográficas:

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

Lantz, B. (2013). *Machine learning with R*. Packt Publishing Ltd.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. (Eds.). (2011). *Handbook of educational data mining*. CRC Press.

Sampieri, R. H., Collado, C. F. & Lucio, P. B. (2014). *Metodología de la investigación* (Vol. 6). México: Mcgraw-hill.

Hemerográficas:

Aluko, R. O., Adenuga, O. A., Kukoyi, P. O., Soyingbe, A. A., & Oyedeji, J. O. (2016). Predicting the academic success of architecture students by pre-enrolment requirement: using machine-learning techniques. *Construction Economics and Building*, *16*(4), 86-98.

Arsad, P. M., & Buniyamin, N. (2013). A neural network students' performance prediction model (NNSPPM). In *Smart Instrumentation, Measurement and Applications (ICSIMA), 2013 IEEE International Conference on* (pp. 1-5). IEEE.

Bayer, J., Bydzovská, H., Géryk, J., Obsivac, T., & Popelinsky, L. (2012). Predicting Drop-Out from Social Behaviour of Students. *International Educational Data Mining Society*.

Bienkowski, M., Feng, M., & Means, B. (2012). Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. US Department of Education, Office of Educational Technology, 1, 1-57.

Bin Mat, U., Buniyamin, N., Arsad, P. M., & Kassim, R. (2013, December). An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention. In *Engineering Education (ICEED), 2013 IEEE 5th Conference on* (pp. 126-130). IEEE.

Brito-Jiménez, I. T., & Palacio-Sañudo, J. (2016). Calidad de vida, desempeño académico y variables sociodemográficas en estudiantes universitarios de Santa Marta-Colombia. *Duazary*, 13(2), 133-141.

Brown, J. L. (2012). Developing a freshman orientation survey to improve student retention within a college. *College Student Journal*, 46(4), 834-851.

Budny, D. D., Paul, C. A., & Newborg, B. B. (2014). Involving Parents at Step One in the Freshman Engineering Experience. *International Journal Of Engineering Pedagogy*, 4(2), 10-17.

Calisir, F., Basak, E., & Comertoglu, S. (2016). Predicting academic performance of master's students in engineering management. *College Student Journal*, 50(4), 501-513.

Chalaris, M., Gritzalis, S., Maragoudakis, M., Sgouropoulou, C., & Tsolakidis, A. (2014). Improving quality of educational processes providing new knowledge using data mining techniques. *Procedia-Social and Behavioral Sciences*, *147*, 390-397.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (1999). *The CRISP-DM user guide*. In 4th CRISP-DM SIG Workshop in Brussels in March (Vol. 1999).

Cheewaprakobkit, P. (2013). Study of Factors Analysis Affecting Academic Achievement of Undergraduate Students in International Program. *In Proceedings of the International MultiConference of Engineers and Computer Scientists* (Vol. 1, pp. 13-15).

Chen, J. F., & Do, Q. H. (2014). A cooperative Cuckoo Search–hierarchical adaptive neuro-fuzzy inference system approach for predicting student academic performance. *Journal of Intelligent & Fuzzy Systems*, *27*(5), 2551-2561.

Cupani, M., Garrido, S., & Tavella, J. (2013). El Modelo de los Cinco Factores de Personalidad: contribución predictiva al rendimiento académico The Five Factor Model of Personality and its contribution to the prediction of academic per. *Revista de Psicología*, *9*(17), 67-86.

Ecklund, A. P. (2013) Enhancing Incoming Male Student Retention: An Analysis of the Experiences of Persistence in Engineering.

Elakia, G., & Aarthi, N. J. (2014). Application of data mining in educational database for predicting behavioural patterns of the students. International Journal of Computer Science and Information Technologies (IJCSIT), 5(3), 4649-4652.

ElGamal, A.F. (2013). An Educational Data Mining Model for Predicting Student Performance in Programming Course. International Journal of Computer Applications, 70(17).

Gray, G., McGuinness, C., & Owende, P. (2014). An application of classification models to predict learner progression in tertiary education. In *Advance Computing Conference (IACC), 2014 IEEE International* (pp. 549-554). IEEE.

Honken, N. B., & Ralston, P. A. (2013). High-Achieving High School Students and Not So High-Achieving College Students A Look at Lack of Self-Control,

Academic Ability, and Performance in College. Journal of Advanced Academics, 24(2), 108-124.

Jin, Q., Imbrie, P. K., Lin, J. J., & Chen, X. (2011). A multi-outcome hybrid model for predicting student success in engineering. In American Society for Engineering Education. American Society for Engineering Education.

Kumar S., Vijayalakshmi M.N. (2012). Appraising the significance of self-regulated learning in higher education using neural networks, International Journal of Engineering Research and Development Volume 1 (Issue 1) 09-15.

Lang, C., Siemens, G., Wise, A., & Gasevic, D. (Eds.). (2017). Handbook of learning analytics. SOLAR, Society for Learning Analytics and Research.

