ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

"IDENTIFICACIÓN DE LOS FACTORES DE DESERCIÓN DE UNA CARRERA UNIVERSITARIA EN UNA INSTITUCIÓN DE EDUCACIÓN SUPERIOR, USANDO TÉCNICAS DE MINERÍA DE DATOS"

TRABAJO DE TITULACIÓN

PREVIO A LA OBTENCIÓN DEL TÍTULO DE:

MAGISTER EN SISTEMAS DE INFORMACIÓN GERENCIAL

Autores:

ORDÓÑEZ URGILÉS MILTON XAVIER

NOBOA CISNEROS CÉSAR EMILIANO

Guayaquil – Ecuador

2018

AGRADECIMIENTOS

A Dios, por darme el regalo de la vida y permitirme seguir cumpliendo con mis objetivos propuestos.

A toda mi familia, por ser el soporte necesario y motivo de superación personal.

A mi director de tesis M.Sc. Jorge Magallanes, por su asesoría en la elaboración de este trabajo.

A mis compañeros de trabajo GTSI - ESPOL, por estar siempre prestos a colaborar con la información y asistencia necesaria para el desarrollo de este trabajo.

A la Escuela Superior Politécnica del Litoral, por darme su aval para la realización de la Maestría.

A todos mis profesores en el trascurso de la Maestría.

Milton Xavier Ordóñez Urgilés

A DIOS Padre, Soberano Señor de quien procede el amor y la sabiduría, pero también la justicia. Imposible condensar en estas líneas la gratitud que yo le debo, como imposible es que lo ínfimo alcance lo supremo. Imposible enumerar todos sus favores: el aire, la luz del sol, el arco iris tras la lluvia, el trabajo, el alimento, la salud, los estudios, la vida de mi madre, el perdón de mis pecados, una lista sin fin.

Al bendito Hijo de DIOS Jesucristo, por medio de quien DIOS me ha ofrecido perdón por mis faltas y salvación de la muerte. Gracias a Jesús, porque muchas veces cambió mi desesperación en inmensa alegría. La culminación de esta carrera es una de las innumerables razones para volver a decir: ¡Gracias, gracias Señor!

A mis padres, por haberme ayudado a lo largo de mi vida.

A mis familiares, por todos sus gestos amables.

A mi compañero Milton Ordóñez y a nuestro director Jorge Magallanes, por haber participado conmigo de este proyecto.

César Emiliano Noboa Cisneros

DEDICATORIAS

Dedico este trabajo a mi amada esposa Annabell.

A mis hijas Valeria, Gabriela y Gianella que son mis tesoros.

A mis padres Milton y Gloria, mis guías y ejemplos de vida.

Milton Xavier Ordóñez Urgilés

A Jesucristo, mi salvador.

A mi madre, quien es para mí un hermoso regalo de DIOS.

César Emiliano Noboa Cisneros

TRIBUNAL DE SUSTENTACIÓN

Mgs. Lenin Freire C.

DIRECTOR MSIG

M.Sc. Jorge Magallanes B.

DIRECTOR DEL PROYECTO

DE GRADUACIÓN

Mgs. Omar Maldonado D.

MIEMBRO DEL TRIBUNAL

DECLARACIÓN EXPRESA

"La responsabilidad del contenido de este Trabajo de Titulación, nos corresponde exclusivamente; y el patrimonio intelectual de la misma a la ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL".

(Reglamento de Graduación de ESPOL)

……..
Milton Xavier Ordóñez Urgilés

……..
César Emiliano Noboa Cisneros

RESUMEN

En el presente trabajo se propone identificar los factores que inciden en la deserción universitaria por medio del análisis de los datos contenidos, principalmente en la base del Sistema Académico. Para ello, se lleva a cabo un análisis discriminante con el uso de una técnica estadística convencional conocida como Regresión Logística y 3 técnicas de minería de datos que son: Árboles de decisión, k-vecinos más cercanos y Naive Bayes. Puesto que la base teórica de las 4 técnicas aplicadas es diferente, sus resultados no son los mismos, pero sí complementarios.

La minería de datos no sólo permite aumentar la comprensión de los datos explicando fenómenos que éstos esconden; sino también permite estimar con cierto nivel de confianza futuras tendencias. En este sentido, este trabajo también lleva a cabo un análisis experimental de los resultados obtenidos por las técnicas estadísticas tanto de manera individual como en conjunto. Este análisis permite estimar la capacidad del proyecto completo para detectar potenciales desertores; así como estimar cuántos estudiantes actualmente registrados son potenciales desertores.

En el Capítulo 2 de este trabajo se exponen los aspectos teóricos tanto del análisis discriminante de forma general como de las técnicas estadísticas que llevan a cabo la tarea de discriminación. Si bien no se profundiza en la teoría más allá de lo requerido, el Capítulo 2 es importante para comprender las fases del proceso de minería de datos que se sigue en los capítulos posteriores.

El Capítulo 3 es el inicio del proyecto como tal. En este capítulo se expone el levantamiento de la información y el análisis exploratorio preliminar de los datos disponibles. Se realizan conjeturas iniciales que luego deberán ser corroboradas o refutadas por los resultados finales. En este capítulo, es especialmente importante la selección del conjunto de datos objetivo; en este apartado se explica cuál es ese conjunto y algunas de sus características.

En el Capítulo 4 se presentan las variables seleccionadas para el análisis discriminante, se da un detalle pormenorizado de la forma de obtener estas variables y su significado. Ya al final del capítulo se expone brevemente el modelo general que se va a utilizar y el significado de sus resultados.

El Capítulo 5 es el inicio de la experimentación. Este capítulo presenta los resultados obtenidos por los 4 métodos en todos los conjuntos de entrenamiento, así como la validación de la mayoría de los resultados por medio de la técnica de validación cruzada. Con los modelos ya construidos y entrenados, en este capítulo ya se está en capacidad de detectar los posibles desertores dentro del conjunto actual de estudiantes; se presentan estos resultados como antesala a las conclusiones finales del proyecto.

Finalmente, el Capítulo 6 presenta los resultados obtenidos por el proyecto en los conjuntos de prueba previamente seleccionados. Es en este capítulo donde se estima la capacidad total del proyecto de minería de datos para identificar los potenciales desertores; capacidad que podría incrementarse considerablemente con la implementación de las recomendaciones propuestas al final de este trabajo.

ÍNDICE GENERAL

ABREVIATURAS Y SIMBOLOGÍA

CEAACES: Consejo de Evaluación, Acreditación y Aseguramiento de la Calidad de la Educación Superior en el Ecuador.

CRISP-DM: Cross Industry Standard Process for Data Mining (Proceso estándar de la industria para la minería de datos). Se trata de un modelo o metodología de procesos de minería de datos que describe los enfoques comunes que utilizan los expertos en el área.

KDD: Knowledge Discovery in Databases (Descubrimiento de conocimientos en bases de datos). Metodología general y secuencial que se sigue para descubrir conocimiento en un conjunto de datos en bruto.

KNN: K-nearest neighbors (K vecinos más cercanos). Es un método de clasificación supervisada en la minería de datos que estima la clase a la que pertenece un elemento conociendo sus variables predictoras.

SAAC: Sistema Administrativo Académico de la ESPOL.

SEMMA: Sample, Explore, Modify, Model, Assess (Muestra, Explora, Modifica, Modela, Evalúa). Metodología desarrollada por la empresa SAS para procesos de minería de datos.

SIB: Sistema de Información Bibliotecario de la ESPOL.

TIOBE: The Importance of Being Earnest (La importancia de ser serio). Empresa dedicada a la medición de parámetros de software.

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

Es un hecho casi irrefutable que la deserción universitaria tiene efectos negativos no sólo a nivel cultural sino también a nivel económico, efectos negativos no sólo a nivel individual sino a nivel de toda la sociedad en su conjunto.

En concreto, un estudiante no graduado no será fácilmente insertado en el campo laboral y aunque esto ocurriera, el ingreso percibido por éste sería mucho menor que el obtenido por sus competidores ya graduados. A causa de este bajo ingreso, el Estado recaudaría menos impuestos y ofrecería menos servicios, desatándose así un círculo vicioso sin duda nocivo para un país.

Si la deserción provoca millonarias pérdidas en países donde la educación universitaria no es gratuita, en Ecuador, donde se adoptó la gratuidad de la educación universitaria a partir del año 2009, los efectos negativos de esta deserción se magnifican.

Las universidades públicas del Ecuador tienen como gran reto reducir el nivel de deserción de sus estudiantes. En particular, la Escuela Superior Politécnica del Litoral, una de las 3 universidades con categoría A en el Ecuador, está llamada a ser pionera en esta iniciativa.

El presente trabajo es presentado como una contribución a la loable iniciativa de reducir el nivel de deserción en la población estudiantil. Este trabajo no está en la línea de la sicología educativa o en la línea del análisis organizacional, sino que su enfoque e intención principal es reflejar lo que nos revelan los datos y registros informáticos acerca de la deserción. Con este trabajo, se propone analizar las bases de datos disponibles para analizar la situación actual y proyectar tendencias. Evidentemente, la deserción universitaria va más allá de simples datos y estimaciones; pero es siempre necesario incluir en el análisis integral el componente objetivo que los datos proporcionan.

CAPÍTULO 1

GENERALIDADES

1.1 ANTECEDENTES

La Escuela Superior Politécnica del Litoral es una institución de educación superior, líder a nivel nacional y actualmente ubicada en la categoría “A” de acuerdo al CEAACES (Consejo de Evaluación, Acreditación y Aseguramiento de la Calidad de la Educación Superior en el Ecuador). Esta institución en la actualidad cuenta con alrededor de 9,600 estudiantes registrados y una planta docente de 660 profesores.

La búsqueda constante de la calidad y excelencia en sus procesos, conlleva la mejora continua en un área clave como son las tecnologías y los sistemas de información. Actualmente se cuenta con un sistema de

gestión de indicadores de calidad, donde periódicamente se evalúan indicadores claves como: tasa de titulación, tasa de retención, tasa de deserción, entre otros. El manejo y control de estos indicadores es de suma importancia para los continuos procesos de auditorías y acreditación a los que está sometido la institución.

Sin embargo, un tema clave de los indicadores antes mencionados es la tasa de deserción estudiantil. Con relación a este indicador sería muy importante no sólo llevar un control del mismo, sino también aplicar técnicas modernas de computación; para poder anticiparse a las posibles deserciones estudiantiles y tomar los correctivos necesarios antes de que ocurran.

Después de un análisis de la situación, las técnicas de minería de datos se perfilan como la mejor opción para solventar este problema en particular.

1.2 DESCRIPCIÓN DEL PROBLEMA

De acuerdo a un reporte con fecha noviembre 2016 del Diario El Telégrafo [1], los índices de deserción universitaria en el Ecuador son altos. Esta deserción ha sido un problema ampliamente discutido y conlleva una serie de efectos negativos tales como la disminución de profesionales calificados, desperdicio de recursos tanto para el sector

público como el privado, desmotivación de la potencial fuerza laboral entre otros.

De acuerdo a la base de datos del Sistema Académico de la Escuela Superior Politécnica, de los estudiantes que ingresaron en el primer semestre del año 2009, el 27.56% abandonó la universidad. Aunque esta tasa ha decrecido considerablemente (14.80% para los estudiantes que ingresaron en 2012-2S) sigue implicando efectos negativos que deben ser contrarrestados por medio de una gestión adecuada de las universidades.

Sin duda la gratuidad de la educación universitaria impone una mayor presión sobre los efectos de una deserción, especialmente si ésta ocurre en las etapas finales de la carrera universitaria.

1.3 SOLUCIÓN PROPUESTA

La identificación de los factores que inciden en la deserción universitaria es clave para la toma de una serie de decisiones que produzcan una reducción de este problema.

Las instituciones cuentan en su mayoría con un sistema académico que recolecta tanto las características demográficas, socioeconómicas y académicas de los estudiantes antes y durante su recorrido en las diferentes carreras. Sin embargo, estos datos no son lo suficientemente

analizados para producir un conocimiento que desemboque en decisiones adecuadas.

El presente trabajo de titulación propone la implementación de técnicas de minería de datos a la base SAAC (Sistema Administrativo Académico de la ESPOL) para generar un modelo que clasifique al conjunto de estudiantes de acuerdo al riesgo de deserción. La detección temprana de un estudiante con alto riesgo de deserción permitirá a las autoridades y consejeros académicos aplicar medidas correctivas para evitar que los estudiantes deserten. Se debe considerar también que una disminución de la tasa de deserción mejorará indirectamente no sólo los índices de evaluación de la universidad (tales como la tasa de titulación, tasa de retención, etc.) sino también su imagen institucional.

Se propone inicialmente realizar una selección de las características relevantes. De acuerdo a Zlatko Kovacic [2], los factores preponderantes pueden estar relacionados tanto a las variables socio-demográficas del estudiante como a su ambiente académico. Por ejemplo, en el caso de la ESPOL, un somero análisis de la cantidad de desertores por género revela que los hombres son más propensos a desertar (Tabla 1) y que el 75% de estas deserciones ocurren aproximadamente en los 3 primeros semestres de la carrera (Figura 1.1 y Figura 1.2).

Tabla 1. Cantidad de desertores por género desde el año 2009

	Desertores		No Desertores		Todos	
Género	Cantidad	%	Cantidad	%	Cantidad	%
F	568	31.20	5,343	41.80	5,911	40.50
M	1,252	68.80	7,434	58.20	8,686	59.50
Total	1,820	100.00	12,777	100.00	14,597	100.00

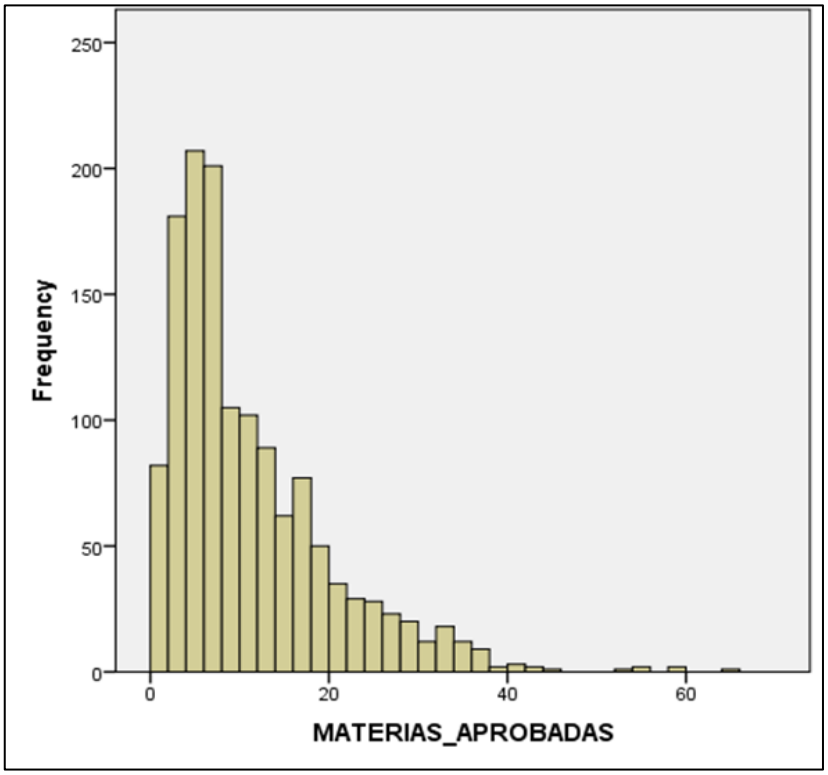


Figura 1.1. Histograma de la cantidad de materias aprobadas por los desertores

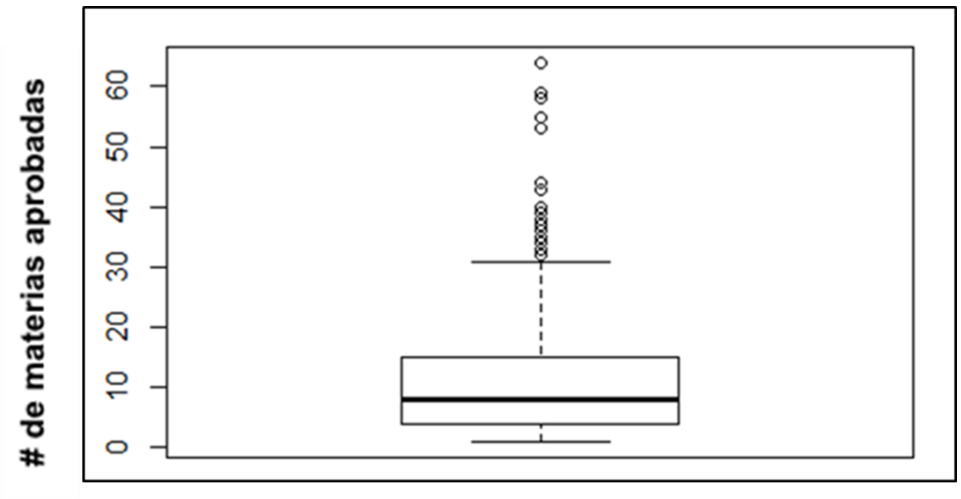


Figura 1.2. Cantidad de materias aprobadas por los desertores (diagrama de caja)

Por ello se propone realizar un análisis de los datos contenidos en la encuesta socio-económica de los estudiantes, de sus características académicas previas al enrolamiento en la universidad y de su record académico en la misma universidad.

Sin perjuicio de escoger otras variables que resulten más preponderantes durante la construcción de los modelos, dentro de las variables que inicialmente se analizarán para definir su factor de incidencia en la deserción estudiantil se encuentran: el nivel socioeconómico, la etnia del estudiante, el género del estudiante, el rendimiento académico promedio, el número de materias reprobadas y el número del semestre en el que se encuentra el estudiante. La selección de variables y, por ende, los modelos que se construyan a partir de ellas, dependerán de la cantidad

y calidad de los datos que mantenga el Sistema Académico. Sin duda, luego de este proyecto surgirán valiosas recomendaciones respecto a futuras recolecciones de datos que deberán hacerse para mejorar los modelos de discriminación.

Luego de realizar la selección de las variables, se propone aplicar métodos de Minería de Datos tanto retardados (lazy) como anticipados para generar un modelo que permita, si es posible, predecir el riesgo de deserción del estudiante. Los métodos que se emplearán para la clasificación son: K-vecinos más cercanos, Naive Bayes, Árboles de Decisión y Regresión Logística. Los modelos se construirán en base al conjunto de entrenamiento, que usualmente consiste en los 2/3 del conjunto total de datos y la precisión de la mayoría de estos modelos será principalmente evaluada mediante validación cruzada (cross-validation).

En esta propuesta se presentarán reportes relacionados a los resultados obtenidos por cada método aplicado, así como el análisis comparativo correspondiente.

Luego de analizar los resultados de cada método, se realizará una estimación de la capacidad del proyecto para detectar posibles desertores. Con el fin de tomar acciones preventivas para evitar la deserción, también se mostrarán las predicciones del proyecto acerca de los potenciales desertores dentro del conjunto actual de estudiantes,

haciendo énfasis en la información relacionada a los estudiantes que hayan sido detectados por más de un método.

1.4 OBJETIVO GENERAL

Implementar técnicas de Minería de Datos para la identificación de los factores de deserción de una carrera universitaria en una Institución de Educación Superior.

1.5 OBJETIVOS ESPECÍFICOS

- Investigar la forma de implementar un proyecto de Minería de Datos para el análisis propuesto.
- Realizar un levantamiento y pre-procesamiento de información respecto a los datos contenidos en la base de datos SAAC.
- Diseñar los modelos mediante la selección de las características relevantes y la definición de los requisitos preliminares asociados a las técnicas de Minería de Datos a utilizar.
- Implementar las técnicas de Minería de Datos para el entrenamiento, validación y prueba de los modelos.
- Realizar un análisis individual y comparativo de los resultados obtenidos por los diferentes modelos.

1.6 METODOLOGÍA

La metodología guía para el desarrollo de este proyecto es CRISP-DM (Cross Industry Standard Process for Data Mining), que es una de las metodologías para proyectos de minería de datos más utilizadas por la comunidad inherente al área [3]. De esta metodología se hablará de forma más amplia en el Capítulo 2 (marco teórico).

Las etapas o fases seguidas para el desarrollo del presente trabajo se muestran en la Figura 1.3.

Figura 1.3. Etapas de la metodología del proyecto de minería de datos

En el análisis exploratorio preliminar se realizan consultas a las bases de datos del sistema académico. Se examina el volumen de los datos, sus características y se realizan pruebas estadísticas básicas que permitan descubrir información oculta a simple vista.

En la selección del conjunto de datos objetivo se elige un subconjunto de los datos explorados y obtenidos previamente, con este subconjunto se continuará el proceso de minería a las siguientes fases. Para realizar esta selección se deben tomar en cuenta criterios como calidad, volumen y características de los datos de acuerdo a las técnicas de minería que se vayan a utilizar.

La limpieza y preparación del conjunto de datos se refiere en primer lugar a examinar los datos del conjunto objetivo, con el fin de encontrar errores, incoherencias o datos fuera de rango que puedan afectar las fases posteriores. Luego se deben implementar técnicas para mejorar la calidad de los datos, solventar incoherencias o errores en los mismos.

A continuación, se realiza la selección de las variables a utilizar en los modelos de minería de datos, tomando en cuenta factores como disponibilidad y confiabilidad de los datos, análisis estadístico preliminar y otras técnicas de selección.

El pre-procesamiento de los datos se refiere a las operaciones que se deben realizar a los datos originales para obtener las variables finales seleccionadas en la fase previa. Se debe tomar en cuenta el formato requerido por las técnicas de minería a utilizar.

Para el entrenamiento de los modelos se separa el conjunto de datos objetivo en una parte para entrenamiento y la otra para prueba. Las

técnicas de minería son aplicadas al conjunto de entrenamiento con el fin de generar modelos que permitan predecir la probabilidad de deserción estudiantil y los factores que la causan.

A continuación, se deben evaluar los modelos generados en la fase anterior con el conjunto de prueba. Para cada modelo se calcula el porcentaje de detección de desertores y el porcentaje de error al clasificar.

Finalmente se muestran y analizan los resultados obtenidos por cada método y el resultado integral del proyecto, es decir de todos los métodos en conjunto.

CAPÍTULO 2

MARCO TEÓRICO

2.1 CONCEPTOS BÁSICOS DE MINERÍA DE DATOS

La minería de datos es el proceso que tiene como objetivo descubrir y extraer conocimiento (información relevante) de grandes volúmenes de datos. Estos datos podrían provenir de grandes bases transaccionales de sistemas de información. La información extraída deberá ser estructurada de forma comprensible para aprovecharla posteriormente.

Se puede decir que la minería de datos está formada por la unión de varias disciplinas como: estadística, inteligencia artificial, aprendizaje automático, reconocimiento de patrones, sistemas de base de datos, visualización de información, entre otras. A modo de resumen, en la

Figura 2.1 se presentan algunas de las disciplinas que conforman el campo de la minería de datos.

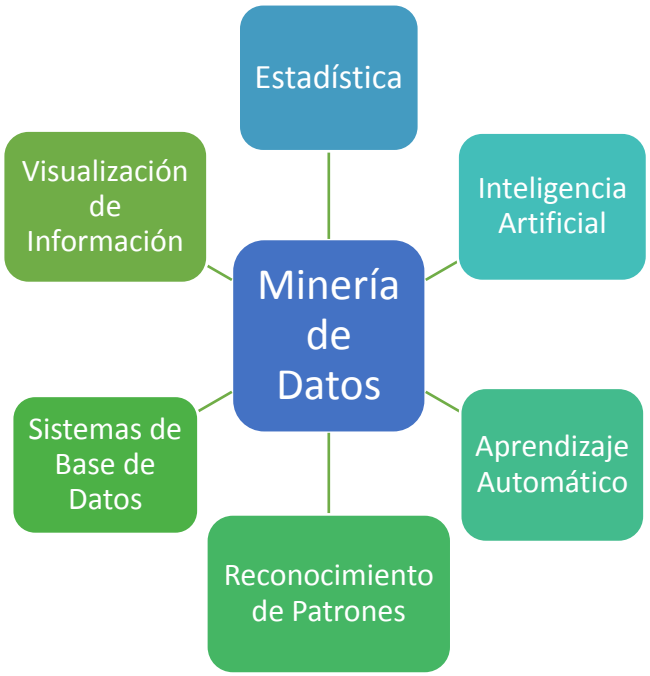


Figura 2.1. Disciplinas que conforman la minería de datos

2.1.1 Clasificación de los métodos según la tarea que realizan

Dependiendo del tipo de tarea que se quiera llevar a cabo, se pueden seleccionar uno o varios métodos de minería de datos que se adecuen al problema propuesto. De acuerdo a la tarea que realizan los métodos o técnicas de minería de datos se clasifican en técnicas de predicción, agrupamiento y reglas de asociación.

Técnicas de predicción

Estas técnicas tratan de predecir el valor de un atributo o variable objetivo (variable dependiente) conociendo los valores de las variables de entrada (variables independientes). Estas técnicas dan como resultado relaciones entre las variables, las cuales se representan como un modelo. Algunas aplicaciones de este tipo de técnicas son: detección de fraudes aduaneros, predicción de rendimientos financieros, análisis de riesgo en la entrega de créditos, detección de spam, entre otros [4].

Las técnicas de predicción también se las llama de aprendizaje supervisado y pueden ser de dos tipos: modelos de clasificación y modelos de regresión.

En este punto vale mencionar que las técnicas de minería de datos a implementarse en el presente trabajo pertenecen a modelos de clasificación o también conocidos como modelos de discriminación.