Li, K. F., Rusk, D., & Song, F. (2013). Predicting student academic performance. In *Complex, Intelligent, and Software Intensive Systems (CISIS), 2013 Seventh International Conference on* (pp. 27-33). IEEE.

Liñán, L. C., & Pérez, Á. A. J. (2015). Educational Data Mining and Learning Analytics: differences, similarities, and time evolution. *International Journal of Educational Technology in Higher Education*, *12*(3), 98-112.

López Bonilla, J. M., López Bonilla, L. M., Serra, F., & Ribeiro, C. (2015). Relación entre actitudes hacia la actividad física y el deporte y rendimiento académico de los estudiantes universitarios españoles y portugueses. *Revista iberoamericana de psicología del ejercicio y el deporte*, *10*(2), 275-284.

Martínez, L. N., Llorca, J. A. S., Tello, F. P. H., & Mira, I. J. (2016). Las metas múltiples: análisis predictivo del rendimiento académico en estudiantes chilenos (multiple goals: predictive analysis of academic achievement in chilean students). *Educación XX1*, *19*(1), 267.

Mishra T., Kumar D. & Gupta S. (2014) Mining Students' Data for Prediction Performance Fourth International Conference *on Advanced Computing & Communication Technologies*, Rohtak, pp. 255-262.

Mohamed, S.A., Husaina, W. & Abdul, R.N. (2015). The Third Information Systems International Conference A Review on Predicting Student’s

Performance using Data Mining Techniques. *Procedia Computer Science*; 72, 414 – 422. Disponible en: https://doi.org/10.1016/j.procs.2015.12.157

Papamitsiou, Z. K., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Educational Technology & Society*, *17*(4), 49-64.

Park, H., Behrman, J. R., & Choi, J. (2013). Causal effects of single-sex schools on college entrance exams and college attendance: Random assignment in Seoul high schools. Demography, 50(2), 447-469.

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, *41*(4), 1432-1462.

Ramesh, V., Parkavi, P., & Ramar, K. (2013). Predicting student performance: a statistical and data mining approach. *International journal of computer applications*, *63*(8).

Rasberry, C. N., Lee, S. M., Robin, L., Laris, B. A., Russell, L. A., Coyle, K. K., & Nihiser, A. J. (2011). The association between school-based physical activity, including physical education, and academic performance: a systematic review of the literature. *Preventive medicine*, *52*, S10-S20.

Rodríguez, Á. P. A., & Arenas, D. A. M. (2016). Programas de intervención para Estudiantes Universitarios con bajo rendimiento académico. *Informes Psicológicos*, *16*(1), 13-34.

Romero, C., & Ventura, S. (2013). Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12-27.

Sembiring, S., Zarlis, M., Hartama, D., Ramliana, S., & Wani, E. (2011). Prediction of student academic performance by an application of data mining techniques. In International Conference on Management and Artificial Intelligence IPEDR (Vol. 6, pp. 110-114).

Sepehrian, F. (2012). Emotional Intelligence as a predictor of academic performance in university. *Journal of Educational Sciences and Psychology*, *2*(2).

Shahiri, A. M., & Husain, W. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, *72*, 414-422.

Shull, P. J., Ford, P., & Carrier, K. (2013). Extended Abstract-Retention of Marginal Students: Effective First Year Interventions.

Siller T., Paguyo C. (2012). “Use of Pass/Fail Grading to increase First Year Retention”, 4th First Year Engineering Experience (FYEE) Conference.

Sin, K., & Muthu, L. (2015). Application of big data in education data mining and learning analytics — A literature review. *ICTACT journal on soft computing*, *5*(4), 1-035.

Singh, W., & Kaur, P. (2016). Comparative Analysis of Classification Techniques for Predicting Computer Engineering Students' Academic Performance. *International Journal of Advanced Research in Computer Science*, *7*(6).

Strecht, P., Cruz, L., Soares, C., Merdes-Moreria, J. & Abren, R. (2015). A Comparative Study of Classification and Regression Algorithms for Modelling Students’ Academic Performance. In 8th International Conference on Educational Data Mining, Madrid, Spain, 392-395.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (pp. 29-39).

Yadav, S. K., & Pal, S. (2012). Data mining: A prediction for performance improvement of engineering students using classification. World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 2, No. 2, 51-56, 2012.

York, T. T., Gibson, C., & Rankin, S. (2015). Defining and measuring academic success. *Practical Assessment, Research & Evaluation*, *20*(5), 2.

Electrónicas:

Lenguaje de programación R
http://www.r-project.org/

Minería de Datos con R
http://www.rdatamining.com/

Exploración de Datos con R
http://www.rdatamining.com/examples/exploration

Árbol de Decisiones con R
http://www.rdatamining.com/examples/decision-tree

Agrupamiento utilizando algoritmo k-means
http://www.rdatamining.com/examples/kmeans-clustering

Detección de anomalías o valores atípicos (outlier detection)
http://www.rdatamining.com/examples/outlier-detection

Manual Paquete C5.0

https://cran.r-project.org/web/packages/C50/C50.pdf

Manual Paquete CaTools

https://cran.r-project.org/web/packages/caTools/caTools.pdf

Manual Paquete Kernlab

https://cran.r-project.org/web/packages/kernlab/kernlab.pdf

Manual Paquete ggplot2

https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf

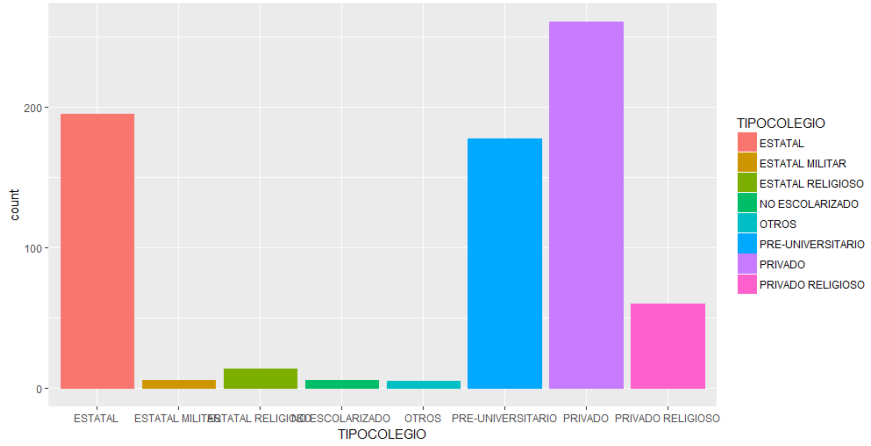
Ministry of Education Malaysia: MEM. (2015). National higher education strategic plan (2015). URL: http://www.ums.edu.my/bendahari/images/dokumen/2.%20Pembentangan%20Ybrs.%20Prof%20Madya%20Dr.%20Norhayati%20Muhamad.pdf

ANEXOS

1. **Resultados de análisis exploratorio**
2. **Scripts desarrollados en R**
3. **Resultados Análisis de Varianza ANOVA**

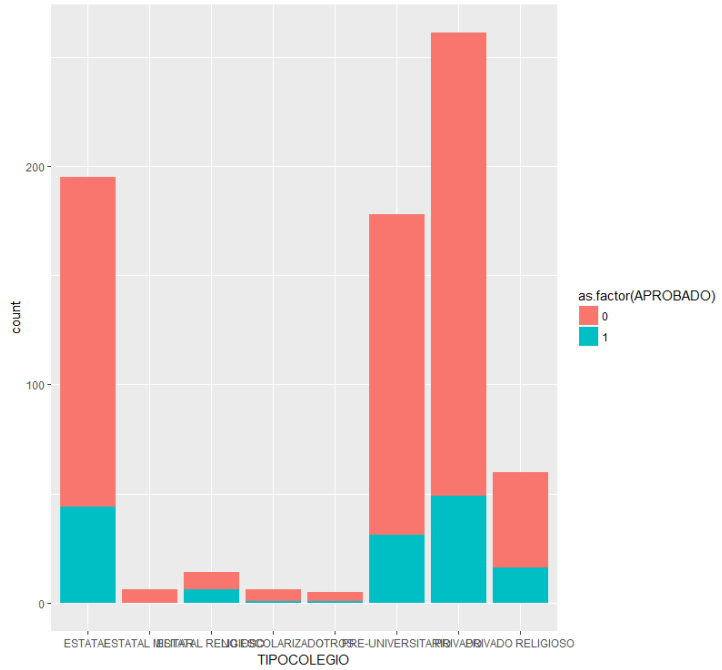
ANEXO N° 1 Resultados de análisis exploratorio

Ingresantes por tipo de colegio



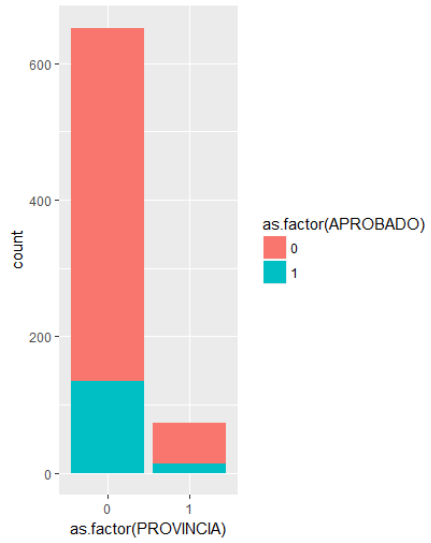

```
gtipocol <- ggplot(old, aes(x=TIPOCOLEGIO))

gtipocol + geom_histogram(aes(color = TIPOCOLEGIO, fill = TIPOCOLEGIO), stat="count")
```

Ingresantes por tipo de colegio aprobados y desaprobados




```
gtipocol + geom_histogram(aes(fill=as.factor(APROBADO)), stat="count")
```