Técnicas de agrupamiento

Estas técnicas agrupan datos dentro de un número de clases. La agrupación se la realiza generalmente partiendo de criterios de distancia o similitud, de manera que exista alta similitud entre los

elementos de un mismo grupo y baja similitud entre los elementos de distinto grupo. Algunas aplicaciones de este tipo de técnicas son: perfilamiento de usuarios, identificación de enfermedades, reducción de dimensiones, reconocimiento de patrones, etc. Puesto que no siempre son conocidas las clases o grupos a determinar estas técnicas se llaman de aprendizaje no supervisado [4].

Técnicas de reglas de asociación

Permiten descubrir posibles relaciones o correlaciones entre distintos sucesos o acciones aparentemente independientes. Dichas relaciones expresan un antecedente y un consecuente, es decir la ocurrencia de un suceso o acción desencadena la aparición de otros. Algunas aplicaciones de este tipo de técnicas son: análisis de canasta de compras, predicción y correlación de variables financieras, relaciones entre factores ambientales y climáticos, entre otros [5].

2.1.2 Clasificación de los métodos según el modo de procesamiento

La minería de datos está relacionada al aprendizaje ya sea supervisado o no supervisado. Estos métodos de aprendizaje construyen sus modelos en base a un conjunto de

entrenamiento, y el aprendizaje obtenido es puesto a prueba con un nuevo conjunto que no participó del entrenamiento del método, a este conjunto se lo conoce como el conjunto de prueba.

Desde el punto de vista del entrenamiento y prueba, los métodos de minería de datos se clasifican en anticipados y retardados.

Los métodos anticipados son aquellos que construyen sus modelos previamente y responden casi instantáneamente cuando son evaluados. Estos métodos emplean la mayoría del tiempo en entrenar, luego de lo cual generan modelos explícitos y comprensibles. Ejemplos de estos métodos son la regresión lineal y regresión logística.

Los métodos retardados son aquellos que no construyen sus modelos con anticipación, sino que lo hacen en el momento en que son evaluados. El entrenamiento y evaluación de estos métodos son llevados a cabo cuando son consultados. Estos métodos generan modelos implícitos y normalmente incomprensibles. Ejemplos de estos métodos son los k-vecinos más cercanos y Naive Bayes [6].

2.2 ANÁLISIS DISCRIMINANTE MEDIANTE TÉCNICAS DE MINERÍA DE DATOS

De forma general, el análisis discriminante trata de predecir la clase a la que pertenece un elemento desconocido, sabiendo las clases a las que pertenecen un grupo conocido de elementos. El grupo de elementos conocidos constituyen el conjunto de entrenamiento, el análisis discriminante basa sus predicciones en este conjunto.

La Figura 2.2 ilustra la idea general del análisis discriminante, en este ejemplo se pretende determinar el color de elemento desconocido, sabiendo los colores de los otros elementos. En este caso, la distancia puede ser un buen criterio para determinar la clase (color) a la que pertenece el nuevo elemento.

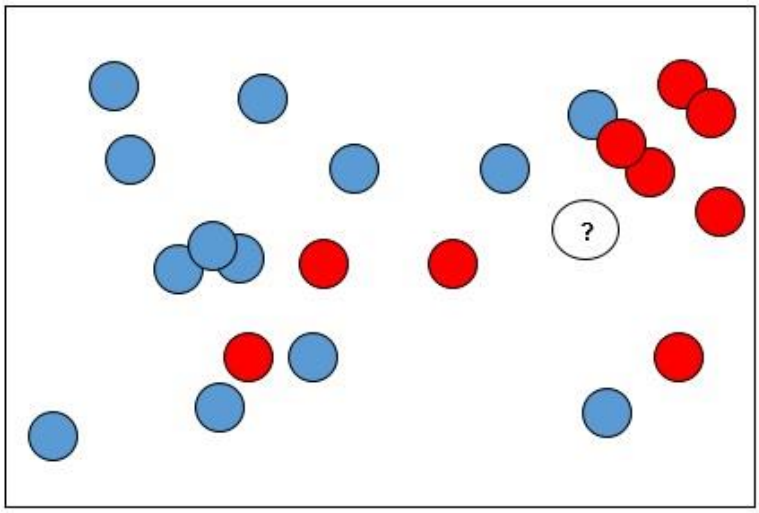


Figura 2.2. Ilustración del análisis discriminante

La capacidad de los métodos de discriminación es evaluada mediante el error de entrenamiento y el error de prueba. El error de entrenamiento es el porcentaje de malas clasificaciones que el método produce en el conjunto de entrenamiento; mientras que el error de prueba es el porcentaje de malas clasificaciones en el conjunto de prueba.

En la Figura 2.3, se muestran 2 ejemplos de discriminación en un conjunto de entrenamiento. La curva discriminante lineal (en verde) produce 5 malas clasificaciones, por tanto, el error en este caso es 5/22 = 22.72%. En cambio, la curva discriminante no lineal (en morado) produce un error menor de 3/22 = 13.63%.

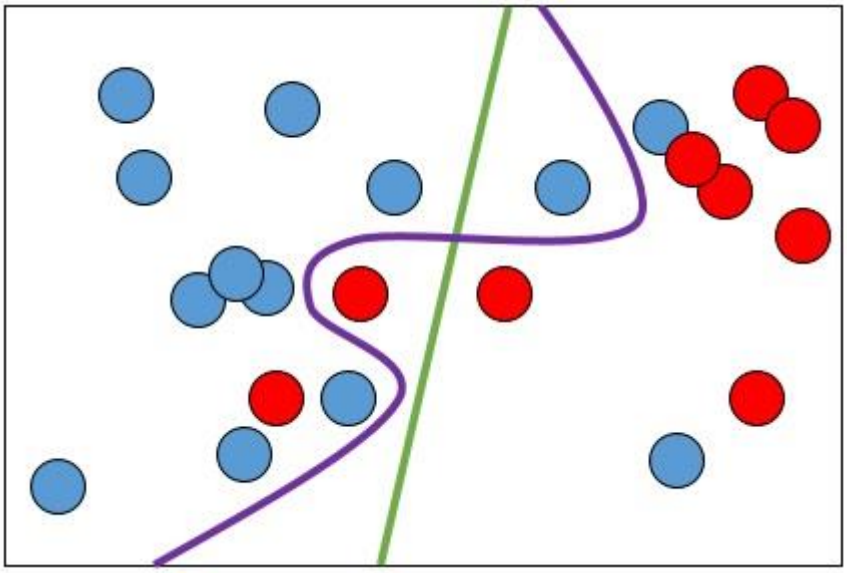


Figura 2.3. Ejemplo de curvas discriminantes

A continuación, se revisará los métodos o técnicas de minería de datos que se aplicarán en el presente trabajo, las cuales son: Árboles de decisión, K-vecinos más cercanos, Naive Bayes y Regresión logística.

2.2.1 Árboles de decisión

El Árbol de Decisión es una técnica que realiza particiones sucesivas del conjunto de entrenamiento de tal manera que cada grupo que se genere luego de cada partición sea más puro que su antecesor. En el caso en que sólo existan 2 clases, la pureza de un conjunto mide el nivel de dominio de una clase sobre la otra.

En la Figura 2.4 se ilustra el procedimiento que sigue esta técnica. Se observa que cada partición produce conjuntos donde un color domina respecto al otro. El proceso concluye hasta que las hojas del árbol son totalmente puras o la cantidad de elementos de las hojas es menor que un valor preestablecido. En el caso de la figura sólo una de las hojas es totalmente pura; de izquierda a derecha los niveles de pureza de las hojas son: 4/7= 57.14%, 9/10 = 90%, 6/8 = 75% y 9/9 = 100%.

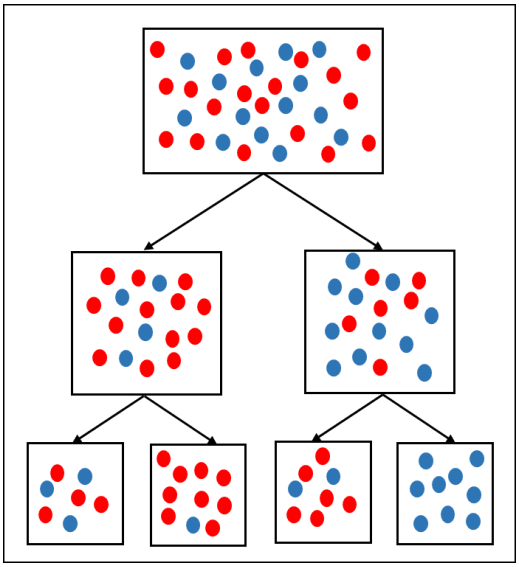


Figura 2.4. Ejemplo de discriminación utilizando árboles de decisión

Si se supone que el árbol de la Figura 2.4 tiene elementos con dos atributos numéricos continuos X y Y, un ejemplo de las particiones en el plano bidimensional se muestra en la Figura 2.5.

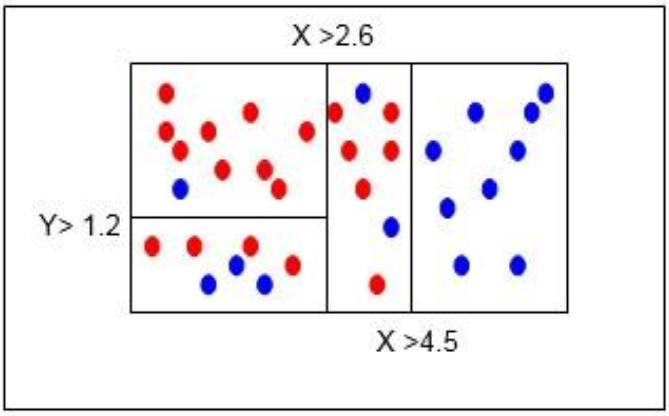


Figura 2.5. Resultado de una discriminación utilizando un árbol de decisión en el plano bidimensional

Considerando las tres particiones de la Figura 2.5, el árbol de la Figura 2.4 etiquetado quedaría como se muestra en la Figura 2.6.

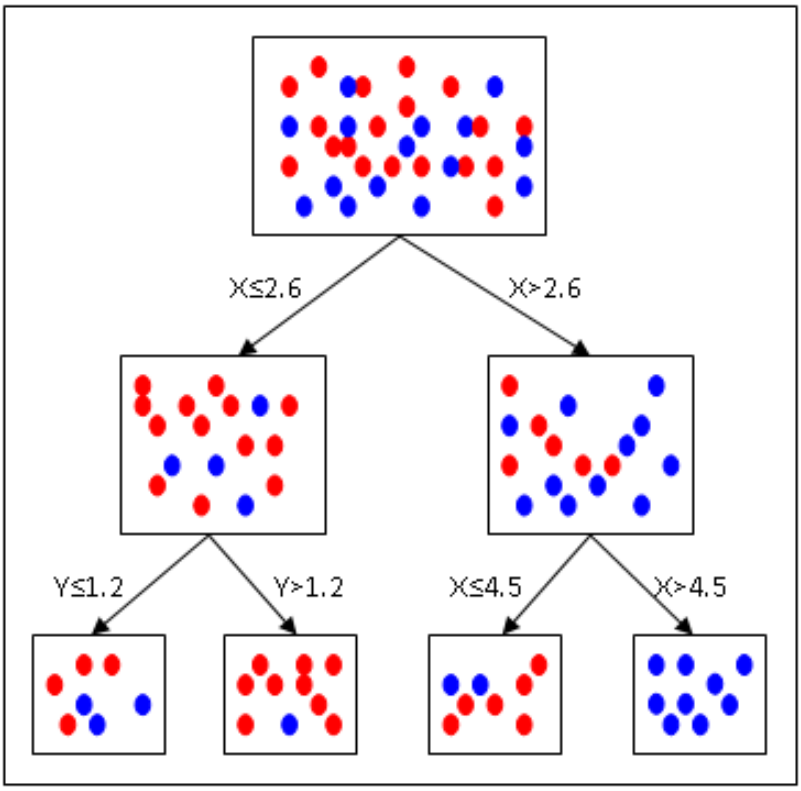


Figura 2.6. Esquema de un árbol de decisión con atributos numéricos

Luego del entrenamiento, la técnica de árbol de decisión produce reglas de discriminación que pueden aplicarse para clasificar un nuevo elemento. Cada hoja del árbol representa una regla. Cuando las hojas no son totalmente puras, las reglas vienen acompañadas de una probabilidad que mide la certeza de la

decisión. De esta manera, las reglas finales que produce el árbol de la Figura 2.6 son:

- Si X>4.5 entonces el elemento es de la clase “azul”.
- Si 2.6<X≤4.5 entonces el elemento es de la clase “roja” con una probabilidad de 6/8 = 0.75.
- Si X≤2.6 y Y≤1.2 entonces el elemento es de la clase “roja” con una probabilidad de 4/7 = 0.57.
- Si X≤2.6 y Y>1.2 entonces el elemento es de la clase “roja” con una probabilidad de 9/10 = 0.9.

De manera general, el algoritmo de árbol de decisión realiza una búsqueda exhaustiva de la variable que se utilizará para la partición y el valor en el cual se realizará la partición. Esta decisión se toma en base a la partición que produzca los conjuntos más puros posibles [7].

2.2.2 K-vecinos más cercanos

El método de los k-vecinos más cercanos o knn (k-nearest neighbors), donde k es un número entero positivo, consiste en predecir la clase de un nuevo elemento en base a las clases de sus k vecinos más cercanos. La clase asignada al nuevo

elemento será la clase dominante de los k vecinos. En el caso en el que sólo existan 2 clases, la clase asignada será aquella que tengan la mayoría de los vecinos.

La solución al problema propuesto en la Figura 2.2, empleando el método de los 7-vecinos más cercanos, se la muestra en la Figura 2.7. El nuevo elemento según este método es de clase “roja”.

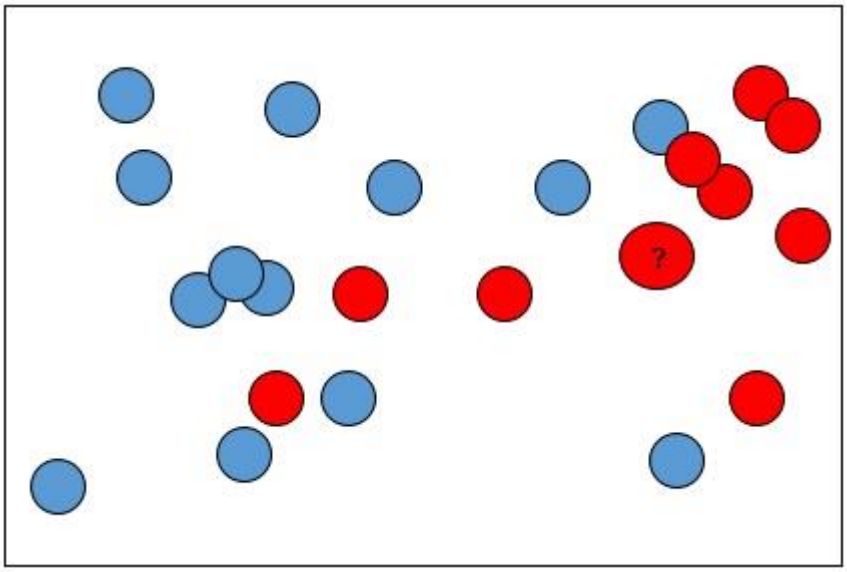


Figura 2.7. Ilustración de la aplicación de los 7-vecinos más cercanos

La elección de k para el método de los k-vecinos más cercanos depende del conjunto con el que se esté trabajando y

normalmente el k idóneo se lo escoge en base a los errores de entrenamiento y de prueba.

En los problemas de la vida real, los elementos del conjunto objetivo tienen más de 2 atributos y por ello no pueden ser visualizados en un plano bidimensional. Sin embargo, se puede definir una función de distancia que permita medir la similitud entre los elementos. Algunas de estas funciones de distancias son: la distancia de Minkowski, la métrica de Canberra, el coeficiente de Czekanowski, entre otras. La selección de la medida de distancia dependerá de la naturaleza de los atributos o variables de los elementos del conjunto objetivo [8].

2.2.3 Naive Bayes

Naive Bayes es un método de discriminación que se basa en el teorema de Bayes:

$$P(A\backslash B) = \frac{P(B\backslash A)P(A)}{P(B)}$$

(2.1)

Este teorema calcula la probabilidad condicional del evento A dado el evento B. Si se supone que el evento A es tener una

enfermedad y el evento B es presentar un síntoma, el teorema de Bayes [9] se traduce a:

$$P(Enfermo \backslash Síntomas) = \frac{P(Síntomas \backslash Enfermo)P(Enfermo)}{P(Síntomas)}$$

(2.2)

De acuerdo a la expresión anterior, la probabilidad de estar enfermo dado que se presentan los síntomas depende de [10]:

- La probabilidad de presentar los síntomas dado que se tiene la enfermedad, esta probabilidad se la estima con el porcentaje de enfermos que presentan los síntomas.
- La probabilidad de tener la enfermedad, esta probabilidad se la estima con la proporción de individuos de la población que tienen la enfermedad.
- La probabilidad de tener los síntomas, esta probabilidad se la estima con la proporción de individuos que presentan los síntomas.

Si se supone que se tienen sólo 2 clases C_1 y C_2 y 2 atributos X_1 y X_2, la aplicación del teorema de Bayes a la discriminación sería la siguiente:

$$P(C_1 \backslash X_1 \wedge X_2) = \frac{P(X_1 \wedge X_2 \backslash C_1)P(C_1)}{P(X_1 \wedge X_2)}$$

(2.3)

$$P(C_2 \backslash X_1 \wedge X_2) = \frac{P(X_1 \wedge X_2 \backslash C_2)P(C_2)}{P(X_1 \wedge X_2)}$$

(2.4)

La clase seleccionada para el nuevo elemento con atributos X_1 y X_2 será aquella que tenga la probabilidad condicional más alta.

El método Naive Bayes aumenta una suposición al análisis anterior, supone que las variables X_1 y X_2 son independientes; con esto $P(X_1 \wedge X_2)=P(X_1)P(X_2)$ [7].

Las nuevas expresiones para las probabilidades condicionales son las siguientes:

$$P(C_1 \backslash X_1 \wedge X_2) = \frac{P(X_1 \wedge X_2 \backslash C_1)P(C_1)}{P(X_1 \wedge X_2)} = \frac{P(X_1 \backslash C_1)P(X_2 \backslash C_1)P(C_1)}{P(X_1)P(X_2)}$$

(2.5)

$$P(C_2 \backslash X_1 \wedge X_2) = \frac{P(X_1 \wedge X_2 \backslash C_2)P(C_2)}{P(X_1 \wedge X_2)} = \frac{P(X_1 \backslash C_2)P(X_2 \backslash C_2)P(C_2)}{P(X_1)P(X_2)}$$

(2.6)

La clase seleccionada será aquella que cumpla con:

$$\max\left\{\frac{P(X_1\backslash C_1)P(X_2\backslash C_1)P(C_1)}{P(X_1)P(X_2)}, \frac{P(X_1\backslash C_2)P(X_2\backslash C_2)P(C_2)}{P(X_1)P(X_2)}\right\}$$

(2.7)

Puesto que las razones anteriores tienen el mismo denominador la expresión se reduce a:

$$\max\{P(X_1\backslash C_1)P(X_2\backslash C_1)P(C_1), P(X_1\backslash C_2)P(X_2\backslash C_2)P(C_2)\}$$

(2.8)

2.2.4 Regresión Logística

Para comprender la regresión logística, es importante presentar primero conceptos relacionados a la regresión lineal.

La regresión lineal pretende estimar el valor de una variable respuesta en base a los valores de los atributos de los elementos del conjunto objetivo. El modelo de regresión lineal es el siguiente:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_n X_n = \beta_0 + \sum_{k=1}^{n} \beta_k X_k$$

(2.9)

En la Ecuación (2.9), Y representa la variable respuesta, las X's representan los atributos o variables independientes y los coeficientes β's son las ponderaciones que representan la influencia individual de cada variable independiente sobre la variable respuesta.

Para el caso de la discriminación entre 2 clases, luego de estimar el valor de Y, se determina que si Y<0.5 entonces el elemento pertenece a la primera clase, caso contrario pertenece a la otra clase. El valor de Y se lo estima aplicando el método de los mínimos cuadrados con los datos del conjunto de entrenamiento. Teniendo presente que en el conjunto de entrenamiento se conoce los valores de Y, se determina los coeficientes de la Ecuación (2.9) tal que se minimice la expresión:

$$\sum_{i=1}^{m} (y_i - \widehat{y_i})^2$$

(2.10)

Donde y_i es el valor de la variable respuesta del i-ésimo elemento y $\widehat{y_i} = \beta_0 + \sum_{k=1}^{n} \beta_k X_k$.

El modelo de regresión logística puede considerarse un modelo de regresión lineal donde la variable respuesta no es Y sino una transformación de Y de la siguiente manera:

$$\ln\left(\frac{Y}{1-Y}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_n X_n = \beta_0 + \sum_{k=1}^{n} \beta_k X_k \quad \textbf{(2.11)}$$

La ventaja de la regresión logística es que en la Ecuación (2.11) el argumento del logaritmo natural debe ser mayor a cero, lo que implica que Y debe estar entre cero y uno. Ahora Y puede ser interpretado como una probabilidad. Nuevamente si Y<0.5 entonces el elemento pertenece a la primera clase, caso contrario el elemento pertenece a la otra clase. En resumen, en regresión logística el valor de Y = P(Y=1) [11].

2.3 PLANIFICACIÓN DE UN PROYECTO DE MINERÍA DE DATOS

En esta sección se describirá las bases teóricas sobre las que se fundamenta el presente proyecto de minería de datos en relación a la metodología utilizada como guía. Se mencionará de forma breve las

metodologías más usadas en proyectos de este tipo, para luego detallar la seleccionada para el presente trabajo: CRISP-DM.

Según las investigaciones realizadas las metodologías más frecuentemente utilizadas para proyectos de minería de datos son: SEMMA, CRISP-DM, KDD y Catalyst [12], [13], [14].

Este proyecto de minería de datos utilizará como guía la metodología CRISP-DM, debido a que es la más seguida a nivel mundial para este tipo de proyectos, es de libre distribución y es independiente de la herramienta o software que se use para llevar a cabo el proceso de minería [14].

La metodología CRISP-DM fue creada en el año 2000 por el grupo de empresas SPSS, NCR y Daimer Chrysler. Consiste en 6 fases no rígidas que funcionan de manera cíclica. Según se observa en la Figura 2.8 las fases son: Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado, Evaluación e Implantación [15], [3].

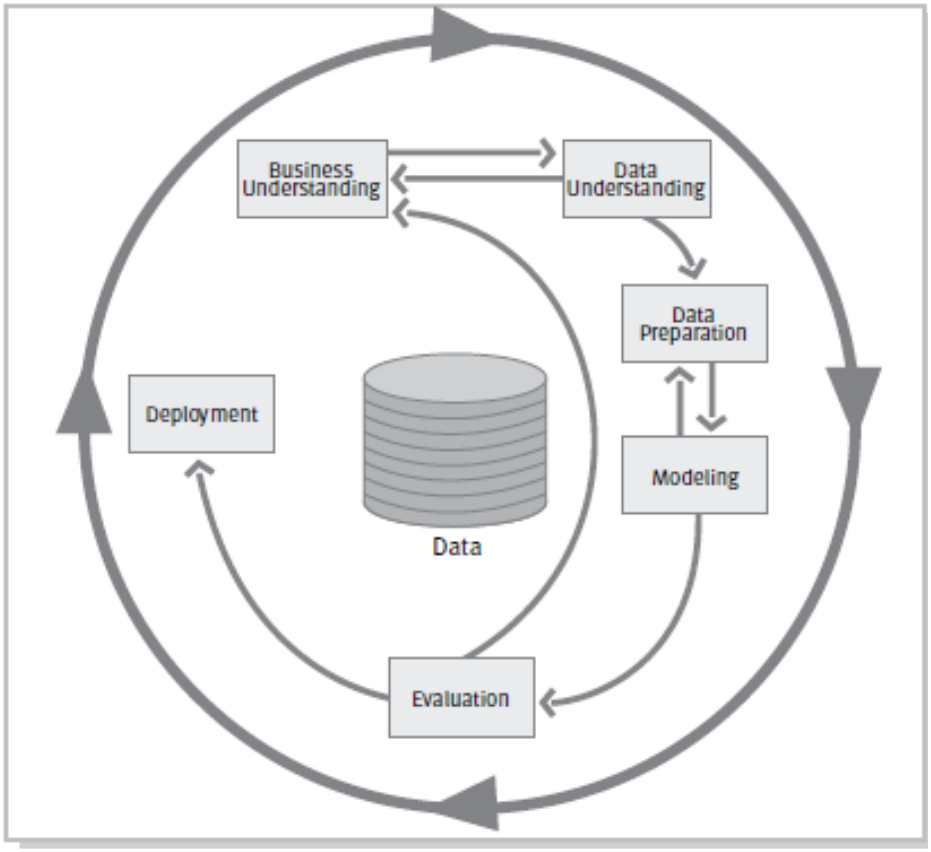


Figura 2.8. Fases de la metodología CRISP-DM. Fuente: SPSS, CRISP-DM 1.0 Step-by-step data mining guide [3]

A continuación, se detallará cada una de las fases que comprende la metodología CRISP-DM:

2.3.1 Comprensión del negocio

En esta primera fase se requiere comprender los objetivos y requisitos del proyecto desde el punto de vista institucional o empresarial, con el fin de transformarlos en objetivos técnicos y realizar la planificación del proyecto. Se debe procurar entender

de forma completa el problema que se quiere resolver, lo cual permitirá luego obtener los datos correctos e interpretar acertadamente los resultados. En resumen se debe poder convertir el conocimiento adquirido del negocio, en un problema de minería de datos [3].

Esta fase se divide en las siguientes tareas [15]:

- Determinar los objetivos del negocio.
- Valoración de la situación.
- Determinar los objetivos de minería de datos.
- Realizar el plan del proyecto.

2.3.2 Comprensión de los datos

En esta fase se debe realizar la primera exploración y recolección de datos, de esta forma se establece un contacto inicial con el problema, se debe identificar la calidad de los datos, establecer relaciones entre ellos que permitan idear las primeras hipótesis. Generalmente esta fase junto con las 2 siguientes, son las que demandan el mayor tiempo y esfuerzo en un proyecto de minería de datos. Podría ser necesario generar a partir de la base de

datos corporativa una nueva base dedicada sólo para el proyecto de minería [3].

Esta fase se divide en las siguientes tareas [15]:

- Recolectar los datos iniciales.
- Descripción de los datos.
- Exploración de los datos.
- Verificar la calidad de los datos.

2.3.3 Preparación de los datos

Esta fase consiste en preparar los datos para adaptarlos a las técnicas de minería de datos que se utilicen en las siguientes fases. Las tareas generales a realizar para la preparación de datos incluyen selección de datos, limpieza de datos, creación de otras variables, integrar datos de diferentes orígenes y cambios en el formato. Esta fase se encuentra muy relacionada con la fase de modelado, ya que dependiendo de la técnica de minería de datos a utilizar, será el procesamiento previo de los datos [3].

Esta fase de divide en las siguientes tareas [15]:

- Seleccionar los datos.

- Limpiar los datos.
- Estructurar e Integrar los datos.
- Formateo de los datos.

2.3.4 Modelado

En esta fase se eligen las técnicas de modelado más adecuadas para el proyecto de minería de datos en cuestión. Se debe considerar previo al modelamiento, un método que permita evaluar los modelos seleccionados y determinar que tan buenos resultaron ser. Luego se procede a la generación y evaluación del modelo [3].

Las tareas en las que se divide esta fase son [15]:

- Seleccionar técnica de modelado.
- Generar el plan de prueba.
- Construir el modelo.
- Evaluar el modelo.

2.3.5 Evaluación

En esta fase se evalúan él o los modelos que componen el proyecto de minería de datos de forma global, teniendo en cuenta si se cumplen los criterios de éxito definidos inicialmente en el problema. Puede ser necesario revisar o repetir los pasos anteriores en busca de errores. Si él o los modelos generados son válidos en función de los indicadores de éxito definidos anteriormente se procede a la implementación del modelo [3].

Esta fase de divide en las siguientes tareas [15]:

- Evaluar los resultados.
- Revisión del proceso.
- Determinar próximos pasos.

2.3.6 Implantación

En esta última fase, luego de que él o los modelos han sido construidos y validados, el conocimiento adquirido debe derivarse en gestiones dentro del proceso de negocio. Estas gestiones podrían ser seguir las recomendaciones basadas en los resultados de los modelos o aplicar el modelo a conjuntos de datos de producción. Adicionalmente se debe documentar y

presentar los resultados de forma amigable para el usuario. También debe asegurarse el mantenimiento de la aplicación y la comunicación de los resultados a las áreas correspondientes [3].

Las tareas que conforman esta fase son [15]:

- Plan de implementación.
- Plan de monitoreo y mantenimiento.
- Informe final y Revisión del proyecto.

2.4 SOFTWARE DE MINERÍA DE DATOS A UTILIZAR

El software que principalmente se utiliza en este proyecto es R, se trata de un software libre que proporciona un lenguaje de programación y un ambiente para el desarrollo de estadística computacional incluida la minería de datos.

El uso y la popularidad de R se han incrementado en los últimos años no sólo en el campo comercial sino también en el campo académico. De acuerdo a un reporte reciente [16], el software R es el segundo más utilizado en los artículos académicos luego del software estadístico de IBM y por encima de conocidos programas estadísticos como SAS, Stata, Minitab, entre otros. Como lenguaje de programación R ocupa el puesto

número 12 en el ranking general y su posición se encuentra a la alza, según lo revela el indicador TIOBE para abril del 2018 [17].

Dos de los aspectos que más atraen a los entusiastas y profesionales de la minería de datos para usar R son:

- Su versatilidad para poder personalizar los modelos y gráficos estadísticos.
- La enorme riqueza de librerías adicionales que pueden ser fácilmente agregadas al ambiente de desarrollo y ejecución de comandos. Estas librerías ofrecen herramientas acerca de los más variados temas de estadística y minería de datos, librerías principalmente desarrolladas por la comunidad R alrededor del mundo.

En relación a la gran cantidad de librerías fáciles de usar que R ofrece, quizás este es uno de los aspectos que más llama la atención. En este proyecto en particular, se han empleado una serie de librerías que han facilitado considerablemente el trabajo.

La Tabla 2 muestra las principales librerías empleadas en este proyecto junto con los comandos más utilizados. Nótese que todas las librerías a excepción de “dplyr” están relacionadas con los métodos de discriminación, mientras que la librería dplyr es empleada en este

proyecto, principalmente para realizar el muestreo de los conjuntos de entrenamiento.

Tabla 2. Principales librerías y comandos utilizados en el proyecto

Librería	Descripción	Comandos utilizados en el proyecto
dplyr	Librería para manipular datos y data frames	sample_n, setdiff, unión
tree	Librería para árboles de clasificación y regresión	tree, plot, text
class	Librería con varias funciones para clasificación incluido knn y mapas auto-organizados	knn, predict,knn.cv
naivebayes	Librería para implementar el algoritmo de Naive Bayes	naive_bayes, predict
stats	Librería incluida en R para varias funciones estadísticas	glm, predict

CAPÍTULO 3

LEVANTAMIENTO DE INFORMACIÓN

3.1 ANÁLISIS EXPLORATORIO PRELIMINAR

La fuente principal de los datos para este proyecto es la base de datos SAAC. Una fuente adicional es la base de datos SIB (Sistema de información bibliotecario). En esta sección se tratará acerca de estas bases de datos y de la información relevante para el proyecto que éstas contienen.

La base de datos SAAC contiene 683 tablas de usuario. Sin embargo, no todas estas tablas contienen datos relevantes para el proyecto. Las tablas con datos que atañen a este proyecto son las siguientes:

- Tabla ESTUDIANTE_ACAD, que contiene datos personales, demográficos y académicos del estudiante. Estos datos cambian muy poco en el transcurso de la carrera del estudiante.
- Tabla HISTORIA_DETAL, que contiene datos acerca del comportamiento académico del estudiante en el transcurso de la carrera.
- Tabla CARRERA_ESTUDIANTE, que contiene datos académicos resumidos respecto a la o las carreras del estudiante.

En relación a la base de datos del SIB se utiliza la tabla PRÉSTAMOS, la cual contiene datos acerca del material bibliográfico consultado por el estudiante en el transcurso de la carrera.

La Tabla 3 muestra un breve resumen de los estudiantes registrados a partir del año 2009; este año es de especial interés para el proyecto, puesto que es el año en el que empezó la gratuidad en las instituciones públicas de educación superior.

Tabla 3.Cantidad de estudiantes por año de ingreso y sexo

Año de Ingreso	Mujeres	% Mujeres por año	Hombres	% Hombres por año	Total
2009	622	41.41%	880	58.59%	1,502
2010	679	40.20%	1,010	59.80%	1,689
2011	746	40.37%	1,102	59.63%	1,848
2012	628	38.50%	1,003	61.50%	1,631

Año de Ingreso	Mujeres	% Mujeres por año	Hombres	% Hombres por año	Total
2013	701	44.03%	891	55.97%	1,592
2014	925	41.93%	1,281	58.07%	2,206
2015	731	41.68%	1,023	58.32%	1,754
2016	587	37.32%	986	62.68%	1,573
2017	583	38.01%	951	61.99%	1,534
Totales	6,202		9,127		15,329
Promedio	689.11	40.38%	1,014.11	59.62%	1,703.22

Tal como lo muestra la Tabla 3 en total han ingresado 15,329 estudiantes de los cuales 9,127 son varones, que representan el 59.54% del total. Este porcentaje es casi igual al porcentaje promedio de estudiantes varones que ingresaron anualmente (celda amarilla). En general, el porcentaje anual de ingreso de estudiantes varones y de mujeres es aproximadamente estable, hecho que se puede visualizar en la Figura 3.1.

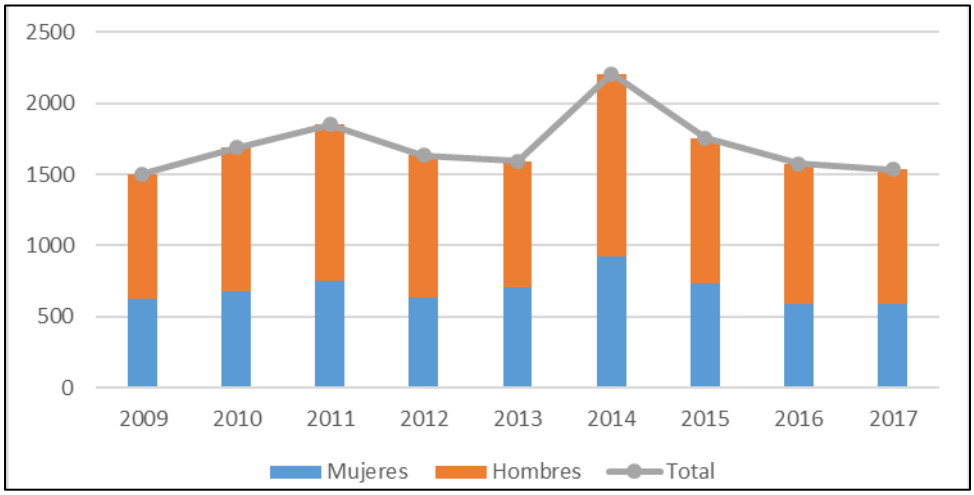


Figura 3.1. Cantidad de estudiantes por año de ingreso y sexo

La Tabla 4 muestra los estudiantes por año de ingreso y estado civil. Aunque la cantidad de estudiantes que no son solteros es considerable, desafortunadamente la actualización de estos datos en el Sistema Académico no es obligatoria salvo para los novatos. Por tanto, un análisis basado en esta variable carecería de confiabilidad. Si un estudiante soltero contrajo matrimonio en el transcurso de su carrera, nada asegura que su estado civil fue cambiado en la base de datos justo el semestre que se casó.

Tabla 4. Cantidad de estudiantes por año de ingreso y estado civil

	Estado civil					
Año de ingreso	Casado	Divorciado	Soltero	Unión libre	Viudo	Total
2009	179	10	1,312	1		1,502
2010	132	8	1,548	1		1,689

	Estado civil					
Año de ingreso	**Casado**	**Divorciado**	**Soltero**	**Unión libre**	**Viudo**	**Total**
2011	96	7	1,743	2		1,848
2012	50	4	1,576	1		1,631
2013	24		1,567		1	1,592
2014	28	3	2,174	1		2,206
2015	6	2	1,746			1,754
2016	12		1,561			1,573
2017	3		1,531			1,534
Total	**530**	**34**	**14,758**	**6**	**1**	**15,329**

En la Tabla 5 se observa la cantidad de estudiantes por año de ingreso y nacionalidad. En las celdas amarillas se observa la cantidad de extranjeros por año de ingreso; aunque estos valores no son significativos, tal como se observa en la Figura 3.2, van en ascenso.

Tabla 5. Cantidad de estudiantes por año de ingreso y nacionalidad

PAÍS	2009	2010	2011	2012	2013	2014	2015	2016	2017	Total
ALEM							3		6	9
ARGE	1		1	1	1	1	1	1	3	10
BOGO	1									1
BOLI								1		1
BULG									1	1
CANA							1			1
CHIL	1			1		1			1	4
CHIN	1	2	1				2		1	7
COLO	2	3	2	2	1	3	4	2	2	21
CORE						1				1
CUBA				1						1

PAÍS	2009	2010	2011	2012	2013	2014	2015	2016	2017	Total
ECUA	1,496	1,683	1,839	1,624	1,578	2,192	1,730	1,556	1,502	15,200
ESPA						2	4	1	2	9
ESTA		1			2		3	1	2	9
FINL							1			1
FRAN					1					1
HAIT					1					1
MEXI			1			1		1	5	8
PANA			1		1				1	3
PERU			2	1	4	1	2	7	5	22
REDO							1			1
RUSI								1		1
TAIW					1	1				2
VENE			1	1	2	3	2	2	3	14
Total Extranjeros	6	6	9	7	14	14	24	17	32	129
Total	**1,502**	**1,689**	**1,848**	**1,631**	**1,592**	**2,206**	**1,754**	**1,573**	**1,534**	**15,329**

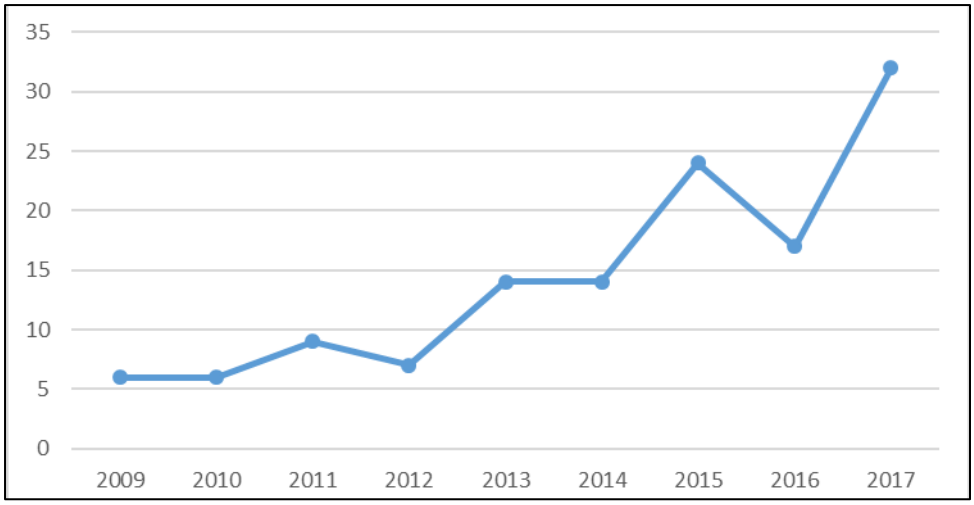


Figura 3.2. Cantidad de estudiantes extranjeros por año de ingreso

A continuación, se realizará un análisis preliminar de la deserción estudiantil en la ESPOL, revisando las estadísticas de ciertas variables

de los estudiantes que en primera instancia se podría llegar a pensar son influyentes sobre la deserción.

En esencia este proyecto trata de comparar las características de los estudiantes desertores versus las características de aquellos que no lo son, para luego de esto definir una regla o modelo que los diferencie. Para etiquetar a un estudiante como desertor se requiere un procesamiento previo de los campos de la base de datos. En este trabajo, un estudiante es considerado desertor si ha dejado de estudiar los últimos 3 años (2015, 2016, y 2017) y no se ha graduado.

En la Tabla 6 se observa el porcentaje de deserción de los estudiantes que ingresaron en el período 2009-2013 clasificados por género. Tal como se observa en esta tabla y en la Figura 3.3, el porcentaje de deserción ha ido decreciendo hasta situarse en el 10.56% para las mujeres y el 14.48% para los hombres; sin embargo, aún son porcentajes altos, especialmente, si se los traduce a cantidades absolutas.

Tabla 6. Porcentaje de deserción por año de ingreso

Año	Mujeres	Hombres
2009	21.38%	30.57%
2010	15.91%	28.12%
2011	15.15%	23.05%
2012	11.15%	16.85%
2013	10.56%	14.48%

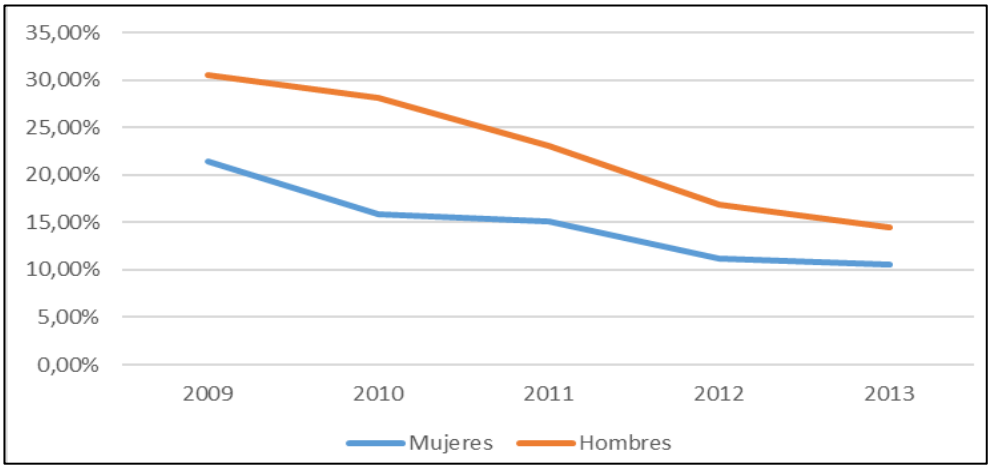


Figura 3.3. Comportamiento de la deserción por año y sexo

La Tabla 6 y la Figura 3.3 también reflejan que existe una considerable diferencia en la deserción de ambos grupos. La figura muestra no sólo que ha existido mayor porcentaje de desertores varones que de mujeres, sino que la brecha entre estos 2 grupos se ha acortado, pero siempre ha existido. Aparentemente, las mujeres son menos propensas a desertar de sus estudios.

Aunque en el Capítulo 5 se construirán los modelos para analizar las características más influyentes en la deserción de un estudiante incluido el sexo, en esta sección se discutirá muy brevemente la influencia del género en la deserción de los estudiantes.

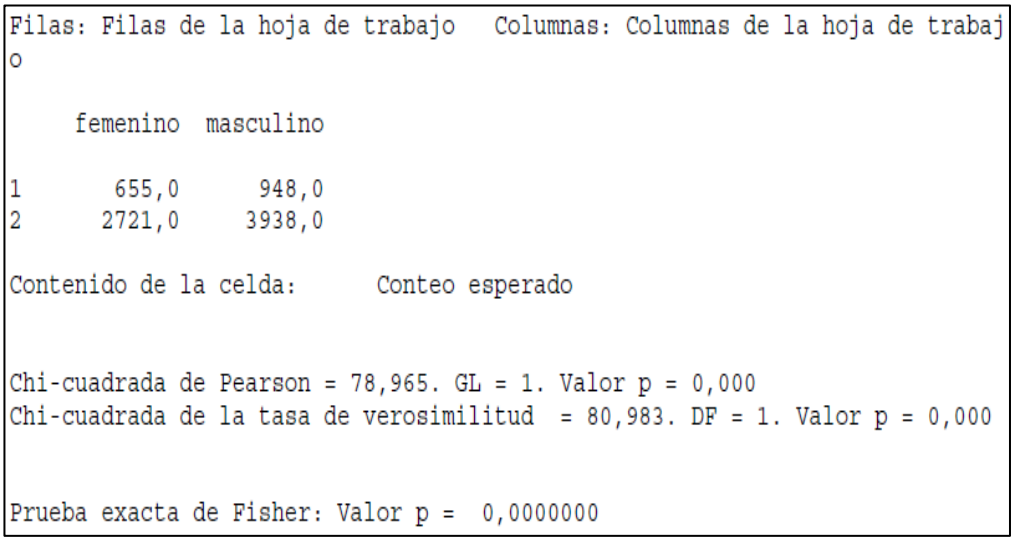
La Tabla 7 muestra una comparación entre desertores y no-desertores clasificados por género. En general los desertores representan casi el

20% del total de estudiantes (celda azul). Si el género no tuviera ninguna influencia en el porcentaje de deserción, este 20% de deserción debería aproximadamente mantenerse al analizar el grupo de varones y el grupo de mujeres por separado. Sin embargo, tal como se refleja en las celdas amarillas de la tabla, el porcentaje de deserción de las mujeres es considerablemente menor al 20% mientras que el de los hombres es ligeramente mayor.

Tabla 7. Deserción de estudiantes que ingresaron desde el 2009 al 2013

	Femenino		Masculino		Todos	
	Cantidad	%	Cantidad	%	Cantidad	%
Desertor	498	14.75%	1,105	22.62%	1,603	19.40%
NO Desertor	2,878	85.25%	3,781	77.38%	6,659	80.60%
Totales	3,376	100.00%	4,886	100.00%	8,262	100.00%

La Figura 3.4 muestra una prueba estadística realizada en el software Minitab para la Tabla 7. El valor p de 0.0000 indica que la posibilidad de que los resultados de la tabla (celdas amarillas) sean fruto del azar es de menos de 1 en un millón. Lo que indica que la deserción y el género no son variables independientes.



```
Filas: Filas de la hoja de trabajo   Columnas: Columnas de la hoja de trabaj
o

     femenino  masculino

1       655,0      948,0
2      2721,0     3938,0

Contenido de la celda:      Conteo esperado


Chi-cuadrada de Pearson = 78,965. GL = 1. Valor p = 0,000
Chi-cuadrada de la tasa de verosimilitud  = 80,983. DF = 1. Valor p = 0,000


Prueba exacta de Fisher: Valor p =  0,0000000
```

Figura 3.4. Prueba estadística de independencia para sexo vs deserción (Software Minitab)

Tal como se lo hizo con el género, también se pudiera revisar otras características y verificar si éstas son independientes de la deserción, siempre y cuando estas características tengan la significancia y confiabilidad requerida. Por ejemplo, la cantidad de estudiantes casados es significativa pero no es confiable; por otro lado, la nacionalidad es confiable pero la cantidad de extranjeros no es muy significativa con respecto al total de estudiantes. Sin embargo, el análisis discriminante que se realizará en los Capítulos 5 y 6 aborda de forma más amplia estas posibilidades.

3.2 SELECCIÓN DEL CONJUNTO DE DATOS OBJETIVO

En el presente proyecto no se realizará un estudio longitudinal de un conjunto de estudiantes a través del tiempo, sino más bien se realizará el análisis de un conjunto de estudiantes en un instante de tiempo, en este caso el instante de tiempo es un año y semestre particular. La idea es tomar las características de un conjunto de estudiantes en un semestre determinado y predecir su posible deserción.

Tal como se ilustra en la Figura 3.5, se tomará como “instante de tiempo” el segundo semestre del año 2011. Es como si se tomara una fotografía al término 2011-2S y se analizaran las características de todos los estudiantes que “salieron en la foto”. En concreto, el conjunto de datos objetivo para este proyecto son todos los estudiantes que ingresaron desde el año 2009 y estudiaron en el semestre 2011-2S (véase la Figura 3.5).

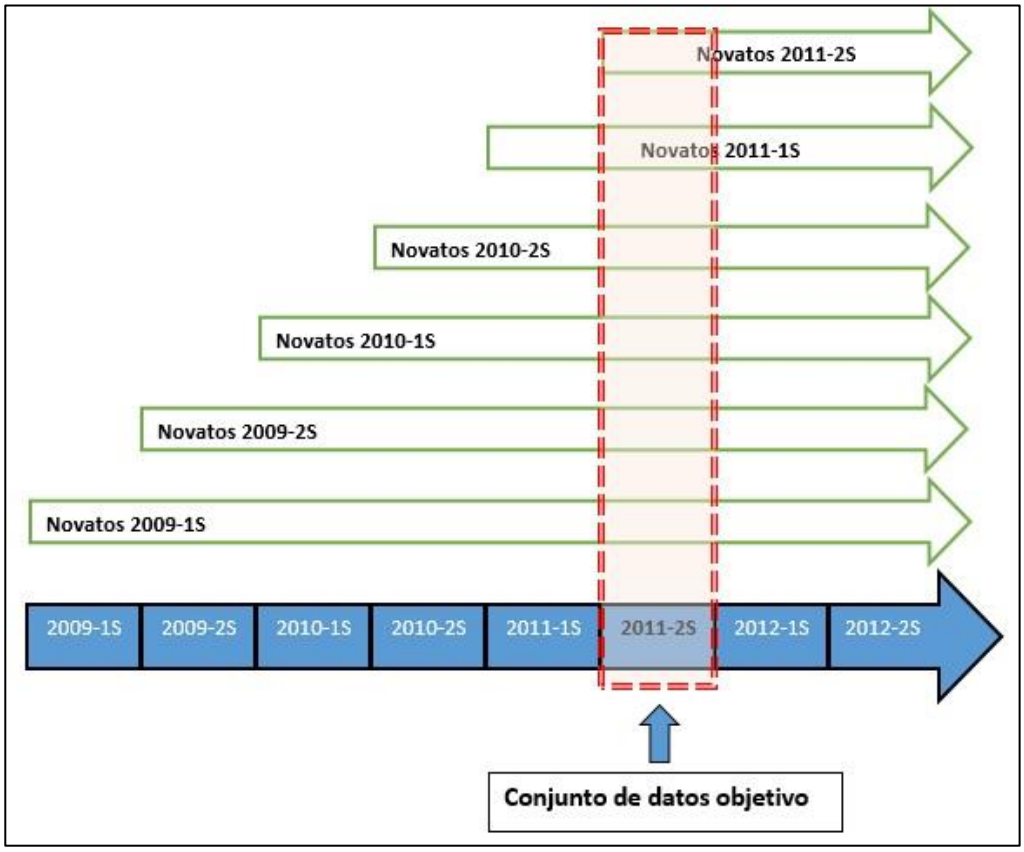
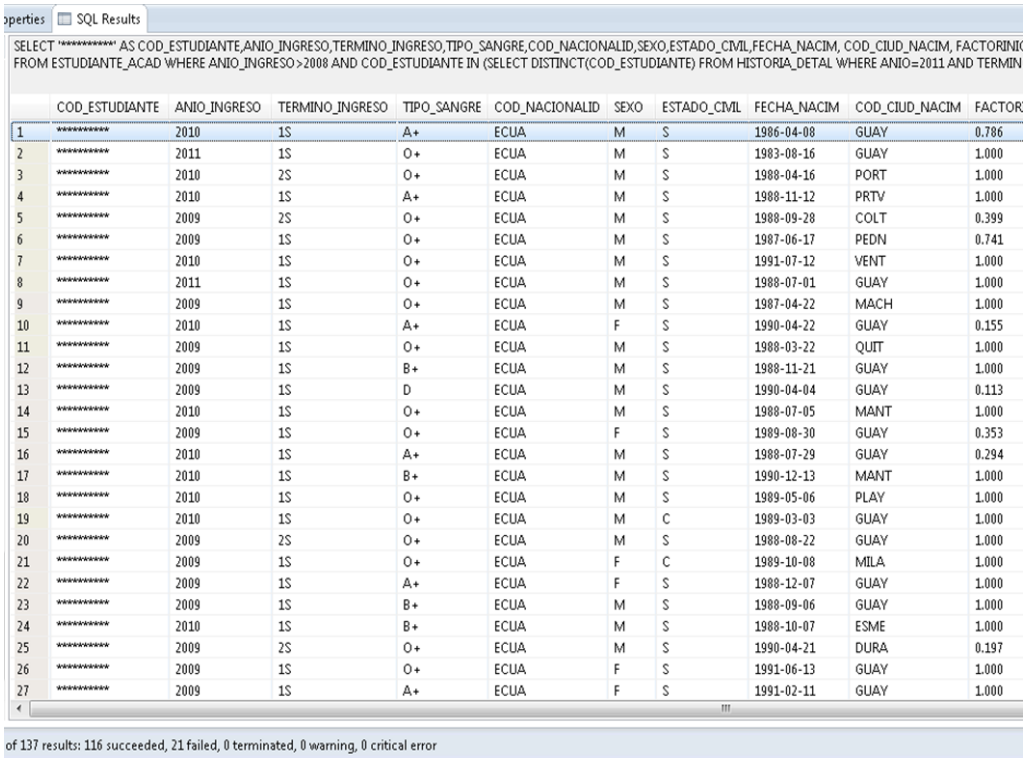


Figura 3.5. Esquema de selección del conjunto de datos objetivo

En cuanto a la elección del “instante de tiempo” 2011-2S, no tiene nada particular, tan sólo que garantiza que los estudiantes etiquetados como desertores o no-desertores efectivamente lo serán, pues ya han pasado más de los 3 años que se requiere para que sean considerados como tal.