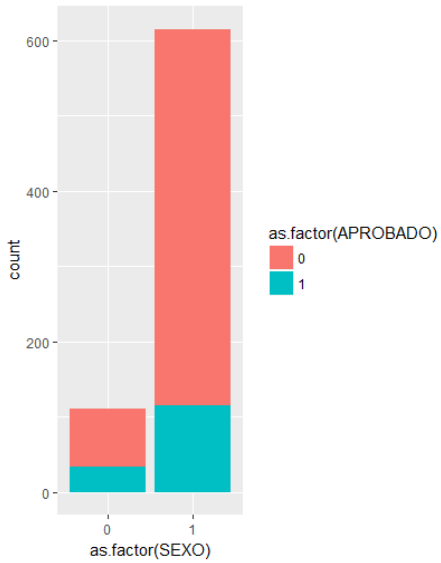
Ingresantes por condición procedente de provincias aprobados y desaprobados




```
gcolprov <- ggplot(old, aes(as.factor(PROVINCIA)))
```

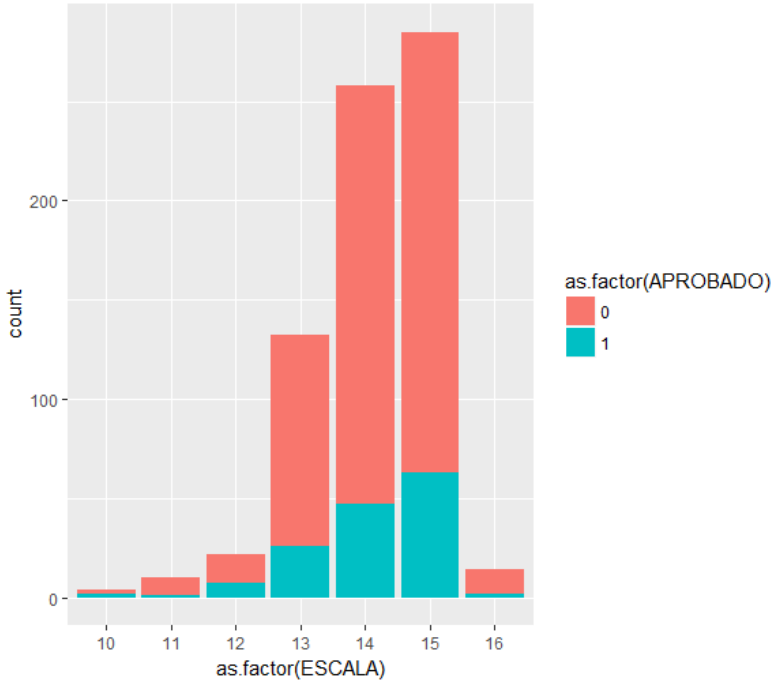
```
gcolprov + geom_histogram(aes(fill=as.factor(APROBADO)), stat ="coun
t")
```

Ingresantes según sexo aprobados y desaprobados



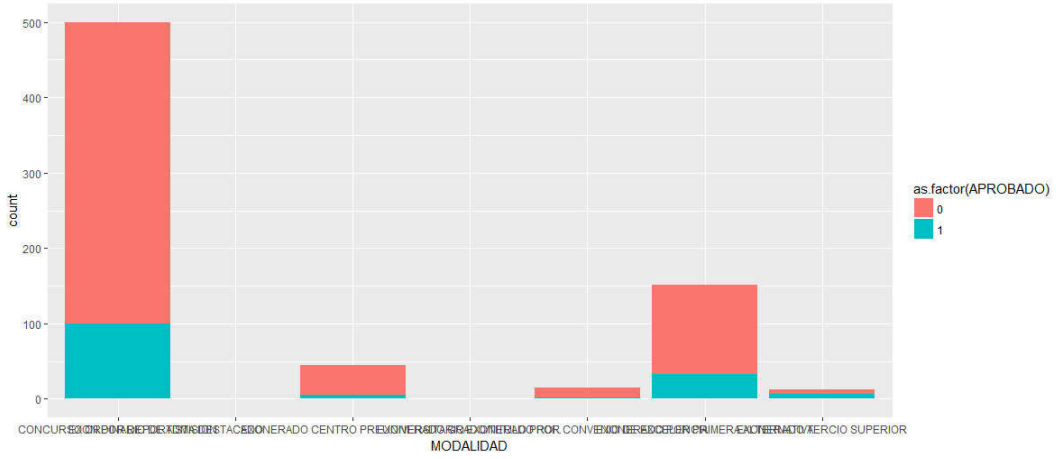

```
gsexo <- ggplot(old, aes(as.factor(SEXO)))
gsexo + geom_histogram(aes(fill=as.factor(APROBADO)), stat ="count")
```