En la Figura 3.6 se observa una muestra anonimizada de este conjunto de datos objetivo tomados de la tabla ESTUDIANTE_ACAD de la base

de datos del Sistema Académico; mientras que la Tabla 8 muestra algunas estadísticas de este conjunto.



SQL Results

SELECT '**********' AS COD_ESTUDIANTE,ANIO_INGRESO,TERMINO_INGRESO,TIPO_SANGRE,COD_NACIONALID,SEXO,ESTADO_CIVIL,FECHA_NACIM, COD_CIUD_NACIM, FACTORINI(
FROM ESTUDIANTE_ACAD WHERE ANIO_INGRESO>2008 AND COD_ESTUDIANTE IN (SELECT DISTINCT(COD_ESTUDIANTE) FROM HISTORIA_DETAL WHERE ANIO=2011 AND TERMIN

	COD_ESTUDIANTE	ANIO_INGRESO	TERMINO_INGRESO	TIPO_SANGRE	COD_NACIONALID	SEXO	ESTADO_CIVIL	FECHA_NACIM	COD_CIUD_NACIM	FACTOR
1	**********	2010	1S	A+	ECUA	M	S	1986-04-08	GUAY	0.786
2	**********	2011	1S	O+	ECUA	M	S	1983-08-16	GUAY	1.000
3	**********	2010	2S	O+	ECUA	M	S	1988-04-16	PORT	1.000
4	**********	2010	1S	A+	ECUA	M	S	1988-11-12	PRTV	1.000
5	**********	2009	2S	O+	ECUA	M	S	1988-09-28	COLT	0.399
6	**********	2009	1S	O+	ECUA	M	S	1987-06-17	PEDN	0.741
7	**********	2010	1S	O+	ECUA	M	S	1991-07-12	VENT	1.000
8	**********	2011	1S	O+	ECUA	M	S	1988-07-01	GUAY	1.000
9	**********	2009	1S	O+	ECUA	M	S	1987-04-22	MACH	1.000
10	**********	2010	1S	A+	ECUA	F	S	1990-04-22	GUAY	0.155
11	**********	2009	1S	O+	ECUA	M	S	1988-03-22	QUIT	1.000
12	**********	2009	1S	B+	ECUA	M	S	1988-11-21	GUAY	1.000
13	**********	2009	1S	D	ECUA	M	S	1990-04-04	GUAY	0.113
14	**********	2010	1S	O+	ECUA	M	S	1988-07-05	MANT	1.000
15	**********	2009	1S	O+	ECUA	F	S	1989-08-30	GUAY	0.353
16	**********	2010	1S	A+	ECUA	M	S	1988-07-29	GUAY	0.294
17	**********	2010	1S	B+	ECUA	M	S	1990-12-13	MANT	1.000
18	**********	2010	1S	O+	ECUA	M	S	1989-05-06	PLAY	1.000
19	**********	2010	1S	O+	ECUA	M	C	1989-03-03	GUAY	1.000
20	**********	2009	2S	O+	ECUA	M	S	1988-08-22	GUAY	1.000
21	**********	2009	1S	O+	ECUA	F	C	1989-10-08	MILA	1.000
22	**********	2009	1S	A+	ECUA	F	S	1988-12-07	GUAY	1.000
23	**********	2009	1S	B+	ECUA	M	S	1988-09-06	GUAY	1.000
24	**********	2010	1S	B+	ECUA	M	S	1988-10-07	ESME	1.000
25	**********	2009	2S	O+	ECUA	M	S	1990-04-21	DURA	0.197
26	**********	2009	1S	O+	ECUA	F	S	1991-06-13	GUAY	1.000
27	**********	2009	1S	A+	ECUA	F	S	1991-02-11	GUAY	1.000

of 137 results: 116 succeeded, 21 failed, 0 terminated, 0 warning, 0 critical error

Figura 3.6. Muestra del conjunto de datos objetivo

Tabla 8. Resumen estadístico básico del conjunto de datos objetivo

Año de ingreso	# de estudiantes	Sexo		Nacionalidad		Estado civil	
2009	1,165	**Femenino**	1,805	**Ecuatoriana**	4,282	**Soltero**	3,994
2010	1,413	**Masculino**	2,489	**Extranjera**	12	**Casado**	281
2011	1,716					**Otros**	19
Totales	**4,294**		**4,294**		**4,294**		**4,294**

Existen características o variables de los estudiantes que no se encuentran en la base de datos, sino que requieren ser calculadas previamente. Nótese que en la Figura 3.6 sólo aparece la fecha de nacimiento, pero la edad tiene que ser calculada. Algunas características como la edad no requieren mayor cálculo; sin embargo, existen otras características que requerirán mayor procesamiento, especialmente, los atributos que tienen que ver con el rendimiento académico del estudiante.

Los atributos más importantes respecto al rendimiento del estudiante son promedio, materias aprobadas, materias reprobadas entre otros. Nótese que no se requiere el promedio general del estudiante sino más bien el promedio que éste tenía en el semestre 2011-2S que es el instante de tiempo escogido para el análisis.

En el Capítulo 4 se discutirán las características seleccionadas para el presente estudio y el pre-procesamiento requerido para aquellas que no pueden ser tomadas directamente de la base de datos del Sistema Académico.

3.3 LIMPIEZA Y PREPARACIÓN DEL CONJUNTO DE DATOS OBJETIVO

La calidad de la información que se obtenga en un proyecto de minería dependerá en gran medida de la calidad de los datos con los que se trabaje. Por esta razón, es necesario realizar una limpieza y preparación

adecuada del conjunto de datos. En el caso de este proyecto, se deben examinar principalmente las tablas relevantes de la base del Sistema Académico en busca de datos erróneos e incoherencias que puedan afectar el futuro análisis.

Una forma de revisar la presencia de datos erróneos es verificar que todos los campos numéricos importantes de las tablas relevantes se encuentren en un rango razonable. Examinar los rangos puede reflejar posibles datos erróneos o aumentar nuestra comprensión del conjunto de datos objetivo.

A continuación, se expondrá de forma detallada uno de los casos detectados en este proyecto. El caso en particular es la aparente incoherencia en las materias convalidadas por los estudiantes del conjunto de datos en análisis.

La Tabla 9 refleja los rangos en los que se encuentran algunos de los campos más importantes de las tablas ESTUDIANTE_ACAD e HISTORIA_DETAL. Si se tiene presente que el conjunto de datos objetivo son los estudiantes que ingresaron a partir del año 2009, el valor mínimo del campo ANIO en la Tabla 9 (celda en amarillo) se encuentra fuera del rango razonable, pues es menor que el año de ingreso del estudiante.

Tabla 9. Rango de valores de algunas características numéricas del conjunto de datos objetivo

Campo	Descripción	Mínimo	Máximo
FECHA_NACIM	Fecha de nacimiento	4/16/1971	7/25/1995
P_ANTERIOR	Factor socioeconómico	0	40
PROMEDIO	Nota final de la materia	0	10
VEZ_TOMADA	# de vez en que se toma la materia	0	4
COD_ESTUDIANTE	# de matrícula	2005*****	2011*****
ANIO	Año en que se tomó la materia	2007	2011

Una forma de aclarar la aparente incoherencia que se muestra en la Tabla 9, es realizar una búsqueda exhaustiva de los casos en los que se presenta esta situación. Al buscar los registros en los que el campo ANIO (año en que se tomó la materia) es menor al año de ingreso, se obtienen los 20 registros mostrados en la Tabla 10. Tal como se observa en esta tabla, la aparente incoherencia no es tal; pues se tratan de materias convalidadas que el estudiante aprobó antes de ingresar a la Espol.

Tabla 10. Datos en los que el año de ingreso es mayor al año en que se tomó una materia

#	COD_ESTUDIANTE	ANIO	TERMINO	ESTADO_MAT_TOMADA	PROMEDIO
1	2009*****	2007	1S	CV	0
2	2009*****	2007	2S	CV	0
3	2009*****	2007	1S	CV	0
4	2009*****	2008	2S	CV	0

#	COD_ESTUDIANTE	ANIO	TERMINO	ESTADO_MAT_TOMADA	PROMEDIO
5	2009*****	2008	1S	CV	0
6	2010*****	2007	1S	CV	0
7	2010*****	2007	2S	CV	0
8	2010*****	2007	1S	CV	0
9	2010*****	2008	2S	CV	0
10	2010*****	2008	1S	CV	0
11	2010*****	2007	1S	CV	0
12	2010*****	2007	2S	CV	0
13	2010*****	2007	1S	CV	0
14	2010*****	2008	2S	CV	0
15	2010*****	2008	1S	CV	0
16	2010*****	2007	1S	CV	0
17	2010*****	2007	2S	CV	0
18	2010*****	2007	1S	CV	0
19	2010*****	2008	2S	CV	0
20	2010*****	2008	1S	CV	0

CV: Convalidada

La Tabla 10 refleja un aspecto importante que debe considerarse: las materias convalidadas tienen como nota final (promedio) el valor de cero, esto no es una coincidencia. En el conjunto de datos objetivo se tienen 1,226 materias convalidadas, de las cuales todas tienen nota final de cero. La importancia de esta observación radica en que las materias convalidadas deben considerarse como las materias que el estudiante

ya tiene aprobadas en la carrera, pero no pueden considerarse para calcular el promedio general del estudiante.

Para ejemplificar la situación expuesta en el párrafo anterior, la Tabla 11 presenta un ejemplo hipotético de un estudiante que ha aprobado 2 materias y le han convalidado 3. De acuerdo a esta tabla, el estudiante tiene en su haber 5 materias que se le consideran como aprobadas, sin embargo, el promedio sólo debe considerar las 2 materias con nota mayor que cero. De esta manera, el promedio correcto no es (0 + 6.5 + 0 + 7.2 + 0) /5 = 2.74 sino (6.5 + 7.2) /2 = 6.85.

Tabla 11. Cálculo correcto del promedio general de notas del estudiante

Materia	Estado	Nota
A	CV	0
B	AP	6.5
C	CV	0
D	AP	7.2
E	CV	0
PROMEDIO INCORRECTO		**2.74**
PROMEDIO CORRECTO		**6.85**

CV: Convalidada AP: Aprobada

Tal como se mostró en el caso de la aparente incoherencia de las materias convalidadas por los estudiantes del conjunto de datos objetivo, algunas de estas incongruencias en los datos pueden resolverse con un

correcto procedimiento al calcular los valores de las variables que participarán del modelo de discriminación. Algunas variables pueden ser tomadas directamente de los campos de las tablas de las bases de datos, mientras que otras requerirán un procesamiento previo; es en este procesamiento que deberán tomarse las precauciones necesarias para evitar errores que afecten la calidad de los datos. Las secciones 4.1 y 4.2 tratarán acerca de las variables seleccionadas para los modelos de discriminación y el procedimiento correcto para calcular sus valores.

CAPÍTULO 4

ANÁLISIS Y DISEÑO DE LA SOLUCIÓN

4.1 SELECCIÓN DE LAS VARIABLES PARA LOS MODELOS

La decisión de cuáles variables utilizar para el análisis discriminante depende, en este proyecto, de varios factores tales como: la revisión de la literatura existente, la disponibilidad de los datos, la confiabilidad de los datos existentes, el análisis estadístico preliminar de los datos, las sugerencias de especialistas en docencia universitaria, entre otros.

El conjunto de variables seleccionadas para este proyecto se divide en 2 grupos: las variables relacionadas a las características personales del estudiante y las variables relacionadas al comportamiento académico del estudiante.

La mayoría de las variables seleccionadas requieren un procesamiento adicional para encontrar sus valores. En la Figura 4.1 se observa un esquema general de las tablas y campos utilizados para producir las variables.

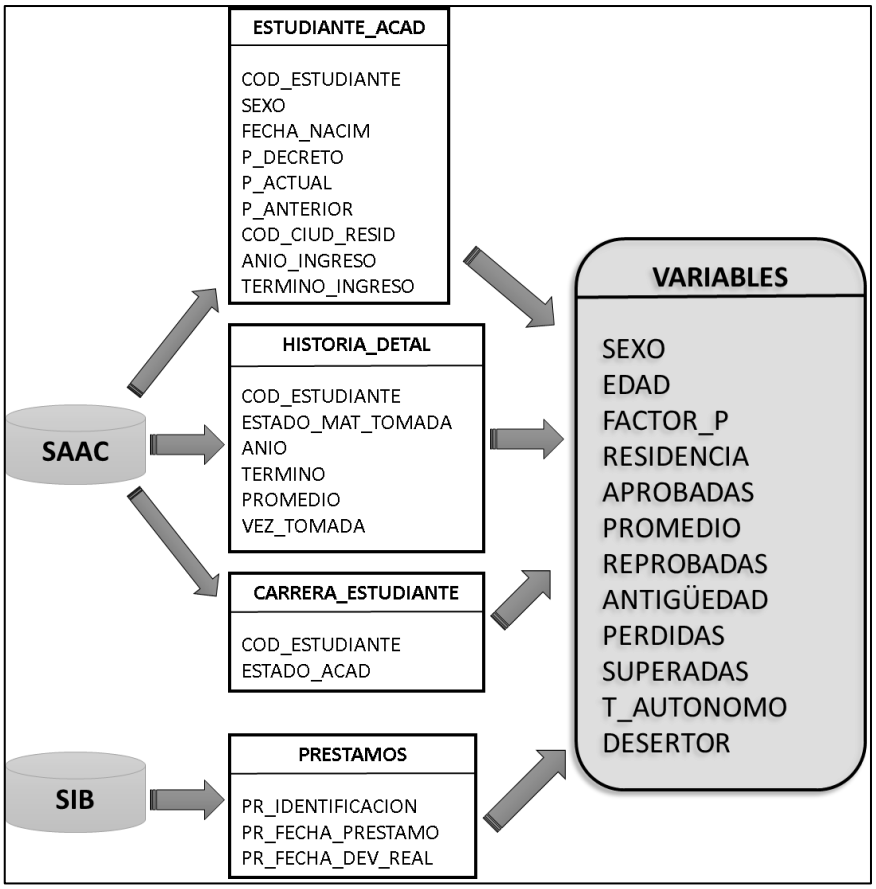


Figura 4.1. Esquema de generación de las variables

Tal como se observa en la Figura 4.1, en este proyecto se trabajará con 12 variables que de manera general provienen de 4 tablas, 3 tablas de la

base SAAC y 1 tabla de la base de datos SIB. En la Tabla 12 se indica la descripción de cada una de estas variables.

Tabla 12. Descripción de las variables seleccionadas

#	Variable	Descripción	Tipo	Posibles valores
1	SEXO	Sexo del estudiante	Categórica	{F, M}
2	EDAD	Edad del estudiante	Numérica, entera	16 en adelante
3	FACTOR_P	Indicador del nivel socioeconómico del estudiante	Numérica, entera	0 a 40
4	RESIDENCIA	Tipo de residencia del estudiante	Categórica	{LOCAL,PROV}
5	APROBADAS	Número de materias aprobadas	Numérica, entera	0 en adelante
6	REPROBADAS	Número de materias reprobadas	Numérica, entera	0 en adelante
7	PROMEDIO	Promedio general del estudiante	Numérica, racional	0 a 10
8	ANTIGÜEDAD	Número de semestres de estudio	Numérica, entera	0 a 5
9	PERDIDAS	Número de veces en que perdió un período de prueba	Numérica, entera	Desde 0 en adelante
10	SUPERADAS	Número de veces en que superó un período de prueba	Numérica, entera	Desde 0 en adelante
11	T_AUTONOMO	Indicador de la cantidad de trabajo autónomo del estudiante en el semestre actual	Numérica, entera	Desde 0 en adelante
12	DESERTOR	Etiqueta que indica si el estudiante ha o no ha desertado	Categórica, variable respuesta	{SI,NO}

La variable DESERTOR se conoce como la variable respuesta, el valor de esta variable podrá ser estimado a partir de los valores que tomen las

11 variables restantes, también conocidas como variables explicativas o predictoras. Precisamente uno de los objetivos de este proyecto es plantear un modelo que permita predecir si un estudiante es desertor dados los valores de sus variables predictoras.

Tal como puede intuirse de la Tabla 12, las variables relacionadas al comportamiento académico del estudiante son: APROBADAS, REPROBADAS, PROMEDIO, PERDIDAS, SUPERADAS y T_AUTONOMO. A excepción de la variable T_AUTONOMO, estas variables hacen un recuento desde el año de ingreso de los estudiantes hasta el "instante de tiempo" escogido (véase la sección 3.2), siendo este el segundo semestre del año 2011.

En la sección 4.2 se explicará con mayor detalle la forma de calcular los valores de las variables a las que se ha hecho referencia en esta sección.

4.2 PRE-PROCESAMIENTO DE LOS DATOS

En esta sección se expondrá el tratamiento y operaciones realizadas sobre los datos contenidos en las bases de datos SAAC y SIB, con el fin de obtener las variables seleccionadas para su posterior uso en la aplicación de los algoritmos de minería.

Se debe mencionar que según la forma de obtención se consideran 2 tipos de variables:

- Variables directas, sus valores son obtenidos directamente de uno de los campos de una tabla específica de la base de datos, es decir no se debe realizar en este caso ningún pre-procesamiento.

- Variables pre-procesadas, para obtener los valores requeridos por los modelos de minería se debe realizar previamente ciertas operaciones sobre los datos del repositorio. Es decir, estos valores se los obtiene de forma indirecta.

En la Tabla 13 se muestra un resumen de todas las variables. Se indica la forma de obtención (D: directa o P: pre-procesada), los nombres de los campos y tablas de donde se obtuvo los datos necesarios. Además, se indica entre paréntesis al lado del nombre de las tablas, el nombre de la base de datos de donde se extrae la información (SAAC o SIB).

Tabla 13. Resumen de las variables, forma y origen de obtención

Descripción de la Variable	Variable	Forma	Campo	Tabla
Sexo	SEXO	D	SEXO	ESTUDIANTE_ACAD (SAAC)
Edad	EDAD	P	FECHA_NACIM	ESTUDIANTE_ACAD (SAAC)
Factor Socioeconó-mico	FACTOR_P	P	P_ACTUAL P_DECRETO P_ANTERIOR	ESTUDIANTE_ACAD (SAAC)
Lugar de Residencia	RESIDENCIA	P	COD_CIUD_RESID	ESTUDIANTE_ACAD (SAAC)
Número de materias Aprobadas	APROBADAS	P	COD_ESTUDIANTE ESTADO_MAT_TOMADA ANIO	ESTUDIANTE_ACAD HISTORIA_DETAL (SAAC)
Promedio de materias tomadas	PROMEDIO	P	COD_ESTUDIANTE PROMEDIO ANIO ESTADO_MAT_TOMADA	ESTUDIANTE_ACAD HISTORIA_DETAL (SAAC)
Número de materias Reprobadas	REPROBADAS	P	COD_ESTUDIANTE ESTADO_MAT_TOMADA ANIO	ESTUDIANTE_ACAD HISTORIA_DETAL (SAAC)
Número de semestres cursados	ANTIGUEDAD	P	ANIO_INGRESO TERMINO_INGRESO	ESTUDIANTE_ACAD (SAAC)
Número de veces que perdió materia a prueba	PERDIDAS	P	COD_ESTUDIANTE VEZ_TOMADA ESTADO_MAT_TOMADA ANIO	ESTUDIANTE_ACAD HISTORIA_DETAL (SAAC)
Número de veces que aprobó materia a prueba	SUPERADAS	P	COD_ESTUDIANTE VEZ_TOMADA ESTADO_MAT_TOMADA ANIO	ESTUDIANTE_ACAD HISTORIA_DETAL (SAAC)
¿Es desertor?	DESERTOR	P	COD_ESTUDIANTE ANIO_INGRESO ANIO TERMINO ESTADO_ACAD	ESTUDIANTE_ACAD HISTORIA_DETAL CARRERA_ESTUDIANTE (SAAC)
Índice de trabajo autónomo	T_AUTONOMO	P	PR_IDENTIFICACION PR_ID_PRESTAMO PR_FECHA_DEV_REAL PR_FECHA_PRESTAMO	PRESTAMO (SIB)

D: Directa P: Pre-procesada

Tal como se indica en la Tabla 13, la variable Sexo es la única que se obtiene de forma directa. Esta variable se toma de la tabla ESTUDIANTE_ACAD, la cual contiene la mayoría de datos personales de los estudiantes.

Se puede observar en la Figura 4.2 que la variable Edad se obtiene de forma indirecta del campo FECHA_NACIM de la tabla ESTUDIANTE_ACAD. La operación que se utiliza en el pre-procesamiento se muestra en la misma figura.

Figura 4.2. Pre-procesamiento de la variable Edad

De igual manera, puede observarse en la Figura 4.3 que la variable Factor Socioeconómico (FACTOR_P) se obtiene de forma indirecta de los campos P_ACTUAL, P_DECRETO y P_ANTERIOR que pertenecen a la tabla ESTUDIANTE_ACAD. La operación que se utiliza sobre los

campos mencionados para llegar a obtener la variable final también es mostrada en la misma figura. El valor obtenido será un número entero entre 0 y 40, el cual es un indicador del nivel socioeconómico.

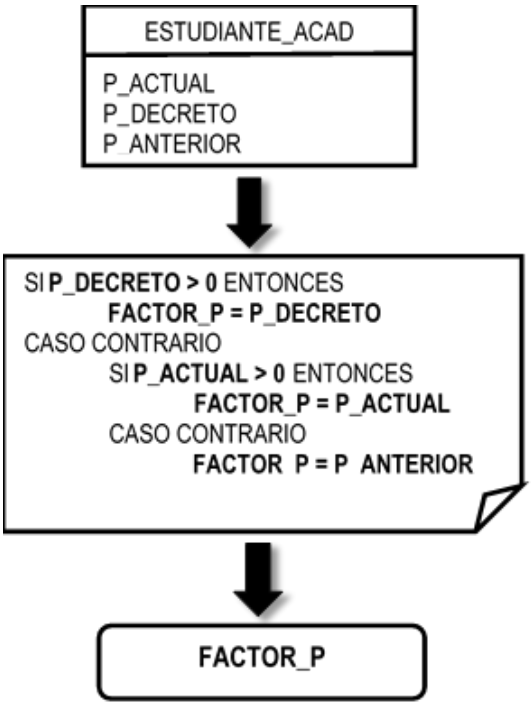


Figura 4.3. Pre-procesamiento de la variable Factor Socioeconómico

La variable Lugar de Residencia (RESIDENCIA) se obtiene de forma indirecta del campo COD_CIUD_RESID de la tabla ESTUDIANTE_ACAD. La operación que se utiliza en el pre-procesamiento se muestra en la Figura 4.4. El valor finalmente obtenido será LOCAL o PROVINCIA.

Figura 4.4. Pre-procesamiento de la variable Lugar de Residencia

La variable Número de materias Aprobadas (APROBADAS) se obtiene de forma indirecta del campo COD_ESTUDIANTE de la tabla ESTUDIANTE_ACAD y de los campos COD_ESTUDIANTE, ESTADO_MAT_TOMADA y ANIO de la tabla HISTORIA_DETAL. La operación que se utiliza sobre los campos mencionados para llegar a obtener la variable final se muestra en la Figura 4.5. Cabe mencionar que ambas tablas se relacionan entre sí por medio del campo COD_ESTUDIANTE. El valor finalmente obtenido será un número entero que indica la cantidad de materias aprobadas por un estudiante hasta la fecha de corte.

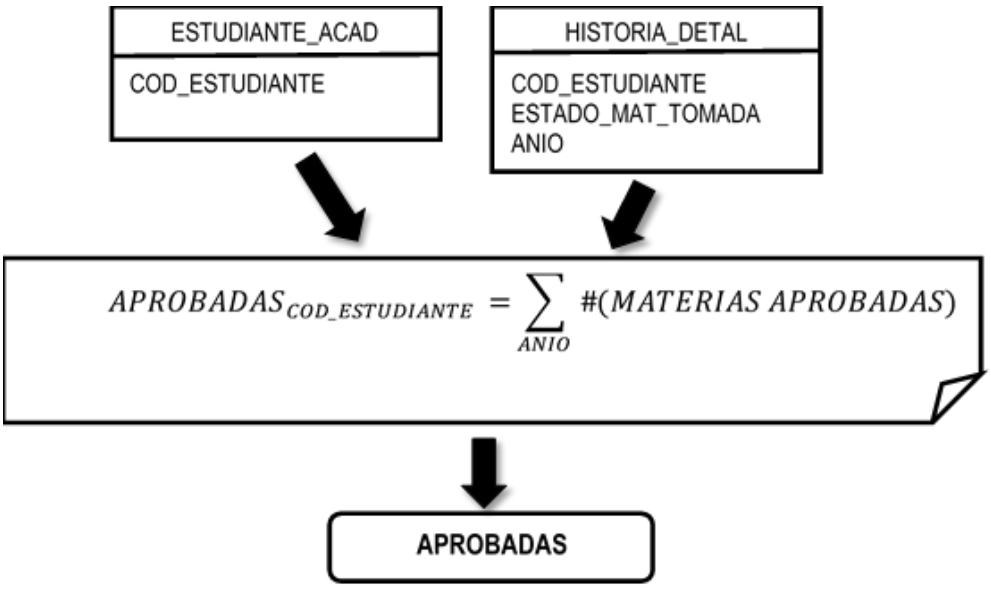


Figura 4.5. Pre-procesamiento de la variable Número de materias Aprobadas

La operación que se utiliza para obtener la variable Promedio de materias tomadas (PROMEDIO) se muestra en la Figura 4.6. El valor finalmente obtenido será el promedio de todas las notas finales de las materias tomadas por el estudiante hasta la fecha de corte.

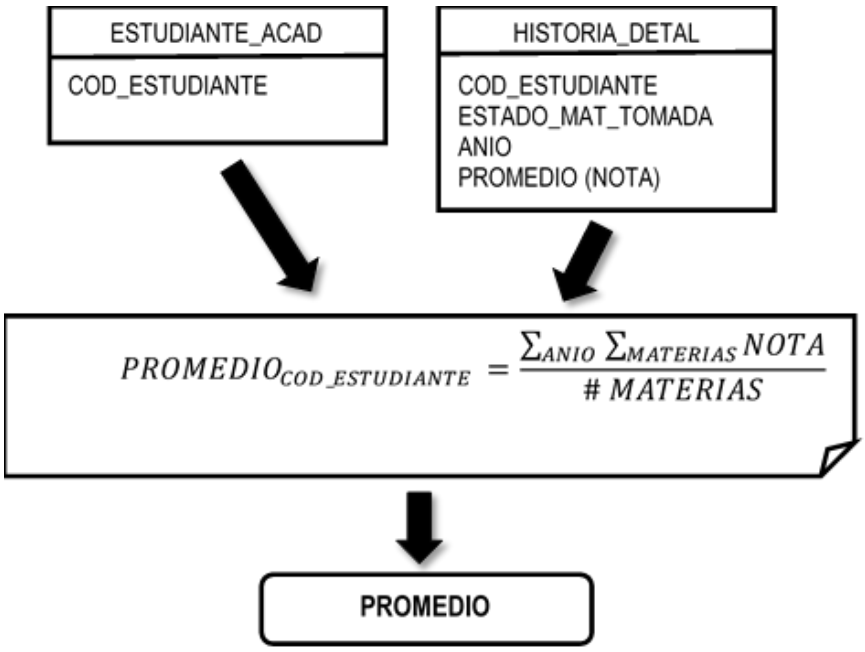


Figura 4.6. Pre-procesamiento de la variable Promedio de materias tomadas

La operación que se utiliza para obtener la variable Número de materias Reprobadas (REPROBADAS) se muestra en la Figura 4.7. El valor final será un número entero que indica la cantidad de materias reprobadas por un estudiante hasta la fecha de corte.

Figura 4.7. Pre-procesamiento de la variable Número de materias Reprobadas

La operación que se utiliza para obtener la variable Desertor se muestra en la Figura 4.8. Las 3 tablas que se utilizan para el pre-procesamiento se relacionan entre sí por medio del campo COD_ESTUDIANTE. Se obtendrá como respuesta un “SI” o “NO” que indica si el estudiante es o no desertor.

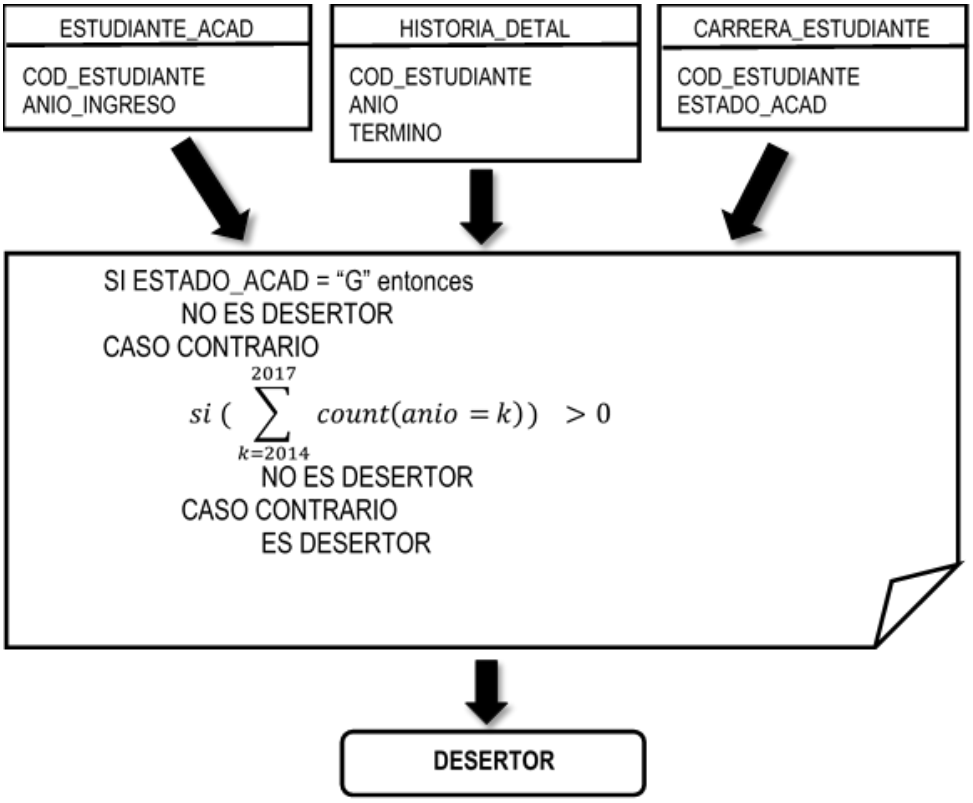


Figura 4.8. Pre-procesamiento de la variable Desertor

De la misma forma ya detallada en los párrafos anteriores se obtienen a través de un pre-procesamiento las variables: ANTIGÜEDAD, PERDIDAS y SUPERADAS.

Finalmente, la variable Índice de trabajo autónomo (T_AUTONOMO) es la única que se obtiene de la base de datos SIB. La Figura 4.9 muestra el pre-procesamiento realizado para obtener esta variable. El valor obtenido será un indicador del trabajo autónomo realizado por el estudiante.

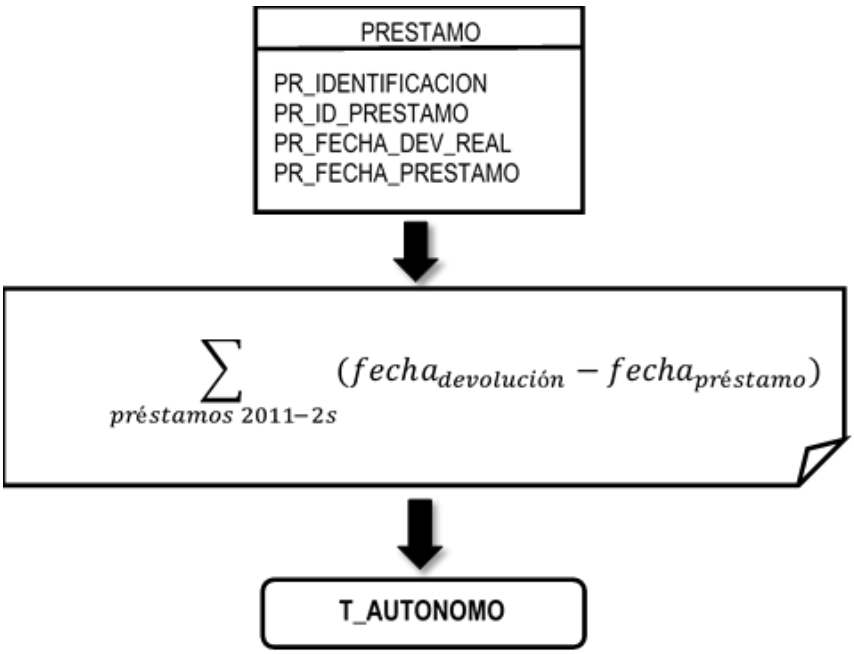


Figura 4.9. Pre-procesamiento de la variable Trabajo autónomo

4.3 DISEÑO GENERAL DEL MODELO DE DISCRIMINACIÓN

La estructura general del modelo para el análisis discriminante de este proyecto se lo ilustra en la Figura 4.10. Tal como se observa en la figura para predecir si el estudiante A es desertor el modelo utiliza 3 insumos:

- Los valores de las variables del mismo estudiante A.
- Los valores de las variables de algunos o de todos los estudiantes que se conocen que son desertores.
- Los valores de las variables de algunos o de todos los estudiantes que se conocen que no son desertores.

En estricto sentido, se puede decir que el modelo no responde si el estudiante es o no desertor; sino más bien responde un indicador de deserción que pudiera ser una probabilidad. En el caso de la regresión lineal el indicador de deserción no siempre puede ser interpretado como una probabilidad pues puede ser menor que cero o mayor que uno; sin embargo, en el caso de la regresión logística el indicador de deserción siempre es una probabilidad.

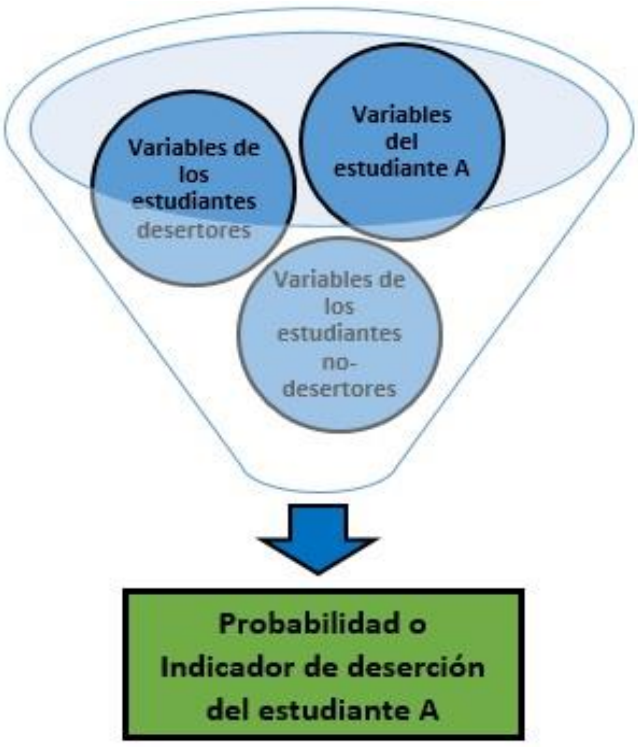


Figura 4.10. Esquema del modelo general de análisis discriminante

Sin embargo, algunas librerías estadísticas del software R no responden la probabilidad de deserción, sino que directamente etiquetan al

estudiante como desertor o no-desertor de acuerdo al valor de indicador de deserción.

CAPÍTULO 5

IMPLEMENTACIÓN Y PRUEBAS

5.1 ENTRENAMIENTO DE LOS MODELOS POR CADA TÉCNICA DE MINERÍA

5.1.1 Esquema de muestreo

Para la aplicación de las técnicas se segmentó el conjunto de datos en 2 grupos: conjunto de entrenamiento y conjunto de prueba, en donde el conjunto de entrenamiento representa aproximadamente el 70% de los registros y el conjunto de prueba el 30%.

Se tomaron 5 muestras diferentes y se aplicaron las técnicas a los conjuntos de entrenamiento de cada una de estas muestras.

Tal como se observa en la Figura 5.1 y en la Figura 5.2 las características de cada conjunto de entrenamiento y prueba cambian de una muestra a otra. En particular la distribución de desertores cambia de muestra a muestra tal como se observa en la Tabla 14.

```
 COD_ESTUDIANTE     SEXO          EDAD          FACTOR_P     RESIDENCIA
 Length:3006        F:1265   Min.   :16.00   Min.   : 0.00   LOCAL:2732
 Class :character   M:1741   1st Qu.:18.00   1st Qu.: 7.00   PROV : 274
 Mode  :character            Median :19.00   Median :10.00
                             Mean   :19.37   Mean   :11.82
                             3rd Qu.:20.00   3rd Qu.:15.00
                             Max.   :40.00   Max.   :40.00
   APROBADAS        PROMEDIO        REPROBADAS       ANTIGUEDAD
 Min.   : 1.00   Min.   :-1.000   Min.   : 0.000   Min.   :0.000
 1st Qu.:12.00   1st Qu.: 4.970   1st Qu.: 0.000   1st Qu.:1.000
 Median :17.00   Median : 5.970   Median : 1.000   Median :3.000
 Mean   :20.08   Mean   : 5.770   Mean   : 2.146   Mean   :2.489
 3rd Qu.:27.00   3rd Qu.: 6.758   3rd Qu.: 3.000   3rd Qu.:4.000
 Max.   :61.00   Max.   : 9.260   Max.   :15.000   Max.   :5.000
    PERDIDAS          SUPERADAS       DESERTOR     T_AUTONOMO
 Min.   :0.000000   Min.   :0.0000   NO:2627   Min.   :   0.00
 1st Qu.:0.000000   1st Qu.:0.0000   SI: 379   1st Qu.:   0.00
 Median :0.000000   Median :0.0000             Median :   0.00
 Mean   :0.004325   Mean   :0.1068             Mean   :  19.43
 3rd Qu.:0.000000   3rd Qu.:0.0000             3rd Qu.:  18.00
 Max.   :2.000000   Max.   :4.0000             Max.   :1663.00
```

Figura 5.1. Características del conjunto de entrenamiento de la muestra 1

```
 COD_ESTUDIANTE      SEXO          EDAD          FACTOR_P      RESIDENCIA
 Length:3006        F:1276   Min.   :16.00   Min.   : 0.00   LOCAL:2728
 Class :character   M:1730   1st Qu.:18.00   1st Qu.: 7.00   PROV : 278
 Mode  :character            Median :19.00   Median :10.00
                             Mean   :19.39   Mean   :11.86
                             3rd Qu.:20.00   3rd Qu.:15.00
                             Max.   :37.00   Max.   :40.00
   APROBADAS        PROMEDIO         REPROBADAS       ANTIGUEDAD
 Min.   : 1.00   Min.   :-1.000   Min.   : 0.000   Min.   :0.000
 1st Qu.:12.00   1st Qu.: 4.960   1st Qu.: 0.000   1st Qu.:1.000
 Median :17.00   Median : 5.935   Median : 1.000   Median :3.000
 Mean   :20.01   Mean   : 5.748   Mean   : 2.177   Mean   :2.475
 3rd Qu.:26.00   3rd Qu.: 6.720   3rd Qu.: 3.000   3rd Qu.:4.000
 Max.   :55.00   Max.   : 9.260   Max.   :15.000   Max.   :5.000
    PERDIDAS          SUPERADAS       DESERTOR    T_AUTONOMO
 Min.   :0.000000   Min.   :0.0000   NO:2631   Min.   :   0.0
 1st Qu.:0.000000   1st Qu.:0.0000   SI: 375   1st Qu.:   0.0
 Median :0.000000   Median :0.0000             Median :   0.0
 Mean   :0.005323   Mean   :0.1111             Mean   :  18.5
 3rd Qu.:0.000000   3rd Qu.:0.0000             3rd Qu.:  16.0
 Max.   :2.000000   Max.   :4.0000             Max.   :1663.0
```

Figura 5.2. Características del conjunto de entrenamiento de la muestra 2

Tabla 14. Distribución de desertores por muestra

Muestra #	Conjunto de entrenamiento	Conjunto de prueba	Total
1	379	146	525
2	375	150	525
3	363	162	525
4	365	160	525
5	378	147	525

5.1.2 Árboles de decisión

Al aplicar la técnica de árbol de decisión a los 5 grupos de entrenamiento se obtuvieron reglas diferentes para catalogar a un estudiante como desertor. Sin embargo, la base de estas reglas es siempre la misma. Las siguientes figuras muestran lo antes dicho (Figura 5.3, Figura 5.4, Figura 5.5, Figura 5.6 y Figura 5.7).

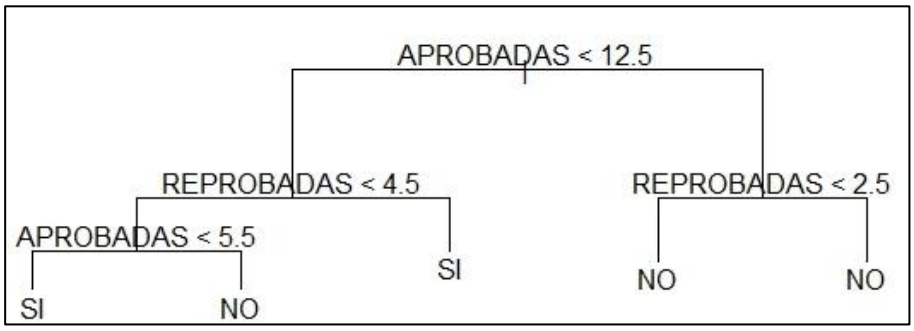


Figura 5.3. Árbol Muestra 1

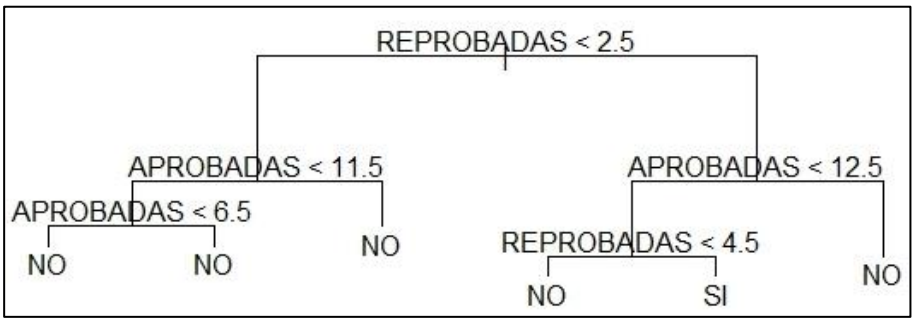


Figura 5.4. Árbol Muestra 2

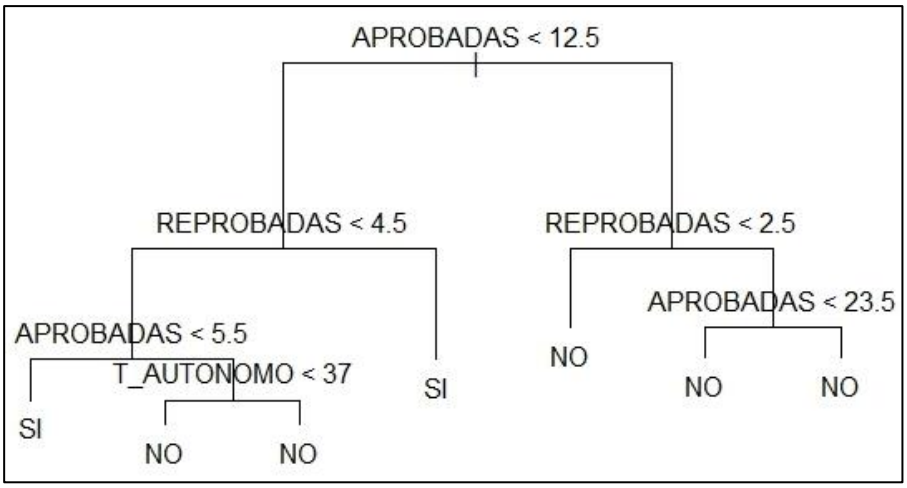


Figura 5.5. Árbol Muestra 3

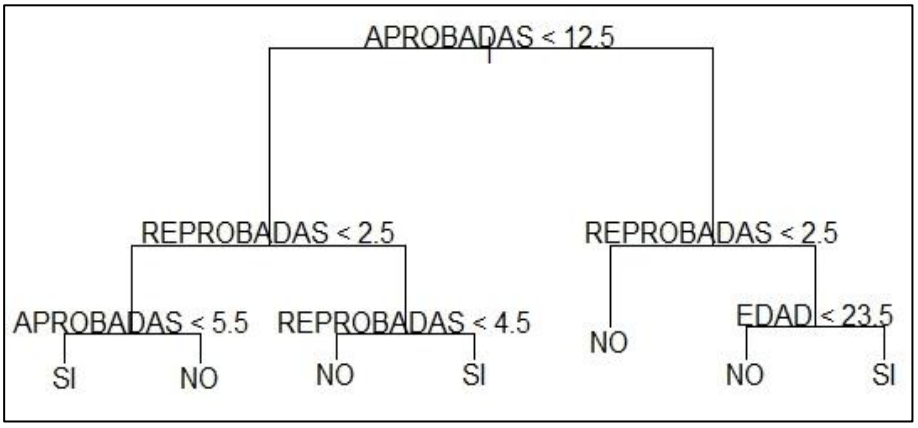


Figura 5.6. Árbol Muestra 4

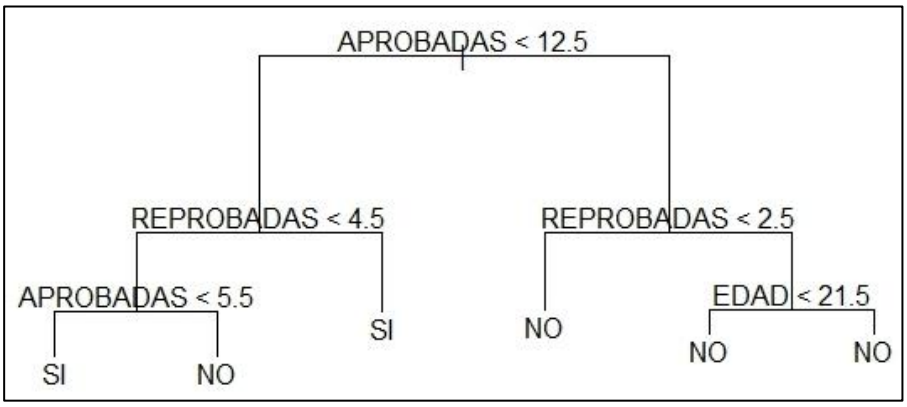


Figura 5.7. Árbol Muestra 5

Con el objetivo de encontrar la regla común que rige a los conjuntos de entrenamiento, adicional a las 5 muestras iniciales se generaron 5 muestras más.

Luego de la aplicación del método en los diferentes conjuntos de entrenamiento de las 10 muestras generadas, se simplificaron las reglas obtenidas; la Tabla 15 muestra un ejemplo de esta situación sobre la muestra 4 (compárese con la Figura 5.6).

Tabla 15. Reglas simplificadas obtenidas de la muestra 4

Regla 1	Si APROBADAS < 12.5 y REPROBADAS > 2.5 y REPROBADAS > 4.5 entonces ES DESERTOR
Regla 1 simplificada	Si APROBADAS < 12.5 y REPROBADAS > 4.5 entonces ES DESERTOR
Regla 2	Si APROBADAS < 12.5 y REPROBADAS < 2.5 y APROBADAS < 5.5 entonces ES DESERTOR
Regla 2 simplificada	Si REPROBADAS < 2.5 y APROBADAS < 5.5 entonces ES DESERTOR
Regla 3	Si APROBADAS > 12.5 y REPROBADAS > 2.5 y EDAD > 23.5 entonces ES DESERTOR
Regla 3 simplificada	No es posible simplificar

Considerando el total de reglas obtenidas en cada muestra es importante notar que algunas de estas reglas son incidentales, es decir, carecen de generalidad y están ajustadas al conjunto de entrenamiento del cual se generaron. Tomar en cuenta estas

reglas para el modelo final introduciría el efecto indeseable conocido como *overfitting* (sobreajuste del modelo).

En la Tabla 16 y Tabla 17 se observa un resumen de las reglas simplificadas que se obtuvieron de las 10 muestras en las que se aplicó el método.