Ingresantes por escala aprobados y desaprobados



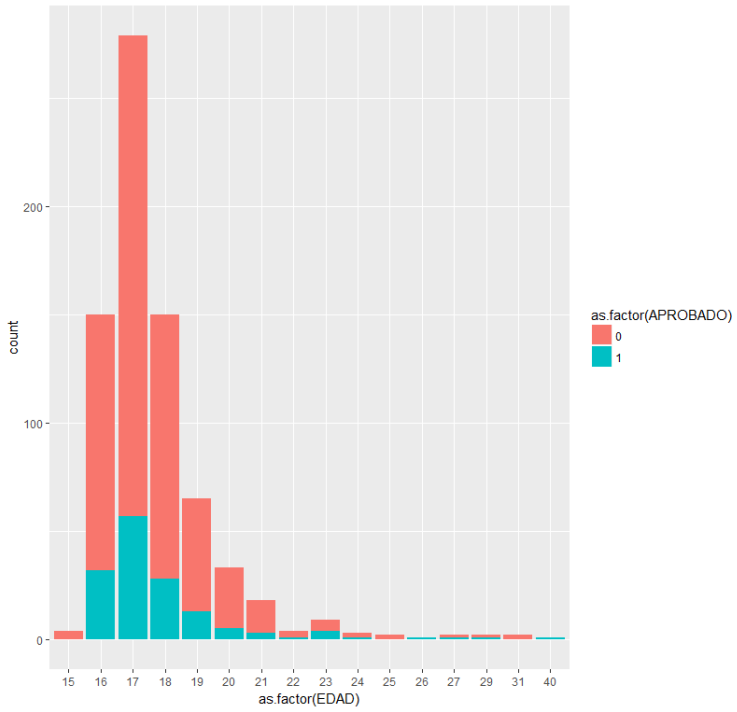

```
gescala <- ggplot(old, aes(as.factor(ESCALA)))
gescala + geom_histogram(aes(fill=as.factor(APROBADO)), stat="count"
)
```

Ingresantes por modalidad aprobados y desaprobados



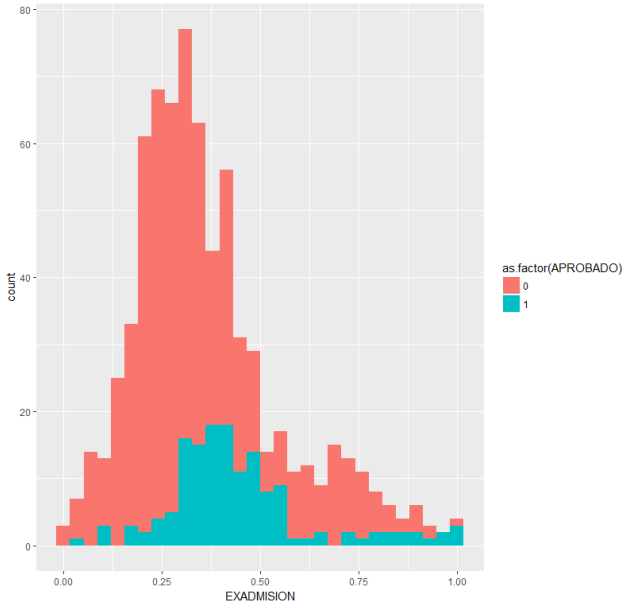

```
gmodalidad = ggplot(old, aes(MODALIDAD))
gmodalidad + geom_histogram(aes(fill=as.factor(APROBADO)), stat="cou
nt")
```

Ingresantes por edad aprobados y desaprobados



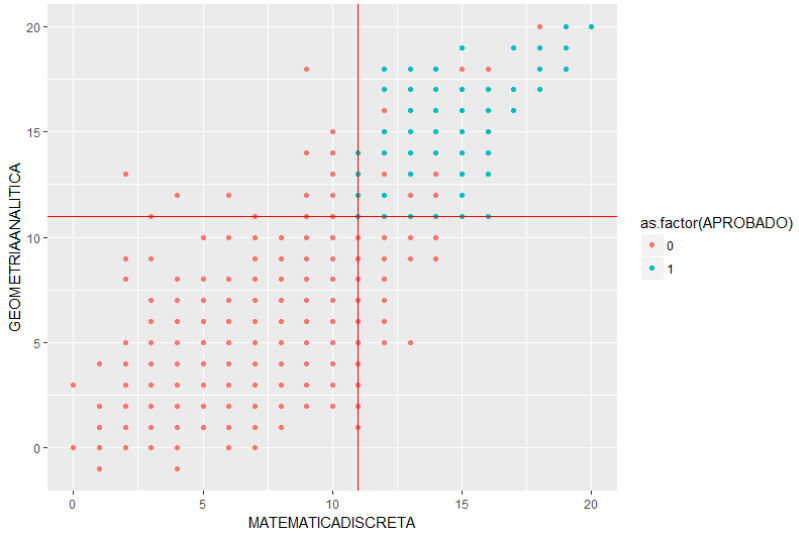

```
gedad <- ggplot(old, aes(as.factor(EDAD)))
gedad + geom_histogram(aes(fill=as.factor(APROBADO)), stat="count")
```