Tabla 16. Reglas simplificadas obtenidas con el método del árbol

Regla 1	Si APROBADAS < 12.5 y REPROBADAS > 4.5 entonces ES DESERTOR
Regla 2	Si REPROBADAS < 4.5 y APROBADAS < 5.5 entonces ES DESERTOR
Regla 3	Si REPROBADAS < 2.5 y APROBADAS < 5.5 entonces ES DESERTOR
Regla 4	Si APROBADAS > 12.5 y REPROBADAS > 2.5 y EDAD > 23.5 entonces ES DESERTOR

Tabla 17. Reglas obtenidas por cada muestra

Muestra 1	Regla 1	Regla 2	
Muestra 2	Regla 1		
Muestra 3	Regla 1	Regla 2	
Muestra 4	Regla 1	Regla 3	Regla 4
Muestra 5	Regla 1	Regla 2	
Muestra 6	Regla 1		
Muestra 7	Regla 1		

Muestra 8	Regla 1		
Muestra 9	Regla 1		
Muestra 10	Regla 1	Regla 3	

De acuerdo a lo que se observa en la Tabla 17, la regla 1 (en verde) se repite en todas las muestras, siendo ésta la regla base que constituye el modelo general. La regla 2 (en amarillo) sólo aparece en 3 muestras, mientras que las reglas 3 y 4 se presentan en menos de 3 muestras.

De esta manera, el modelo resultante de esta técnica sería:

Si (APROBADAS < 12.5 y REPROBADAS > 4.5)
entonces ES DESERTOR
caso contrario NO ES DESERTOR

5.1.3 Regresión logística

Tal como se indicó en secciones anteriores, el entrenamiento en regresión logística consiste en la generación de una expresión que produzca la probabilidad de deserción de un estudiante dados los valores de sus características, de forma general la expresión es de la siguiente forma:

$$P(X \text{ es desertor}) = \frac{1}{1 + e^{-\sum \beta_k x_k}}$$

(5.1)

Siendo x_k la k-ésima característica del estudiante X y β_k una estimación de la influencia individual de dicha característica.

El reporte obtenido en R luego del entrenamiento con la primera muestra se observa en la Figura 5.8.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9148  -0.5105  -0.3360  -0.1696   3.2449

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.073171   0.565245  -7.206 5.76e-13 ***
SEXOM           0.083466   0.133207   0.627 0.530927
EDAD            0.202932   0.026429   7.678 1.61e-14 ***
FACTOR_P       -0.005980   0.009738  -0.614 0.539188
RESIDENCIAPROV -0.249330   0.228413  -1.092 0.275021
APROBADAS      -0.104725   0.011322  -9.250  < 2e-16 ***
PROMEDIO       -0.106775   0.058178  -1.835 0.066460 .
REPROBADAS      0.232576   0.026603   8.743  < 2e-16 ***
PERDIDAS        1.517568   0.643898   2.357 0.018431 *
SUPERADAS      -0.152527   0.143576  -1.062 0.288081
T_AUTONOMO     -0.006271   0.001876  -3.343 0.000828 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2277.8  on 3005  degrees of freedom
Residual deviance: 1812.6  on 2995  degrees of freedom
AIC: 1834.6

Number of Fisher Scoring iterations: 6
```

Figura 5.8. Reporte de Regresión Logística de la muestra 1

De acuerdo a la Figura 5.8 las variables que tienen influencia significativa en la probabilidad de deserción (denotados con ***) son: Edad, Aprobadas, Reprobadas y Trabajo Autónomo (T_Autónomo). El resto de variables no se las considera en la expresión de la muestra 1. De esta manera, la expresión resultante es la siguiente:

$$Prob = \frac{1}{1 + e^{-(-4.07 + 0.20EDAD - 0.10APR + 0.23REP - 0.006TAU)}} \quad \textbf{(5.2)}$$

En la Ecuación (5.2) se puede observar que la influencia individual de cada variable en la probabilidad de deserción varía no sólo en magnitud sino también en signo. Por ejemplo, a mayor edad mayor probabilidad de deserción y a mayor trabajo autónomo menor probabilidad de deserción; sin embargo, la edad afecta en mayor grado que el trabajo autónomo (véanse los coeficientes del reporte de la Figura 5.8).

Los coeficientes de las variables cambian de muestra a muestra. Tal como se observa en la Figura 5.9, el coeficiente de la variable Reprobadas varió en 0.04 pero mantuvo su alto nivel de significancia. En resumen, la significancia estadística se mantuvo para las mismas variables en casi todas las muestras.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0200  -0.5037  -0.3249  -0.1628   3.2139

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.249590   0.579386  -7.335 2.22e-13 ***
SEXOM           0.012140   0.134681   0.090 0.928175
EDAD            0.205872   0.027300   7.541 4.66e-14 ***
FACTOR_P        0.004239   0.009238   0.459 0.646313
RESIDENCIAPROV -0.438601   0.245165  -1.789 0.073615 .
APROBADAS      -0.105933   0.011574  -9.153  < 2e-16 ***
PROMEDIO       -0.109358   0.059933  -1.825 0.068049 .
REPROBADAS      0.273464   0.027301  10.017  < 2e-16 ***
PERDIDAS        0.119809   0.504985   0.237 0.812461
SUPERADAS      -0.344717   0.144622  -2.384 0.017145 *
T_AUTONOMO     -0.008893   0.002369  -3.754 0.000174 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2262.2  on 3005  degrees of freedom
Residual deviance: 1777.3  on 2995  degrees of freedom
AIC: 1799.3

Number of Fisher Scoring iterations: 6
```

Figura 5.9. Reporte de Regresión Logística de la muestra 2

Para generar el modelo resultante se promedia los coeficientes de las variables con alto nivel de significancia. Tomando los promedios de los coeficientes que se muestran en la Tabla 18 se obtiene el siguiente modelo resultante.

$$Prob = \frac{1}{1 + e^{-(-4.07+0.21EDAD-0.11APR+0.26REP-0.0096TAU)}} \quad \textbf{(5.3)}$$

Tabla 18. Promedio de las variables con alta significancia

Muestra #	Edad	Aprobadas	Reprobadas	Trabajo Autónomo
1	0.202932	-0.104725	0.232576	-0.006271
2	0.205872	-0.105933	0.273464	-0.008893
3	0.210054	-0.121198	0.267339	-0.009324
4	0.194807	-0.105821	0.286168	-0.009211
5	0.217949	-0.119258	0.266885	-0.008664
6	0.196164	-0.103633	0.244907	-0.013132
7	0.223350	-0.1048817	0.2478961	-0.009633
8	0.208814	-0.099934	0.246754	-0.011222
9	0.193777	-0.104858	0.255414	-0.009709
Promedio	0.205969	-0.107805	0.257934	-0.009562

5.2 VALIDACIÓN DE LOS MODELOS RESULTANTES

En esta sección se discutirá la validación de los modelos empleando la técnica Cross-Validation (validación cruzada) y una de sus variaciones conocida como Leave-One-Out Cross-Validation. Esta fase de validación constituirá un paso previo a la evaluación final de los modelos.

Aunque normalmente la validación se la ejecuta sólo en el conjunto de entrenamiento, en este proyecto podrá ejecutarse a los conjuntos de entrenamiento o al conjunto total. En cierta medida, servirá de comparación con la etapa de evaluación final.

Adicional a esto, la validación podría servir para afinar alguno de los métodos antes de su uso. En el caso particular de este proyecto, la validación del método de los k-vecinos más cercanos será clave para la determinación del parámetro K.

5.2.1 Esquema de muestreo para la validación cruzada

La validación cruzada requiere segmentar el conjunto de datos en partes aproximadamente iguales conocidas como folds [18]. La validación cruzada que se empleará en este proyecto será 10-fold. La Tabla 19 muestra una ilustración del esquema de muestreo. Cada fila corresponde a un fold, puesto que se trata de 4,294 registros, algunas divisiones tendrán 430 registros y otras 429. Es importante garantizar, si es necesario con muestreo estratificado, que la representatividad del grupo de deserción en cada fold sea cercana al porcentaje de desertores del grupo total que es 525/4,294 = 12.23%. Puesto que los porcentajes de desertores están en un rango aceptable no será necesario aplicar muestreo estratificado.

Tabla 19. Esquema de muestreo para validación cruzada

Número de Fold	Cantidad de registros	% de Desertores
1	429	12.82%
2	430	10.69%

Número de Fold	Cantidad de registros	% de Desertores
3	429	14.21%
4	430	11.86%
5	429	12.12%
6	430	13.48%
7	429	14.21%
8	430	10.69%
9	429	12.12%
10	429	10.02%

La validación cruzada consiste en un proceso de entrenamiento y prueba, donde tal como se muestra en la Tabla 20 uno de los folds se constituye en conjunto de prueba (en Amarillo) y los 9 restantes en conjunto de entrenamiento (en verde). Este proceso continuará hasta que todos los segmentos hayan sido utilizados como prueba para evaluar el método. En la tabla citada se observa el proceso completo y el porcentaje de detección obtenidos en cada una de las 10 iteraciones aplicando el método de Regresión Logística. No siempre los resultados son afortunados, por ejemplo, el porcentaje de detección más bajo (18.97%) se lo obtiene cuando el 6to. fold es considerado como conjunto de prueba.

Tabla 20. Aplicación de validación cruzada 10-fold con el método de Regresión Logística

	# de Fold										
# de Itera-ción	1	2	3	4	5	6	7	8	9	10	% Detec-ción
1	429	430	429	430	429	430	429	430	429	429	38.18%
2	429	430	429	430	429	430	429	430	429	429	36.96%
3	429	430	429	430	429	430	429	430	429	429	36.07%
4	429	430	429	430	429	430	429	430	429	429	23.53%
5	429	430	429	430	429	430	429	430	429	429	23.08%
6	429	430	429	430	429	430	429	430	429	429	18.97%
7	429	430	429	430	429	430	429	430	429	429	19.67%
8	429	430	429	430	429	430	429	430	429	429	23.91%
9	429	430	429	430	429	430	429	430	429	429	32.69%
10	429	430	429	430	429	430	429	430	429	429	32.56%
									Promedio		28.56%

Tal como se observa en la tabla anterior, la validación cruzada pudiera considerarse como preámbulo a la evaluación final de los modelos. En general, en los resultados de los modelos

podríamos esperar un porcentaje de detección similar al 28% para la Regresión Logística (promedio de la Tabla 20).

Aunque es muy utilizado el esquema 10-fold para la validación cruzada, el analista estará en libertad de cambiar el esquema propuesto e incluso prescindir de la validación si lo considera oportuno.

La Tabla 21 muestra un esquema similar al de la Tabla 20 pero aplicando el método de Naive Bayes. Tal como se muestra en esta nueva tabla, el porcentaje de detección de los distintos ensayos también se mantuvo entre el 19 y 40%, siendo su promedio 28.57% casi igual al método anterior.

Tabla 21. Aplicación de validación cruzada 10-fold con el método de Naive Bayes

	# de Fold										
# de Iteración	1	2	3	4	5	6	7	8	9	10	% Detección
1	429	430	429	430	429	430	429	430	429	429	36.36%
2	429	430	429	430	429	430	429	430	429	429	39.13%
3	429	430	429	430	429	430	429	430	429	429	27.87%
4	429	430	429	430	429	430	429	430	429	429	19.61%
5	429	430	429	430	429	430	429	430	429	429	28.85%

	# de Fold										
# de Itera-ción	1	2	3	4	5	6	7	8	9	10	% Detec-ción
6	429	430	429	430	429	430	429	430	429	429	29.31%
7	429	430	429	430	429	430	429	430	429	429	19.67%
8	429	430	429	430	429	430	429	430	429	429	23.91%
9	429	430	429	430	429	430	429	430	429	429	30.77%
10	429	430	429	430	429	430	429	430	429	429	30.23%
									Promedio		**28.57%**

5.2.2 Validación cruzada aplicada al método k-vecinos más cercanos

Como se mencionó en la sección anterior, la técnica de validación cruzada es especialmente útil para la determinación del valor idóneo de k. En resumen, se aplicará validación cruzada para distintos niveles de k y se escogerá aquel que produzca mayor porcentaje de detección.

Dos variantes se emplearán en este caso:

- Se ejecutará el proceso sobre un conjunto de entrenamiento de 3006 registros y no sobre todo el

conjunto 2011-2S, luego se contrastará los resultados con la evaluación final.

- Para evitar la necesidad de muestreo estratificado y aprovechando las capacidades computacionales del proyecto, se empleará una versión de la validación cruzada conocida como Leave-One-Out Cross-Validation [18], la cual pudiera ser interpretada en este caso como 3,006-fold.

Ahora esta nueva versión de validación cruzada arrojará 3,006 porcentajes de detección, uno por cada ensayo, donde el fold de prueba de cada ensayo lo conformará un solo estudiante. En cada ensayo se aplicará el método de k-vecinos más cercanos con un k determinado y se obtendrá el promedio de los 3006 ensayos. La Tabla 22 muestra los resultados de la aplicación de estos ensayos por cada valor de k, los 2 valores más altos (en Amarillo) ocurren en k=1 y k=2. El porcentaje de detección cae considerablemente luego de k=2 y tal como lo muestra la Figura 5.10 la tendencia se mantiene a la baja.

Tabla 22. Promedio del % de detección de Knn para distintos valores de k

Valor K	1	2	3	4	5	6
% Detección	28.50	29.55	19.53	20.05	14.25	12.14
Valor K	7	8	9	10	11	12
% Detección	12.93	12.93	13.19	10.82	11.08	10.03
Valor K	13	14	15	16	17	18
% Detección	9.50	8.71	9.5	9.5	8.71	8.97
Valor K	19	20	21	22	23	24
% Detección	8.44	7.92	6.86	6.86	6.60	6.33

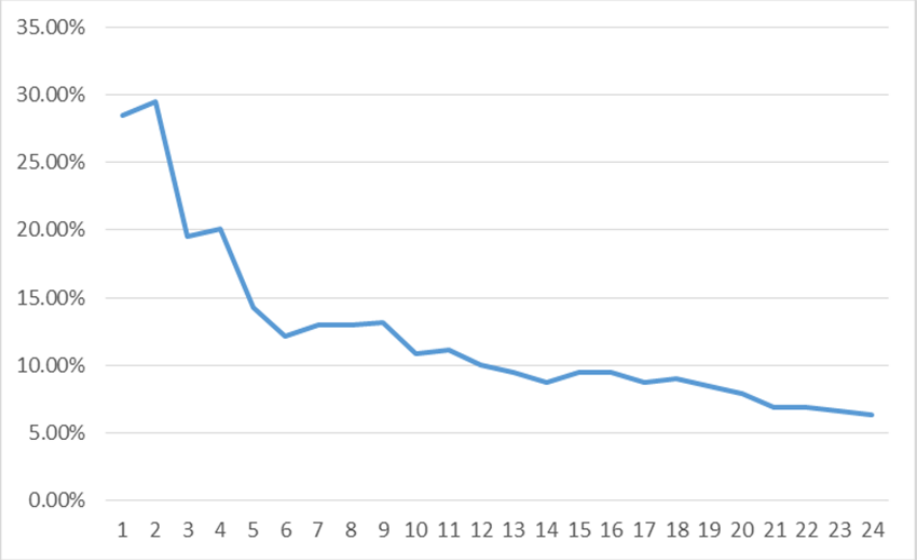


Figura 5.10. Promedio del % de Detección de K-NN para distintos valores de K

En la Tabla 22 y la Figura 5.10, se observa que el k que produce el mayor porcentaje de detección es k=2. Sin embargo, no es conveniente elegir este valor, pues en caso de existir un vecino desertor y otro no-desertor no se podría determinar la clase a la que pertenece el estudiante que se requiere clasificar. Por lo tanto, el k idóneo para este trabajo será k=1, con la ventaja adicional de que el costo computacional será mucho menor.

5.3 GENERACIÓN DE LISTADOS DE ESTUDIANTES CLASIFICADOS SEGÚN SU RIESGO DE DESERCIÓN

Uno de los objetivos de este trabajo es clasificar a los actuales estudiantes de acuerdo a su riesgo de deserción. Para este propósito se empleará parte de los modelos ya entrenados en la sección anterior.

En esta sección se mostrarán los resultados de aplicar los métodos al conjunto de estudiantes del segundo semestre del año 2017. La Tabla 23 y la Tabla 24 muestran un resumen estadístico de este conjunto de datos formado por 9,159 estudiantes.

Tabla 23. Resumen estadístico de estudiantes 2017_2S (variables numéricas)

Variable	Edad	Fact P	Aprob	Reprob	Prom	Antig	Perd	Sup	Trab Aut
Mínimo	16	0	0	0	0	0	0	0	0
Máximo	44	40	96	37	9.81	17	3	7	300
Promedio	21.24	11	35.79	5.06	5.58	5.49	0.04	0.30	11.02

Fact P: Factor P Aprob: Aprobadas Reprob: Reprobadas Prom: Promedio
Antig: Antigüedad Perd: Perdidas Sup: Superadas Trab Aut: Trabajo autónomo

Tabla 24. Resumen estadístico de estudiantes 2017_2S (variables categóricas)

Variable	Sexo		Residencia	
	Femenino	3,620	**Local**	7,787
	Masculino	5,539	**Provincia**	1,372

Se clasificará a cada estudiante como potencial desertor o no desertor, teniendo presente que para este proyecto es preferible un falso positivo a un falso negativo.

5.3.1 Aplicación de árbol de decisión y regresión logística

Tal como se indicó en secciones anteriores, los métodos árbol de decisión y regresión logística generan un modelo de forma

anticipada, modelo que es comprensible ya sea en forma de reglas en el caso del método de árbol o una ecuación en el caso de regresión logística. En esta sección se mostrarán los resultados de la aplicación de estos modelos al conjunto de datos 2017-2S.

En la aplicación de este método de árbol de decisión se considera la regla final obtenida en la sección 5.1.2. La Figura 5.11 muestra los resultados de la aplicación de esta regla.

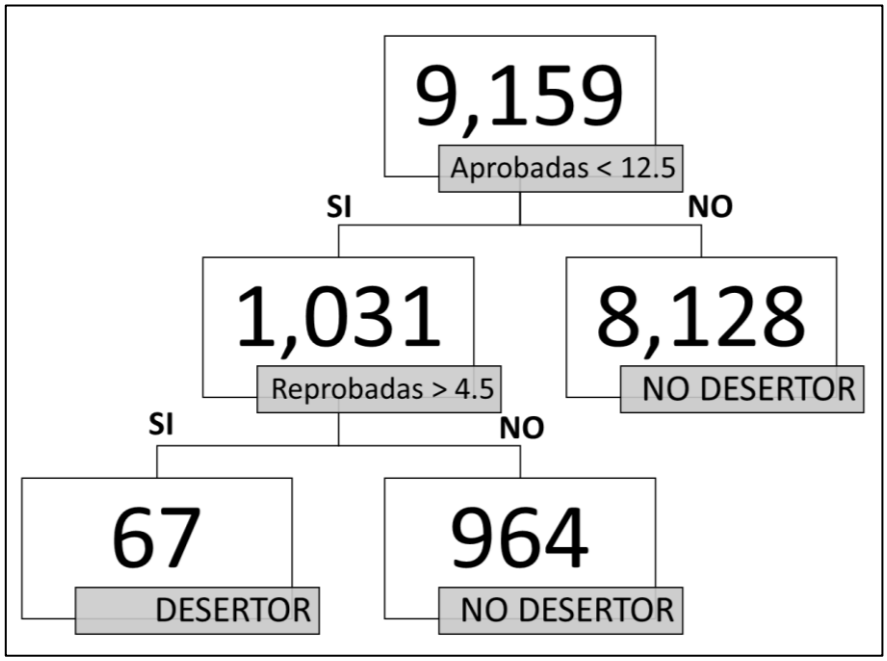


Figura 5.11. Aplicación de regla de método de árbol al conjunto de datos 2017-2S

La Tabla 25 y Tabla 26 muestran un resumen estadístico de los 67 estudiantes detectados como desertores por el método de árbol. Comparando estas tablas con los resúmenes de la Tabla 23 y Tabla 24 referentes al total de estudiantes, se puede indicar que existen ciertas diferencias porcentuales. Por ejemplo, el porcentaje de varones cambia del 60% en el conjunto completo (Tabla 24) al 68% en el conjunto de desertores (Tabla 26). Similar análisis se pudiera realizar con otras variables como Residencia; sin embargo, ninguna de estas variables supera en su capacidad de discriminación a las elegidas por el árbol de decisión: Aprobadas y Reprobadas.

Tabla 25. Resumen estadístico de estudiantes detectados como desertores por el método de árbol (variables numéricas)

Variable	Mínimo	Máximo	Promedio
Edad	18	29	20.87
Factor P	0	22	1.21
Aprobadas	2	12	8.64
Reprobadas	5	20	7.10
Promedio	1.56	6.42	3.71
Antigüedad	0	17	2.68
Perdidas	0	1	0.12
Superadas	0	1	0.15
Trabajo Autónomo	0	176	24.46

Tabla 26. Resumen estadístico de estudiantes detectados como desertores por el método de árbol (variables categóricas)

Variable	Sexo		Residencia	
	Femenino	21	**Local**	52
	Masculino	46	**Provincia**	15

El método de árbol de decisión arroja apenas 67 estudiantes como posibles desertores, que constituye menos del 1% de la población total; no obstante, es importante considerar el presente proyecto de minería de datos bajo los 4 métodos propuestos, puesto que la capacidad de los métodos para predecir varía y es complementaria. A este grupo de 67 potenciales desertores habrá que agregar aquellos detectados por los demás métodos y realizar un análisis integral.

En el método de Regresión Logística se aplicará la ecuación obtenida en el entrenamiento. Puesto que se requiere la mayor precisión posible, no se desechará ninguna de las variables que participan en la ecuación. La Tabla 27 muestra los resultados del cálculo para algunos de los registros del 2017, donde la columna LOGIT representa la probabilidad de deserción del estudiante. En la mayoría de los casos de la muestra la probabilidad permanece por debajo de 0.3.

Tabla 27. Muestra de los resultados de Regresión Logística 2017-2S

SEXO	EDAD	FP	RES	APR	PROM	REP	PER	SUP	TA	LOGIT
F	21	3	LOC	41	5.72	4	0	0	5	0.0143
F	22	9	LOC	33	4.35	11	0	0	0	0.2874
M	18	14	PRV	7	3.34	1	0	0	17	0.1798
M	24	28	LOC	52	5.91	14	0	2	59	0.0401
M	24	27	LOC	60	5.82	16	0	1	0	0.0600
M	22	18	LOC	31	3.49	23	0	2	131	0.7258
F	22	1	LOC	50	5.41	8	0	0	0	0.0190
M	30	0	PRV	34	4.27	8	0	1	63	0.2336
M	21	12	PRV	23	3.77	11	0	0	0	0.4916
M	23	24	PRV	42	6.13	2	0	0	0	0.0105

FP: Factor P Res: Residencia Apr: Aprobadas Prom: Promedio
Per: Perdidas Sup: Superadas TA: Trabajo autónomo

En Regresión Logística se considera que si la probabilidad de desertar es mayor que 0.5 entonces el estudiante será catalogado como potencial desertor. Sin embargo, este umbral (máxima probabilidad permisible para no ser considerado potencial desertor) de 0.5 es discrecional y puede aumentar o disminuir de acuerdo al problema que se esté abordando. La Tabla 28 muestra la cantidad de potenciales desertores para diferentes valores del umbral.

Tabla 28. Cantidad de potenciales desertores según la máxima probabilidad permisible

Umbral	Cantidad de potenciales desertores
0.1	3,022
0.2	1,719
0.3	1,181
0.4	900
0.5	682

En un proyecto como éste en el cual los falsos positivos son preferibles a los falsos negativos se está tentado a considerar el umbral más pequeño posible, sin embargo, no se pueden comprometer demasiados recursos a costa de esta excesiva precaución. Tal como se muestra en la Tabla 28 (en amarillo), el umbral 0.4 es una solución de compromiso que no compromete demasiados recursos (900 < 10%Total de Estudiantes) pero que deja una holgura de probabilidad de 0.1.

Como se mencionó en el diseño de la solución (Capítulo 4), la estrategia debe ser integral. Al emplear los 2 métodos, se obtienen los resultados detallados en la Figura 5.12. Esta figura refleja que existen 50 estudiantes detectados por los 2 métodos, la probabilidad de que éstos sean desertores aumenta considerablemente.

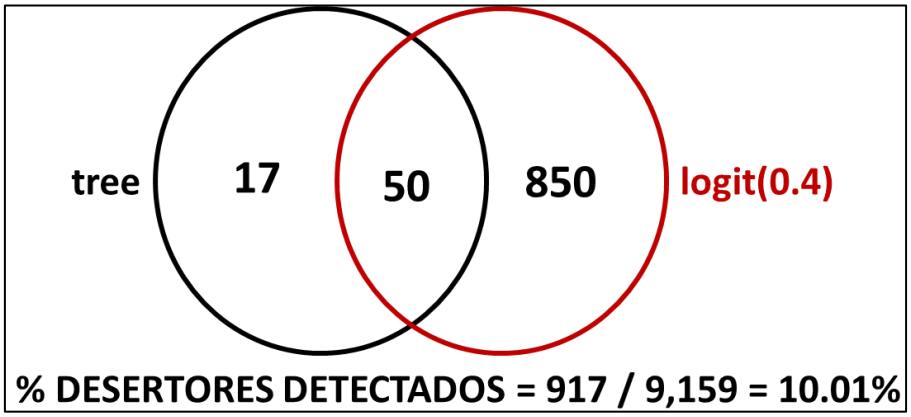


Figura 5.12. Posibles desertores 2017-2S según árbol de Decisión y Regresión Logística

5.3.2 Aplicación de Naive Bayes y K-vecinos más cercanos

Los métodos que se aplicarán en esta sección no tienen ningún modelo explícito a la manera de reglas o ecuaciones. Estos métodos casi no requieren entrenamiento previo y ejecutan la mayor cantidad de sus procesos en el momento en que son consultados por la discriminación de un nuevo elemento.