Ingresantes por nota obtenida en examen de admisión con condición aprobados y desaprobados



```
gexamadm <- ggplot(old, aes(EXADMISION))
gexamadm + geom_histogram(aes(fill = as.factor(APROBADO)))
```

Notas obtenidas en los cursos Geometría Analítica y Matemática Discreta por condición aprobado.




```
g1 = ggplot(old, aes(x=MATEMATICADISCRETA, y=GEOMETRIAANALITICA))

g1 + geom_point(aes(color = as.factor(APROBADO))) + geom_vline(xinte
rcept = 11, color="red")+ geom_hline(yintercept = 11, color = "red")
```

Resumen de datos obtenido por la función summary

```
UniqueID         CICLO            ANIO               SEMESTRE          EDAD
 Min.   :  1.0   Min.   :20101   Min.   :2.010e+03   Min.   :1.000   Min.
:15.0
 1st Qu.:227.0   1st Qu.:20102   1st Qu.:2.010e+03   1st Qu.:1.000   1st Qu
.:17.0
 Median :458.0   Median :20112   Median :2.011e+03   Median :1.000   Median
:17.0
 Mean   :454.9   Mean   :20115   Mean   :2.776e+06   Mean   :1.335   Mean
:17.7
 3rd Qu.:690.0   3rd Qu.:20122   3rd Qu.:2.012e+03   3rd Qu.:2.000   3rd Qu
.:18.0
 Max.   :927.0   Max.   :20132   Max.   :2.011e+09   Max.   :2.000   Max.
:40.0

      SEXO          PROVINCIA
 Min.   :0.0000   Min.   :0.0000
 1st Qu.:1.0000   1st Qu.:0.0000
```

```
 Median :1.0000   Median :0.0000
 Mean   :0.8483   Mean   :0.1007
 3rd Qu.:1.0000   3rd Qu.:0.0000
 Max.   :1.0000   Max.   :1.0000


NOMCOLEGIO
 SACO OLIVEROS         : 35
 TRILCE                : 32
 PAMER                 : 18
 CENTRO PRE USMP       : 16
 TRILCE SALAVERRY      :  6
 (Other)               :617
 NA's                  :  1
            TIPOCOLEGIO     COLEEXC            COLPRE           COLPRIV
 PRIVADO           :261   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
 ESTATAL           :196   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000
 PRE-UNIVERSITARIO:177   Median :0.00000   Median :0.0000   Median :1.0000
 PRIVADO RELIGIOSO: 60   Mean   :0.08414   Mean   :0.2359   Mean   :0.6648
 ESTATAL RELIGIOSO: 14   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:1.0000
 ESTATAL MILITAR  :  6   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
 (Other)          : 11
   EXADMISION         ESCALA
 Min.   :0.0000   Min.   :10.00
 1st Qu.:0.2390   1st Qu.:14.00
 Median :0.3260   Median :14.00
 Mean   :0.3674   Mean   :14.13
 3rd Qu.:0.4445   3rd Qu.:15.00
 Max.   :1.0000   Max.   :16.00


MODALIDAD
 CONCURSO ORDINARIO DE ADMISION   :500
 EXON.POR DEPORTISTA DESTACADO    :  1
 EXONERADO CENTRO PREUNIVERSITARIO: 45
 EXONERADO GRADO/TITULO PROF.     :  1
 EXONERADO POR CONVENIO DE EXCELENCIA: 15
 EXONERADO POR PRIMERA ALTERNATIVA   :151
 EXONERADO TERCIO SUPERIOR           : 12
                  DISTRITO       PROVINCIA.1    DEPARTAMENTO   DISTANCIA
 ATE               : 73   LIMA      :631   LIMA    :638   Min.   : 0.238
 SAN JUAN DE LURIGANCHO: 57  CALLAO    : 51   CALLAO  : 51   1st Qu.: 7.481
 SAN MARTIN DE PORRES : 51   OTROS     : 25   OTROS   : 25   Median :11.000
 LA MOLINA            : 44   CHINCHA   :  5   ICA     :  6   Mean   :11.410
 LIMA                 : 39   HUAROCHIRI:  3   ANCASH  :  1   3rd Qu.:15.740
 SANTA ANITA          : 36   HUARAL    :  2   APURIMAC:  1   Max.   :38.285
 (Other)              :425   (Other)   :  8   (Other) :  3
```

ANEXO N° 2 Scripts desarrollados en R

1. Regresión Logística

```
library(caTools)
set.seed(100)
split = sample.split(old$APROBADO, SplitRatio = 0.75)
oldTrain = subset(old, split=TRUE)
oldTest = subset(old, split==FALSE)
oldTrain = subset(old, split==TRUE)
oldLogP = glm(APROBADO ~ EXADMISION+EDAD+SEXO+DISTANCIA+COLEEXC, dat
a = oldTrain, family=binomial())
summary(oldLogP)
predictoldTrain = predict(oldLogP, type="response")
tapply(predictoldTrain, oldTrain$APROBADO, mean)
table(oldTrain$APROBADO, predictoldTrain > 0.2)

library(ROCR)

ROCRpredold = prediction(predictoldTrain, oldTrain$APROBADO)
ROCRperfold = performance(ROCRpredold, "tpr", "fpr")
plot(ROCRperfold)
plot(ROCRperfold, colorize=TRUE, print.cutoffs.at=seq(0,1,0.1), text
.adj=c(-0.2,1.7))
predictoldTest = predict(oldLogP, type = "response", newdata = oldTe
st)
```

2. Árbol de decisiones

```
library(C50)

ap_old_tree <- C5.0(old[,c(5,6,7,10,11,12,13,14,19)], trials = 10, o
ld$APROBADO2)
summary(ap_old_tree)

plot(ap_old_tree)
ap_oldtrain_tree_pred <- predict(ap_oldtrain_tree, oldTest)
CrossTable(oldTest$APROBADO, ap_oldtrain_tree_pred,
+            prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE,
+            dnn = c('REAL', 'PREDICCION'))
```

3. Support Vector Machines

```
library(kernlab)
ap_old_svm <- ksvm (APROBADO2 ~ EDAD+SEXO+COLEEXC+COLPRE+COLPRIV+EXA
DMISION+ESCALA+DISTANCIA, data = oldTrain, kernel = "vanilladot")
ap_old_svm_pred <- predict(ap_old_svm, oldTest)
table(ap_old_svm_pred, oldTest$APROBADO2)
agreement_old <- ap_old_svm_pred == oldTest$APROBADO2
prop.table(table(agreement_old))
```

ANEXO N° 3 Resultados Análisis de Varianza ANOVA

Plan de estudios 2010

```
> EDAD
             Df Sum Sq Mean Sq F value Pr(>F)
EDAD          1   0.42  0.4161   2.563   0.11
Residuals   723 117.37  0.1623

> DISTANCIA
             Df Sum Sq Mean Sq F value Pr(>F)
DISTANCIA     1   0.67  0.6689   4.129 0.0425 *
Residuals   723 117.12  0.1620
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> SEXO
             Df Sum Sq Mean Sq F value  Pr(>F)
SEXO          1   1.19  1.1917   7.389 0.00672 **
Residuals   723 116.60  0.1613
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> PROVINCIA

             Df Sum Sq Mean Sq F value Pr(>F)
PROVINCIA     1   0.06 0.05511   0.338  0.561
Residuals   723 117.73 0.16284

> TIPOCOLEGIO
             Df Sum Sq Mean Sq F value Pr(>F)
TIPOCOLEGIO   7   1.61  0.2297   1.418  0.195
Residuals   717 116.18  0.1620

> COLEEXC
             Df Sum Sq Mean Sq F value Pr(>F)
COLEEXC       1   0.35  0.3548   2.185   0.14
Residuals   723 117.43  0.1624

> COLPRE
             Df Sum Sq Mean Sq F value Pr(>F)
COLPRE        1   0.12  0.1169   0.718  0.397
Residuals   723 117.67  0.1628

> COLPRIV
             Df Sum Sq Mean Sq F value Pr(>F)
COLPRIV       1   0.07 0.07132   0.438  0.508
Residuals   723 117.72 0.16282
```

```
> EXADMISION
             Df Sum Sq Mean Sq F value   Pr(>F)
EXADMISION    1   5.52   5.521   35.56 3.87e-09 ***
Residuals   723 112.27   0.155
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ESCALA

```
             Df Sum Sq Mean Sq F value Pr(>F)
ESCALA        1   0.01 0.00944   0.058   0.81
Residuals   723 117.78 0.16290
```

```
> MODALIDAD
             Df Sum Sq Mean Sq F value  Pr(>F)
MODALIDAD     6   2.91  0.4842   3.026 0.00628 **
Residuals   718 114.88  0.1600
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> GEOMETRIAANALITICA
                   Df Sum Sq Mean Sq F value Pr(>F)
GEOMETRIAANALITICA   1  59.36   59.36   734.5 <2e-16 ***
Residuals          723  58.43    0.08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> MATEMATICADISCRETA
                   Df Sum Sq Mean Sq F value Pr(>F)
MATEMATICADISCRETA   1  49.10    49.1   516.7 <2e-16 ***
Residuals          723  68.69     0.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> FILOSOFIA
             Df Sum Sq Mean Sq F value Pr(>F)
FILOSOFIA     1  20.27  20.269   150.3 <2e-16 ***
Residuals   723  97.52   0.135
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
APROBADO ~ DISTANCIA*SEXO*EXADMISION*MODALIDAD
                                Df Sum Sq Mean Sq F value   Pr(>F)
DISTANCIA                        1   0.67   0.669   4.955  0.02635 *
SEXO                             1   1.22   1.218   9.022  0.00277 **
EXADMISION                       1   5.23   5.231  38.745 8.41e-10 ***
MODALIDAD                        6   4.64   0.773   5.726 7.97e-06 ***
DISTANCIA:SEXO                   1   0.09   0.090   0.667  0.41430
DISTANCIA:EXADMISION             1   0.21   0.214   1.582  0.20892
```