Inicialmente se aplica el método de Naive Bayes al conjunto de datos 2017-2S. En la Tabla 29 se observa una muestra con los resultados tanto de regresión logística como Naive Bayes. En esta tabla se observa (en Amarillo) estudiantes detectados como desertores por un método y obviados por el otro; esto es razonable porque estos métodos se sostienen sobre diferente

base teórica. El método de Naive se basa directamente en el teorema de Bayes para ejecutar su predicción.

Tabla 29. Muestra 2017-2S con predicciones de deserción realizadas por Regresión Logística y Naive Bayes

SEXO	EDAD	FP	RES	APR	PROM	REP	ANT	TA	LOGIT	NB
F	24	1	PROV	51	7.3	2	11	0	0.0032	NO
M	22	12	LOCAL	43	4.78	5	8	2	0.0228	NO
M	20	4	PROV	39	5.83	4	5	17	0.0106	NO
F	34	7	LOCAL	55	6.99	7	9	0	0.0628	SI
M	22	7	LOCAL	38	4.52	17	9	0	0.2974	SI
M	20	6	PROV	29	5.57	2	3	7	0.0232	NO
F	20	6	PROV	29	6.46	0	2	0	0.0127	NO
F	20	5	LOCAL	23	5.24	5	2	0	0.1288	NO
F	20	17	PROV	31	4.71	10	5	0	0.1696	NO
M	22	9	LOCAL	24	4.95	6	4	0	0.5606	SI
M	23	7	LOCAL	42	5	19	10	0	0.4897	SI
M	23	8	LOCAL	16	5.43	7	11	2	0.5222	NO

FP: Factor P Res: Residencia Apr: Aprobadas Prom: Promedio
Rep: Reprobadas Ant: Antigüedad TA: Trabajo autónomo

Este nuevo método que se agrega a los ya aplicados producen en total 1,716 posibles desertores. La Figura 5.13 muestra la distribución de estos estudiantes. Tal como se observa en la figura, Naive Bayes ha aumentado la detección de 10.01% a 18.73%. Se puede observar también (zona sombreada) que sólo Naive Bayes detecta un poco menos de la mitad de los

desertores. Nuevamente en la figura se observa que 29 desertores son detectados por los 3 métodos, lo cual aumenta la probabilidad de que el evento de deserción efectivamente ocurra.

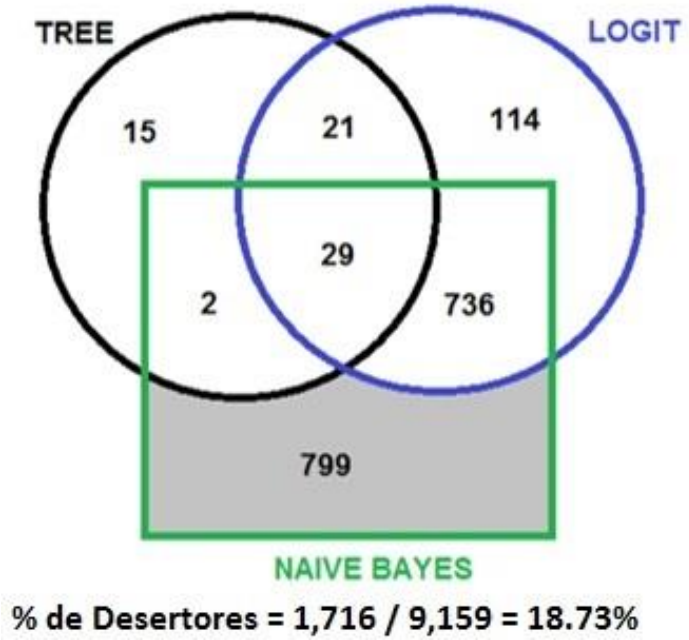


Figura 5.13. Posibles desertores 2017-2S según árbol de Decisión, Regresión Logística y Naive Bayes

Finalmente, para completar la predicción es necesario incluir el método de los k-vecinos más cercanos. En la sección anterior se concluyó que el mejor k para este problema es k=1.

La Tabla 30 muestra unos registros clasificados de acuerdo a los últimos 3 métodos. En esta tabla se observa que el estudiante del 3er registro (en Amarillo) es clasificado como desertor por

Regresión Logística (p>0.4) y por KNN-1 pero para Naive Bayes no es un potencial desertor.

Como ya se ha dicho antes, las predicciones de los métodos con respecto al mismo estudiante pueden variar y la coincidencia de más de un método en el mismo estudiante refuerza la idea de la futura deserción. En este sentido, es más probable que el estudiante del 3er. registro (en amarillo) deserte a que lo haga el estudiante del 4to. Registro (en azul), ya que este último es detectado por solo un método, mientras que el primero es detectado por dos.

Tabla 30. Muestra de estudiantes clasificados según Regresión Logística, Naive Bayes y KNN1

SEX	EDAD	FP	RES	APR	PRO	REP	TA	LOGIT	NB	KNN1
M	22	4	LOC	2	3.46	3	0	0.6811	NO	SI
M	20	12	LOC	6	5.07	7	0	0.6735	NO	NO
M	23	1	LOC	4	6.34	5	17	0.6627	NO	SI
F	23	4	LOC	6	5.23	5	12	0.6594	NO	NO
F	23	0	LOC	37	4.77	10	92	0.6558	NO	NO
F	34	7	LOC	55	6.99	7	0	0.0628	SI	NO
M	22	7	LOC	38	4.52	17	0	0.2974	SI	NO
M	20	6	PRO	29	5.57	2	7	0.0232	NO	NO
M	28	0	PRO	31	5.81	3	0	0.1214	SI	NO
M	20	8	LOC	28	5.5	5	0	0.0742	NO	NO

La contribución del método 1-vecino más cercano al proyecto de detección es significativa. Tal como lo muestra la Figura 5.14, con la aplicación de este nuevo método, 311 estudiantes son agregados como posibles desertores, logrando aumentar el porcentaje de detección del proyecto de 18.73% a 22.13%. Adicionalmente puede observarse que 17 estudiantes son detectados como desertores por todos los métodos, siendo este un grupo de alto riesgo, 203 estudiantes son detectados por 3 métodos y 667 estudiantes son detectados por 2 métodos.

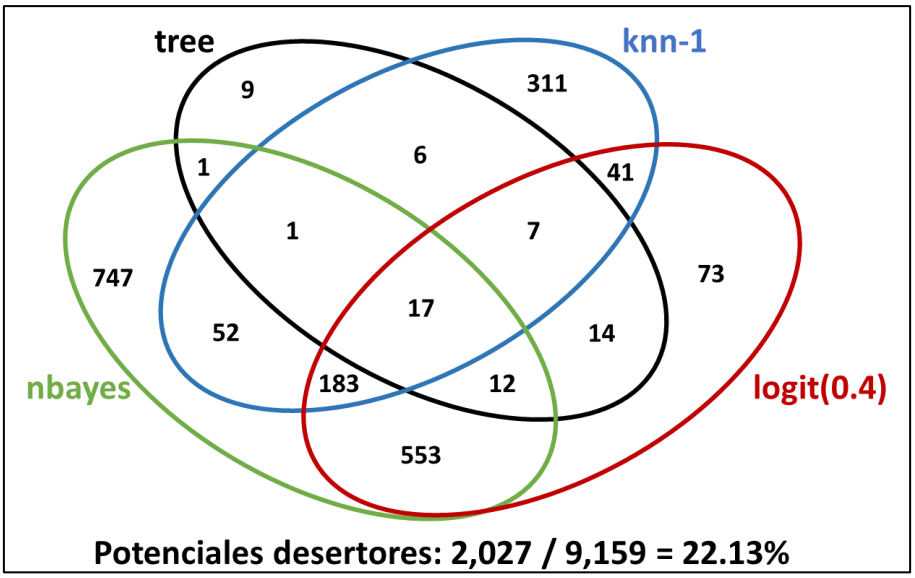


Figura 5.14. Diagrama de Venn de posibles desertores 2017-2S con todos los métodos

CAPÍTULO 6

ANÁLISIS DE RESULTADOS

6.1 ANÁLISIS DE LAS CARACTERÍSTICAS DE CADA MODELO RESULTANTE

En esta sección se presentarán los resultados de evaluar cada método con los 5 conjuntos de prueba. Al evaluar cada método se podrá obtener:

- El porcentaje de error por mala clasificación.
- La cantidad de falsos positivos.
- La cantidad de falsos negativos.
- El porcentaje de detección del método.

6.1.1 Evaluación del método de Árbol de Decisión

Para evaluar la efectividad del método al discriminar entre un desertor y un no-desertor se empleará la regla del modelo resultante de la sección 5.1.2:

> Si (APROBADAS < 12.5 y REPROBADAS > 4.5)
> entonces ES DESERTOR
> caso contrario NO ES DESERTOR

La Tabla 31 muestra los resultados de aplicar esta regla al primer conjunto de prueba. En la tabla se contrasta la realidad con la predicción del método. El método calificará como desertor a aquel estudiante que cumpla la condición de la regla. Se observa en la tabla que en total 54 estudiantes cumplen la regla, pero sólo 38 de ellos son en realidad desertores.

Tabla 31. Resultados de la evaluación del método de árbol en la muestra 1

	Regla		TOTAL
Desertor	**Cumple**	**No cumple**	
SI	38	108	146
NO	16	1,126	1,142
TOTAL	54	1,234	1,288
		% de Detección	26.03%
		Falsos Negativos	73.97%
		Falsos Positivos	1.40%
		% de Error al clasificar	9.63%

Las celdas rojas de la Tabla 31 indican los errores del método. Por ejemplo, 16 estudiantes fueron calificados como desertores por el método (cumplieron la regla) pero en realidad no son desertores; es decir, son falsos positivos (falsos desertores). De igual manera, 108 estudiantes no cumplieron la regla, sin embargo, realmente son desertores; es decir, son falsos negativos. En cambio, las celdas azules indican los aciertos del método.

El error al clasificar se lo obtiene considerando todas las malas clasificaciones (celdas rojas), el cual para este caso queda (16 + 108) /1,288 = 9.63%. Aunque este último es un aceptable error de prueba para un método de discriminación, el interés de este

proyecto no está sólo en clasificar bien sino en detectar la mayor cantidad posible de reales desertores. Tal como se observa en la Tabla 31, de 146 desertores el método detectó 38, es decir, el porcentaje de detección sobre esta muestra es 38/146 = 26.03%

Al cambiar el conjunto de prueba, muy probablemente cambiará tanto el porcentaje de error al clasificar como el porcentaje de detección del método. Véase la Tabla 32, en la cual se muestran los resultados del método sobre otro conjunto de prueba (muestra 3). Tal como se observa, los resultados en este caso son menos afortunados, el porcentaje de error subió y el porcentaje de detección bajó, sin duda una mala muestra.

Tabla 32. Resultados de la evaluación del método de árbol en la muestra 3

	Regla		TOTAL
Desertor	**Cumple**	**No cumple**	
SI	24	138	162
NO	27	1,099	1,126
TOTAL	51	1,237	1,288
		% de Detección	14.81%
		Falsos Negativos	85.19%
		Falsos Positivos	2.40%
		% de Error al clasificar	12.81%

Una de las razones para evaluar el método en varios conjuntos de prueba, es para reducir los efectos adversos de una mala (o buena) muestra que distorsione la realidad. En la Tabla 33 se muestra un resumen de la aplicación del método a los 5 conjuntos de prueba. En esta tabla, se nota que la Tabla 31 y la Tabla 32 eran el caso afortunado y desafortunado respectivamente. Una estimación más realista del comportamiento del método es el promedio de los resultados individuales con cada muestra. Tal como se observa en la Tabla 33, el error promedio al clasificar en el conjunto de prueba es 11.23%, el cual usualmente es siempre mayor que el error de entrenamiento. Así mismo, el porcentaje promedio de detección es 19.79%

Tabla 33. Resumen del método de árbol en los 5 conjuntos de prueba

# de Muestra	Error de entrenamiento	Error de prueba	% de Detección
1	11.34%	9.63%	26.03%
2	11.21%	11.10%	20.00%
3	10.15%	12.81%	14.81%
4	10.58%	11.65%	19.38%
5	10.75%	10.95%	18.73%
Promedio	**10.81%**	**11.23%**	**19.79%**

6.1.2 Evaluación del método de K-Vecinos más cercanos

Este método necesita el parámetro k para su ejecución, en la sección 5.2 se concluyó que el k idóneo para este proyecto es k=1. En todo caso, en esta sección se mostrarán los resultados de la aplicación del método a los 5 conjuntos de prueba para distintos niveles de k.

En la Tabla 34 se muestran los resultados de la aplicación del método al primer conjunto de prueba con k=1, tal como se observa el porcentaje de error al clasificar subió en relación al método anterior (véase la Tabla 31) pero también lo hizo el porcentaje de detección. Lo que ha ocurrido es que este método ha disminuido los falsos negativos (de 138 a 102) a costa de subir los falsos positivos (de 16 a 83); pero el sacrificio vale la pena: el porcentaje de detección se incrementó de 26% a 30%.

Tabla 34. Resultados de la evaluación del método 1-vecino más cercano en la muestra 1

	Predicción		TOTAL
Desertor	**SI**	**NO**	
SI	44	102	146
NO	83	1,059	1,142
TOTAL	127	1,161	1,288
		% de Detección	30.14%
		Falsos Negativos	69.86%
		Falsos Positivos	7.27%
		% de Error al clasificar	14.36%

En la Tabla 35 se observan los resultados de 1-vecino más cercano aplicados a la muestra 3. Comparando estos resultados con los obtenidos con el método de árbol (véase la Tabla 32) se puede concluir que la muestra que es mala para un método no necesariamente lo es para otro. Con el método de árbol se obtuvo apenas el 14.81% como porcentaje de detección; sin embargo, con 1-vecino más cercano el porcentaje de detección es mucho mejor: 30.25%.

Tabla 35. Resultados de la evaluación del método 1-vecino más cercano en la muestra 3

	Predicción		TOTAL
Desertor	**SI**	**NO**	
SI	49	113	162
NO	108	1,018	1,126
TOTAL	157	1,131	1,288
		% de Detección	30.25%
		Falsos Negativos	69.75%
		Falsos Positivos	9.59%
		% de Error al clasificar	17.16%

La Tabla 36 muestra un resumen del porcentaje de detección obtenido al aplicar k-vecinos más cercanos a los 5 conjuntos de prueba. Las celdas azules representan el mejor resultado para cada nivel de k, mientras que las celdas rojas representan el peor resultado por cada valor de k. Tal como se observa, la muestra 4 fue la mejor para 1-vecino más cercano, pero no lo fue para los demás valores de k. Esto confirma lo antes mencionado: una “buena muestra” para un método no necesariamente lo es para otro, incluso en métodos tan similares como 1-NN y 3-NN.

Tabla 36. Resumen del porcentaje de detección del método k-vecinos más cercanos

Métodos	Knn1	Knn3	Knn5	Knn9	Knn11
Muestra 1	30.14%	28.08%	22.60%	16.44%	16.44%
Muestra 2	33.33%	31.33%	23.33%	20.67%	18.67%
Muestra 3	30.25%	20.99%	17.28%	16.05%	14.81%
Muestra 4	34.38%	25.00%	19.38%	16.88%	14.38%
Muestra 5	25.85%	22.45%	19.73%	17.01%	14.97%
Promedio	30.79%	25.57%	20.46%	17.41%	15.85%

La Figura 6.1 muestra una descripción gráfica de la Tabla 36, en esta figura se observa claramente que a mayor valor k menor porcentaje de detección. La figura confirma lo mencionado en la sección 5.2.2, el valor de k que produce los mejores resultados es k=1, produciendo en promedio un porcentaje de detección de 30.79% (véase la Tabla 36). Nótese además que el mejor porcentaje promedio obtenido aquí no dista mucho de aquel obtenido en la sección 5.2.2 (véase la Tabla 22).

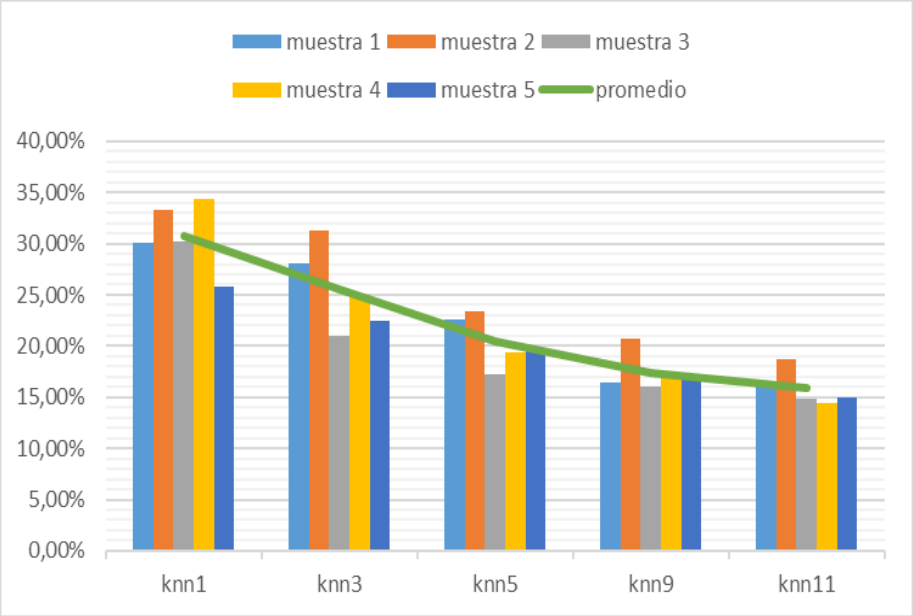


Figura 6.1. Porcentaje de detección de k-vecinos más cercanos para distintos valores de k

6.1.3 Evaluación del método de Naive Bayes

A pesar de la débil suposición de independencia entre las variables, la sobresaliente capacidad de predicción del método Naive Bayes se refleja en la Tabla 37; nótese como se detecta casi el 40% de los desertores, el valor más alto encontrado hasta ahora. Aunque no es de interés primario para el proyecto, el error al clasificar de 12.66% es bastante aceptable.

Tabla 37. Resultados de la evaluación del método de Naive Bayes en la muestra 1

	Predicción		TOTAL
Desertor	**SI**	**NO**	
SI	57	89	146
NO	74	1,068	1,142
TOTAL	131	1,157	1,288
		% de Detección	39.04%
		Falsos Negativos	60.96%
		Falsos Positivos	6.48%
		% de Error al clasificar	12.66%

El método de Bayes se comporta de forma aceptable incluso con una “mala” muestra, su peor resultado lo refleja la Tabla 38. En esta tabla se observa un porcentaje de detección del 26.54%, este resultado supera a los obtenidos en los peores escenarios de Knn (compárese con la Tabla 36).

Tabla 38. Resultados de la evaluación del método de Naive Bayes en la muestra 3

	Predicción		TOTAL
Desertor	**SI**	**NO**	
SI	43	119	162
NO	78	1,048	1,126
TOTAL	121	1,167	1,288
		% de Detección	26.54%
		Falsos Negativos	73.46%
		Falsos Positivos	6.93%
		% de Error al clasificar	15.30%

Un resumen del método para las 5 muestras se encuentra en la Tabla 39, el mejor y peor resultado se resaltan en la celda azul y roja respectivamente. En este método, el porcentaje promedio de detección bordea el 31% ligeramente mejor que el porcentaje promedio de Knn.

Tabla 39. Porcentaje de detección de Naive Bayes por cada muestra

# de Muestra	% de Detección
1	39.04%
2	34.67%
3	26.54%
4	27.50%
5	27.21%
Promedio	**30.99%**

6.1.4 Evaluación del método de Regresión Logística

La discriminación en Regresión Logística depende de un valor discrecional (umbral) que normalmente es 0.5, pero puede ajustarse de acuerdo al proyecto. En la sección 5.3.1 se justificó el uso de 0.4 como umbral o máxima probabilidad permisible para este proyecto.

La Tabla 40 muestra los resultados de la aplicación de Regresión Logística a la muestra 1 con un umbral de 0.4. Este método presenta muy buenos resultados comparados con los casos anteriores, no sólo se obtuvo casi el 40% de detección, sino que el porcentaje de error al clasificar se redujo de 12.66% en Naive Bayes a 9.24% en este método.

Tabla 40. Resultados de la evaluación de Regresión Logística en la muestra 1, umbral 0.4

	Predicción		TOTAL
Desertor	**SI**	**NO**	
SI	57	89	146
NO	30	1,112	1,142
TOTAL	87	1,201	1,288
		% de Detección	39.04%
		Falsos Negativos	60.96%
		Falsos Positivos	2.63%
		% de Error al clasificar	9.24%

En la Tabla 41 se observan los resultados del método para la muestra 3 nuevamente con un umbral de 0.4. En este caso los resultados no son tan buenos como los obtenidos en la muestra 1; tanto los falsos negativos como los falsos positivos aumentaron provocando un descenso en el porcentaje de detección y un aumento en el error de prueba (porcentaje de error al clasificar).

Tabla 41. Resultados de la evaluación de Regresión Logística en la muestra 3, umbral 0.4

	Predicción		TOTAL
Desertor	**SI**	**NO**	
SI	44	118	162
NO	43	1,083	1,126
TOTAL	87	1,201	1,288
		% de Detección	27.16%
		Falsos Negativos	72.84%
		Falsos Positivos	3.82%
		% de Error al clasificar	12.50%

A diferencia de los cambios circunstanciales debido al cambio de la muestra, el comportamiento de la Regresión Logística cambia estructuralmente cuando se varía el umbral. En este sentido, la Tabla 42 muestra los resultados de aplicar este método a la muestra 1 con distintos valores de umbral.

Tabla 42. Resultados de Regresión Logística aplicado a la muestra 1

Umbral	% Falsos negativos	% Falsos positivos	% Detección	% de Error
0.2	36.30%	14.62%	63.70%	17.08%
0.3	52.05%	5.95%	47.95%	11.18%
0.4	60.96%	2.63%	39.04%	9.24%
0.5	75.34%	1.58%	24.66%	9.94%
0.6	85.62%	0.88%	14.38%	10.48%
0.7	91.78%	0.53%	8.22%	10.87%
0.8	95.89%	0.26%	4.11%	11.10%

Como se observa en la Tabla 42, el menor porcentaje de error se lo obtiene con el umbral antes escogido de 0.4 (véase la Tabla 40).

La Figura 6.2 representa gráficamente la Tabla 42, en este gráfico se puede apreciar el comportamiento del modelo a distintos valores de umbral. Como es de esperarse, a menor umbral mayor porcentaje de detección, pero así mismo mayor porcentaje de falsos positivos.

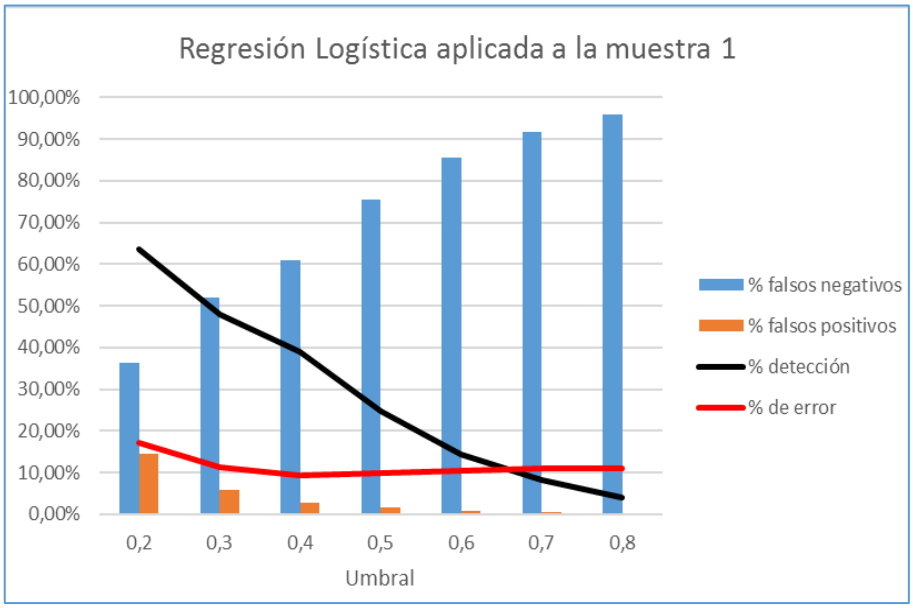


Figura 6.2. Tendencias de los principales indicadores de Regresión Logística, muestra 1

Un enfoque para encontrar el valor idóneo del umbral sería obtener el mayor porcentaje de detección posible sin que el porcentaje de falsos positivos supere el 10%. Esa solución se la encuentra cuando el umbral es 0.3 con porcentaje de detección 47.95% y porcentaje de falsos positivos 5.95% (celdas en amarillo en la Tabla 42).

Aunque se desea el mayor porcentaje de detección posible, también se debe considerar que se deberá emplear recursos de la universidad en cada posible desertor detectado. Teniendo presente el uso mesurado de estos recursos, otro enfoque para encontrar el valor idóneo del umbral sería que el número de

estudiantes detectados como posibles desertores no supere el 10% del total. La Tabla 43 muestra que esto no se cumple con el umbral 0.3, pues los posibles desertores según el método son 138, que representan el 138/1,288 = 10.71% del total. Sin embargo, el umbral 0.4 sí se adapta a este enfoque, en la Tabla 40 se observa que los posibles desertores son 87/1,288 = 6.75% que es menor que el 10% del total. Esto último coincide con el análisis propuesto en la sección 5.3.1 en donde se escogió el umbral 0.4.

Tabla 43. Resultados de la evaluación de Regresión Logística en la muestra 1, umbral 0.3

	Predicción		TOTAL
Desertor	**SI**	**NO**	
SI	70	76	146
NO	68	1,074	1,142
TOTAL	138	1,150	1,288
		% de Detección	47.95%
		Falsos Negativos	52.05%
		Falsos Positivos	5.95%
		% de Error al clasificar	11.18%

En todo caso, escoger uno u otro enfoque e incluso escoger la cota del 10% es un asunto parcialmente discrecional que

depende de los recursos con que cuente el proyecto. En este trabajo se ha elegido el enfoque más austero.

Finalmente, la Figura 6.3 presenta el comportamiento promedio de la Regresión Logística para distintos umbrales, la Tabla 44 muestra los detalles numéricos de esta figura. Nótese que el comportamiento promedio es muy similar al obtenido con la muestra 1 (véase la Figura 6.2).

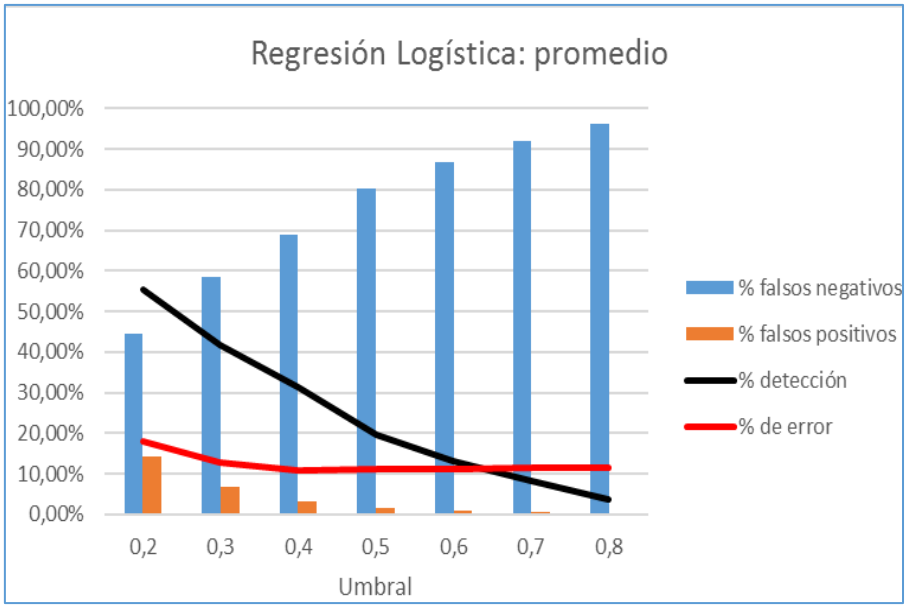


Figura 6.3.Tendencias de los principales indicadores de Regresión Logística, promedio

En la Tabla 44, el porcentaje promedio de detección con umbral 0.4 es 31.24% (celdas en amarillo) lo cual es ligeramente superior a los porcentajes promedio obtenidos por Naive Bayes

y 1-NN (30.99% y 30.79% respectivamente). Nótese además que el umbral 0.4 obtiene el menor porcentaje promedio de error.

Tabla 44. Resultados de la evaluación promedio de Regresión Logística

Umbral	% Falsos negativos	% Falsos positivos	% Detección	% de Error
0.2	44.47%	14.22%	55.53%	17.83%
0.3	58.30%	6.78%	41.70%	12.92%
0.4	68.76%	3.17%	31.24%	10.98%
0.5	80.37%	1.69%	19.63%	11.06%
0.6	86.88%	0.86%	13.12%	11.09%
0.7	91.91%	0.49%	8.10%	11.35%
0.8	96.22%	0.19%	3.78%	11.60%

6.2 ANÁLISIS COMPARATIVO DE LOS MODELOS

En esta parte del trabajo se discutirá de forma integral acerca de los 4 modelos vistos. El interés principal de este proyecto no es el comportamiento individual de los modelos sino el comportamiento del proyecto de minería de datos en su conjunto.

En la Tabla 45 se muestra un resumen comparativo de los porcentajes de detección de las mejores versiones de cada uno de los métodos aplicados a los 5 conjuntos de prueba. La fila sombreada en amarillo de la tabla indica los porcentajes promedio de detección por cada método.

Tal como se observa en la Tabla 45 y en la Figura 6.4, los métodos Knn1, Naive Bayes y Regresión Logística tienen porcentajes promedio similares, mientras que el método de Árbol tiene un porcentaje promedio de detección considerablemente menor.

Tabla 45. Porcentajes de detección por método y por muestra

Métodos	Árbol	Knn1	Naive Bayes	Logit (0.4)
Muestra 1	26.03%	30.14%	39.04%	39.04%
Muestra 2	20.00%	33.33%	34.67%	31.33%
Muestra 3	14.81%	30.25%	26.54%	27.16%
Muestra 4	19.38%	34.38%	27.50%	28.75%
Muestra 5	18.37%	25.85%	27.21%	29.93%
Promedio	19.72%	30.79%	30.99%	31.24%

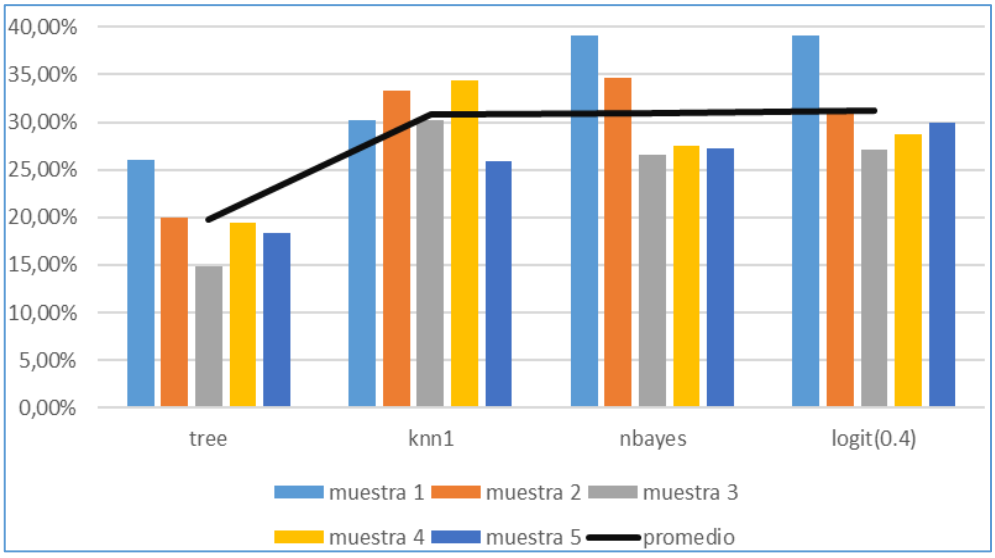


Figura 6.4. Porcentaje de detección por método y por muestra

En la Tabla 45, los mejores resultados se encuentran en las celdas azules mientras que los peores en las celdas rojas. Tal como se mencionó en la sección 6.1, la mejor (o peor) muestra para un método no lo es necesariamente para otro método. En la tabla citada, por ejemplo, la muestra 1 es la mejor para el método de Árbol, Naive Bayes y Regresión Logística pero no lo es para Knn1. Sin embargo, si existe una cierta generalidad: la muestra 1 es la mejor para 3 de 4 métodos y la muestra 3 es la peor también para 3 de 4 métodos.

A continuación, se determinará el porcentaje de detección del proyecto en su conjunto, es decir, congregando todo los métodos o técnicas de minerías de datos utilizadas.

Para encontrar el porcentaje de detección de desertores en este proyecto, es necesario aplicar de forma secuencial cada método e ir contabilizando los nuevos desertores que cada método proporciona.

En el caso de la muestra 1, el método de árbol predijo 38 de 146 desertores, siendo el porcentaje individual de detección 38/146 = 26.03% (véase la Tabla 31). Para la misma muestra 1, knn1 predijo 44 desertores con porcentaje individual de detección de 44/146 = 30.14% (véase la Tabla 34). Sin embargo, algunos desertores fueron encontrados por ambos métodos. Por lo tanto, el porcentaje de detección conjunto de ambos métodos sería algo menor que la suma de los porcentajes de

detección individuales. Tal como lo indica la Figura 6.5, el porcentaje de detección conjunto de los métodos de árbol y knn1 para la muestra 1 es 41.78%. Nótese que el método knn1 agregó 23 nuevos desertores.

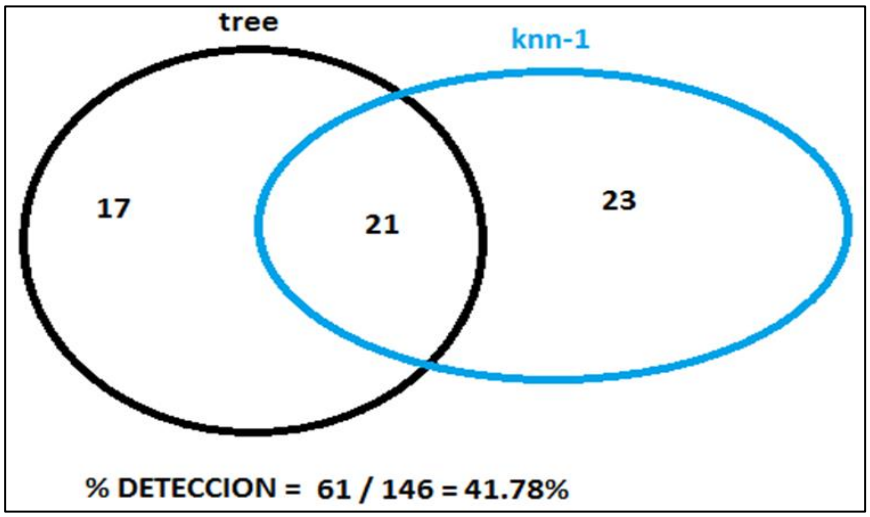


Figura 6.5. Porcentaje de detección conjunto de Árbol y Knn-1 para la muestra 1

Luego de haber aplicado los métodos de árbol y knn1 se agregará los nuevos desertores encontrados por el método Naive Bayes. Tal como se observa en la Figura 6.6, el método Naive Bayes agrega 20 nuevos desertores quedando un porcentaje de detección de (61 + 20) /146 = 55.48%.

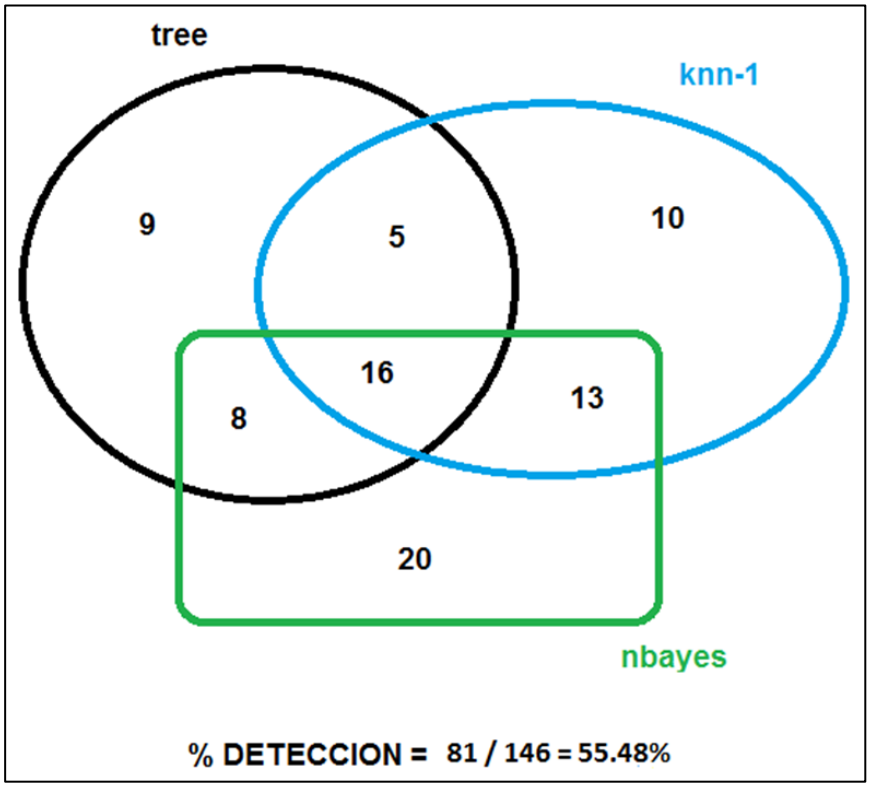


Figura 6.6. Porcentaje de detección conjunto de Árbol, Knn1 y Naive Bayes para la muestra 1

Finalmente, se agregará el método de Regresión Logística con umbral 0.4. Tal como muestra la Figura 6.7, este método apenas agrega 3 nuevos desertores incrementando el porcentaje de detección ligeramente a 57.53%.

La Figura 6.7 desvela un comportamiento importante respecto a los modelos del proyecto: individuamente los métodos con modelos explícitos como árbol y regresión logística agregan menor cantidad de desertores que aquellos métodos retardados con modelos incomprensibles (2 + 3 = 5 vs 8 + 12 = 20). Es decir, los métodos

retardados tienen mayor capacidad para encontrar nuevos desertores. Sin embargo, los métodos con modelos explícitos aumentan la comprensión del problema tratado en este proyecto de minería de datos.

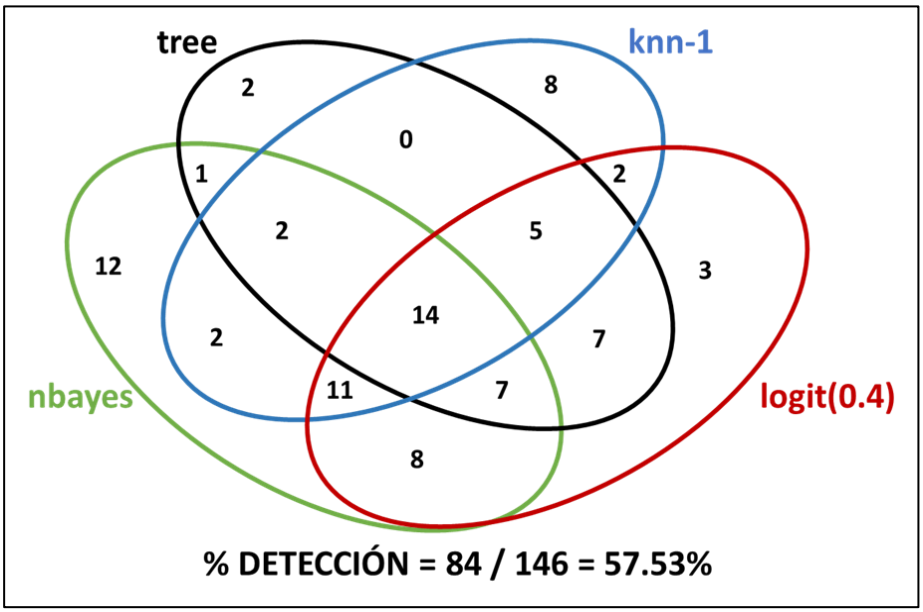


Figura 6.7. Porcentaje de detección conjunto de los 4 métodos para la muestra 1

El análisis debe continuar para las otras 4 muestras. La Tabla 46 muestra los porcentajes de detección acumulados por cada muestra. En esta tabla se pueden observar los distintos porcentajes de detección del proyecto conjunto para cada muestra. Nótese que el valor para la columna Logit (0.4) para la muestra 1 es 57.53% (celda en azul), precisamente el porcentaje de detección que se obtuvo en la Figura 6.7.

Tabla 46. Porcentaje de detección acumulado por muestra

Métodos	Árbol	Knn1	Naive Bayes	Logit (0.4)
Muestra 1	26.03%	41.78%	55.48%	57.53%
Muestra 2	20.00%	39.33%	52.66%	53.33%
Muestra 3	14.81%	34.57%	47.53%	48.77%
Muestra 4	19.38%	32.50%	41.88%	49.38%
Muestra 5	18.37%	36.73%	48.98%	51.70%
Promedio	19.72%	36.98%	49.31%	52.14%

La fila de la 5ta muestra de acuerdo a la Tabla 46 se podría leer de la siguiente manera:

- Para la muestra 5, el método de árbol detectó el 18.37% de desertores.
- Al agregar el método knn1, se incrementó el porcentaje de detección de 18.37% a 36.73%.
- Con el método de Naive Bayes el proyecto alcanzó un porcentaje de detección conjunto de 48.98%, permitiendo Bayes incrementar la detección en 12.25% (48.98% - 36.73%).
- Finalmente, al agregar Regresión Logística el proyecto total alcanzó un porcentaje de detección de 51.70%.

Tal como lo muestran la Tabla 46 (en amarillo) y la Figura 6.8, el porcentaje promedio de detección de todo el proyecto supera el 50%,

exactamente es 52.14%. Es decir, en promedio se esperaría que el actual proyecto de minería determine acertadamente más del 50% de los reales desertores. Un resultado prometedor si se tiene en cuenta que los métodos de discriminación no pretenden maximizar el porcentaje de detección de una clase determinada sino más bien minimizar el error conjunto por mala clasificación.

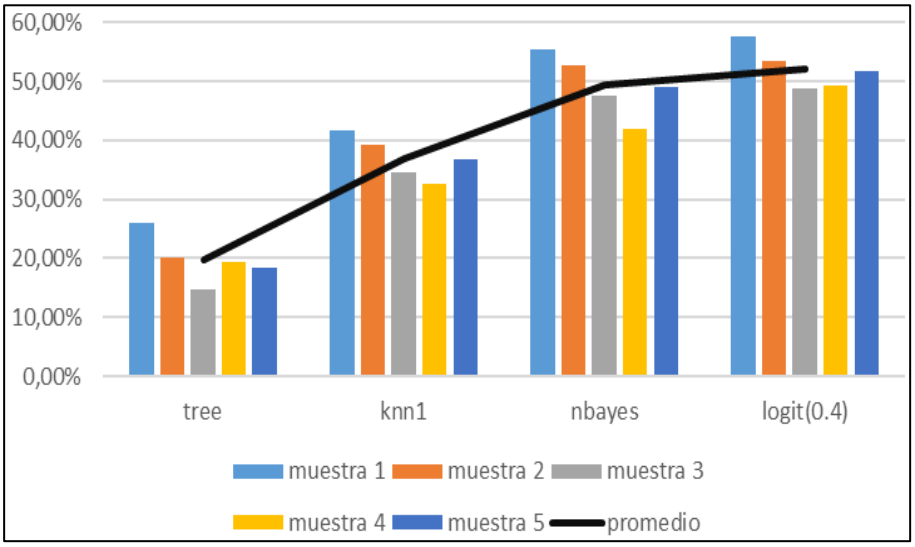


Figura 6.8. Porcentaje de detección acumulado del Proyecto

CONCLUSIONES Y RECOMENDACIONES

CONCLUSIONES

1. De acuerdo al método de árbol de decisión, las características personales de los estudiantes poco inciden en la deserción estudiantil. Sin embargo, su comportamiento académico es determinante, en particular, se obtuvo que reprobar más de 4 materias en las primeras fases de la carrera contribuyen significativamente a la deserción estudiantil.
2. De acuerdo al método de regresión logística, las variables que más contribuyen a la deserción estudiantil son: edad y cantidad de materias reprobadas. En promedio, se obtuvo que los estudiantes con mayor edad tienen 22% más posibilidades de desertar frente a los que son un

año menor; y, por cada materia reprobada las posibilidades de desertar se incrementan en un 28%.

3. De acuerdo a los métodos de árbol de decisión y de regresión logística, reprobar una materia estando a prueba no es garantía de deserción universitaria; así como superar un período de prueba tampoco implica mayor resiliencia en los estudios.
4. De acuerdo al método de regresión logística, la consulta de material bibliográfico pudiera contribuir levemente a evitar la deserción estudiantil. En promedio, las posibilidades de desertar se reducen en 1% por cada día de consulta.
5. De acuerdo a los métodos de árbol de decisión y de regresión logística, mientras más avanza el estudiante en sus estudios menos probable es su deserción. En particular, la regresión logística indica que, en promedio, cada materia aprobada reduce las posibilidades de deserción en un 16%.
6. Luego de los experimentos realizados, se estima que la capacidad promedio del proyecto para detectar un posible desertor es del 52.14%; y, la capacidad promedio para clasificar a un estudiante en el grupo correcto es del 87,11%.
7. De acuerdo a los métodos de regresión logística y naive bayes, de 9,159 estudiantes registrados en el segundo semestre del 2017, 1,391 estudiantes tienen una probabilidad de desertar de más del 50%;

mientras que 567 estudiantes tienen una probabilidad de desertar de más del 75%.

8. Luego de aplicar los 4 métodos de discriminación a los 9,159 estudiantes registrados en el segundo semestre del 2017, se obtuvo que alrededor del 22% de los estudiantes fueron detectados como posibles desertores por al menos uno de los métodos; mientras que 220 estudiantes fueron detectados por más de 2 métodos aumentando así su riesgo de deserción.

RECOMENDACIONES

1. Se recomienda que la universidad incremente la recolección de datos y asegure la confiabilidad de los mismos en tiempo cercano al real, esto permitirá llevar a cabo estudios más profundos y más ajustados a la realidad. En el caso puntual del análisis de la deserción universitaria, esta recomendación permitirá incluir mayor cantidad de factores que puedan explicar más claramente la deserción para proponer soluciones más eficaces. Por ejemplo, adicional al préstamo de libros el trabajo autónomo podría incluir asistencia a las ayudantías académicas, el uso de laboratorios de computación, consultas al profesor por medio de alguna plataforma como sidweb, entre otros aspectos. Por otra parte, para garantizar la confiabilidad de los datos es importante que la

recolección sea lo más objetiva posible, por ejemplo, es un error permitir que el mismo estudiante ingrese o actualice su etnia.

2. Se recomienda desarrollar e implementar una herramienta informática que permita monitorear constantemente el riesgo de deserción de los estudiantes y los factores que más están afectando a ese riesgo. Podría ser una herramienta que vaya agregando nuevos elementos a los modelos para mejorar las predicciones, teniendo presente que las predicciones como el comportamiento académico no son constantes.
3. El presente trabajo es apenas un inicio de varios estudios que pudieran realizarse respecto a la deserción, se recomienda no sólo aplicar técnicas de minería más sofisticadas para predecir y explicar el riesgo de deserción sino llevar a cabo un estudio longitudinal que considere las diferentes etapas de los estudiantes.
4. Con el objetivo de facilitar el análisis de datos del Sistema Académico, se recomienda implementar un data warehouse (almacén de datos) con sus respectivos cubos de información, donde especialmente se construyan tablas temporales que permitan obtener datos en un punto en el tiempo, como por ejemplo, la cantidad de estudiantes que estaban laborando al finalizar el primer semestre del 2016.

BIBLIOGRAFÍA

[1] El Telégrafo, La deserción universitaria bordea el 40%, https://www.eltelegrafo.com.ec/noticias/sociedad/4/la-desercion-universitaria-bordea-el-40, 2016.

[2] Kovačić Z., Early Prediction of Student Success, https://pdfs.semanticscholar.org/e48e/ba98bde33586c20442d46ab9a59c411196e5.pdf, 2010.

[3] SPSS, CRISP-DM 1.0 Step-by-step data mining guide, https://www.the-modeling-agency.com/crisp-dm.pdf, 2000.

[4] James G., Witten D., Hastie T. y Tibshirani R., An Introduction to Statistical Learning, Springer 7th Ed, 2014.

[5] Leskovec J., Rajaraman A. y Ullman J., Minig of Massive Datasets, Cambridge University Press 2nd Ed, 2014.

[6] Hernández J., Ramírez M. y Ferri C., Introducción a la minería de datos, Pearson Education, 2004.

[7] Noboa C., Diseño e implementación de un aplicativo web para el aprendizaje de análisis discriminante, ESPOL, 2008.

[8] Johnson R. y Wichern D., Applied Multivariate, Prentice Hall 4th Ed, 1998.

[9] Anderson D., Sweeney D. y Williams T., Estadística para negocios y economía, Cengage Learning 11a Ed, 2011.

[10] Russell S. y Norvig P., Artificial Intelligence A Modern Approach, Prentice Hall 3rd Ed, 2010.

[11] Frees E., Data Analysis Using Regression Models The Business Perspective, Prentice Hall, 1996.

[12] Jaramillo A. y Paz-Arias H., Aplicación de Técnicas de Minería de Datos para Determinar las Interacciones de los Estudiantes en un Entorno Virtual de Aprendizaje, http://www.rte.espol.edu.ec/index.php/tecnologica/article/view/351/229, 2015.

[13] Moine J., Metodologías para el descubrimiento, http://sedici.unlp.edu.ar/bitstream/handle/10915/29582/Documento_completo.pdf?sequence=1, 2013.

[14] Rodríguez C. y García M., Adecuación a metodología de minería de datos para aplicar a problemas no supervisados tipo atributo-valor, http://scielo.sld.cu/pdf/rus/v8n4/rus05416.pdf, 2016.

[15] IBM, Manual CRISP-DM de IBM SPSS, ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/es/CRISP-DM.pdf, 2012.

[16] Muenchen R., The Popularity of Data Science Software, http://r4stats.com/articles/popularity/, 2017.

[17] TIOBE, Index for April 2018, https://www.tiobe.com/tiobe-index/, 2018.

[18] Witten I., Frank E. y Hall M., Data Mining Practical Machine Learning Tools and Techniques, Elsevier 3rd Ed, 2011.

[19] Samarasinghe S., Neural Networks for Applied Sciences and Engineering, Auerbach Publications, 2007.