```
SEXO:EXADMISION                         1   0.12   0.120   0.891  0.34564
DISTANCIA:MODALIDAD                     4   1.19   0.297   2.196  0.06787 .
SEXO:MODALIDAD                          4   1.94   0.485   3.590  0.00658 **
EXADMISION:MODALIDAD                    4   7.56   1.891  14.008 5.42e-11 ***
DISTANCIA:SEXO:EXADMISION               1   0.19   0.192   1.424  0.23313
DISTANCIA:SEXO:MODALIDAD                3   0.61   0.204   1.510  0.21078
DISTANCIA:EXADMISION:MODALIDAD          4   0.65   0.163   1.206  0.30676
SEXO:EXADMISION:MODALIDAD               3   0.43   0.142   1.054  0.36792
DISTANCIA:SEXO:EXADMISION:MODALIDAD     3   0.42   0.142   1.049  0.37038
Residuals                             686  92.61   0.135
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Plan de estudios 2014

```
> EDAD
             Df Sum Sq Mean Sq F value Pr(>F)
EDAD          1   0.09 0.08728   0.401  0.527
Residuals   273  59.39 0.21754

> SEXO
             Df Sum Sq Mean Sq F value Pr(>F)
SEXO          1   0.05 0.05233    0.24  0.624
Residuals   273  59.42 0.21767

> PROVINCIA
             Df Sum Sq Mean Sq F value Pr(>F)
PROVINCIA     1   0.51  0.5051   2.339  0.127
Residuals   273  58.97  0.2160

> TIPOCOLEGIO
             Df Sum Sq Mean Sq F value Pr(>F)
TIPOCOLEGIO   7   1.87  0.2670   1.237  0.282
Residuals   267  57.61  0.2158

> COLEEXC
             Df Sum Sq Mean Sq F value Pr(>F)
COLEEXC       1   0.76  0.7584   3.526 0.0615 .
Residuals   273  58.72  0.2151
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> COLPRE
             Df Sum Sq Mean Sq F value Pr(>F)
COLPRE        1   0.66  0.6647   3.085 0.0801 .
Residuals   273  58.81  0.2154
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


> COLPRIV
             Df Sum Sq Mean Sq F value Pr(>F)
```

```
COLPRIV       1   0.05 0.05419   0.248  0.619
Residuals   272  59.32 0.21809
1 observation deleted due to missingness

> EXADMISION
             Df Sum Sq Mean Sq F value   Pr(>F)
EXADMISION    1   3.39   3.392   16.51 6.33e-05 ***
Residuals   273  56.08   0.205
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> ESCALA
             Df Sum Sq Mean Sq F value Pr(>F)
ESCALA        1   0.80   0.795   3.699 0.0555 .
Residuals   273  58.68   0.215
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> MODALIDAD
             Df Sum Sq Mean Sq F value Pr(>F)
MODALIDAD     5   2.41  0.4821   2.273 0.0477 *
Residuals   269  57.07  0.2121
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> DISTANCIA
             Df Sum Sq Mean Sq F value Pr(>F)
DISTANCIA     1   0.03 0.03084   0.142  0.707
Residuals   273  59.45 0.21775

> anova1 = aov(APROBADO ~ MODALIDAD*EXADMISION*COLPRE*COLEEXC, data=
new)
> summary(anova1)
                                    Df Sum Sq Mean Sq F value   Pr(>F)
MODALIDAD                            5   2.41   0.482   2.540  0.02902 *
EXADMISION                           1   3.60   3.599  18.961 1.96e-05 ***
COLPRE                               1   0.46   0.461   2.426  0.12060
COLEEXC                              1   0.58   0.581   3.060  0.08149 .
MODALIDAD:EXADMISION                 4   2.82   0.705   3.717  0.00588 **
MODALIDAD:COLPRE                     3   1.21   0.402   2.118  0.09857 .
EXADMISION:COLPRE                    1   0.27   0.271   1.430  0.23288
MODALIDAD:COLEEXC                    3   0.17   0.057   0.302  0.82386
EXADMISION:COLEEXC                   1   0.00   0.003   0.014  0.90658
COLPRE:COLEEXC                       1   0.07   0.066   0.349  0.55507
MODALIDAD:EXADMISION:COLPRE          3   0.61   0.202   1.063  0.36547
MODALIDAD:EXADMISION:COLEEXC         2   0.24   0.121   0.635  0.53081
MODALIDAD:COLPRE:COLEEXC             1   0.23   0.229   1.209  0.27264
EXADMISION:COLPRE:COLEEXC            1   0.30   0.301   1.586  0.20907
MODALIDAD:EXADMISION:COLPRE:COLEEXC  1   0.01   0.010   0.051  0.82151
Residuals                          245  46.50   0.190
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```