UNIVERSIDAD AUTONOMA DE MANIZALES

MAESTRIA EN GESTIÓN Y DESARROLLO DE PROYECTOS DE SOFTWARE

SISTEMA DE APOYO PARA LA ACREDITACIÓN DE LA CALIDAD DE PROGRAMAS ACADEMICOS DE LA UNIVERSIDAD DE CALDAS, APLICANDO TÉCNICAS EN MINERÍA DE DATOS

JUAN CARLOS GONZÁLEZ CARDONA

MANIZALES, CALDAS – COLOMBIA

NOVIEMBRE 2011

SISTEMA DE APOYO PARA LA ACREDITACIÓN DE LA CALIDAD DE PROGRAMAS ACADEMICOS DE LA UNIVERSIDAD DE CALDAS, APLICANDO TÉCNICAS EN MINERÍA DE DATOS

JUAN CARLOS GONZÁLEZ CARDONA

Informe Final

Trabajo de Grado para optar al título de Magister en Gestión y Desarrollo de Proyectos de Software

Asesor Temático y Metodológico

Msc. Javier Hernández Cáceres

UNIVERSIDAD AUTÓNOMA DE MANIZALES
MAESTRIA EN GESTIÓN Y DESARROLLO DE PROYECTOS DE SOFTWARE
MANIZALES
2011

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

RESUMEN

Una de las principales dificultades que enfrenta el sistema educativo actual es la deserción, su valor acumulado llega a niveles del 45% a nivel nacional y del 33% en la Universidad de Caldas; una de sus principales causas es la deserción de tipo académico, por lo que se hace primordial la definición de un indicador que logre medir el rendimiento académico en sus diferentes dimensiones (la excelencia, la eficiencia y la eficacia). Una vez definido el indicador, es fundamental determinar los factores que inciden en este con el objeto de tomar acciones que tengan el mayor impacto posible; estos factores incluyen elementos de identidad, socioeconómicos, vocacionales, de estudios previos, del entorno familiar entre otros y sus relaciones, por lo que un análisis multivariado debe ser el tipo de modelo que los caracterice, la regresión logística y su amplio uso en la determinación de factores de riesgo o de protección fue el análisis seleccionado, tanto por esta característica como por su manejo de variables de tipo numérico como categóricas. Este proceso de extracción de conocimiento a partir de los datos KDD, que ha estado en auge en los últimos años en los ambientes educativos, se desarrolló utilizando la metodología CRISP-DM y fue implementado sobre la aplicación de uso libre RapidMiner y comprende desde la extracción de información de la base de datos del Sistema de Información Académica SIA, su transformación, validación, el cálculo de los índices y el indicador de rendimiento académico IAR, el análisis de Regresión

Logística por programa y nivel de avance en créditos del programa, hasta la generación de informes de tipo descriptivo como del modelo. Los resultados finales muestran los factores de riesgo y de protección en el rendimiento académico para los estudiantes de cada programa presencial en diferentes momentos de su paso por la universidad.

INTRODUCCIÓN

La deserción estudiantil es una de las principales preocupaciones que viven las universidades colombianas, el Ministerio de Educación Nacional ha realizado numerosos estudios sobre esta (Ministerio de Educación Nacional de Colombia, 2009) y creado el aplicativo SPADIES para su gestión. Una de las principales causas de la deserción estudiantil es la de tipo académico, relacionada a esta última, están problemáticas que deben enfrentar las universidades como la repitencia, los prolongados tiempos para graduarse de algunos estudiantes o la no búsqueda de la excelencia en los estudios.

La Universidad de Caldas ha realizado estudios recientes sobre la deserción propia a nivel descriptivo (Candamil Calle, Palomá Parra, & Sánchez Buitrago, 2009), pero hacen falta estudios que busquen relaciones causales y que tengan en cuenta su origen multifactorial, es así como se planteó un proyecto de extracción de conocimiento a partir de bases de datos KDD aprovechando la vasta información alojada en el Sistema de Información Académico SIA y el auge de de la minería de datos en ambientes educativos (Winters, 2006).

El estudio parte de la información disponible en el SIA de tipo institucional, socioeconómica, demográfica y académica de 17.206 estudiantes de los

diferentes programas de pregrado presencial que ingresaron entre el año 2001 y 2010; a dicha información se le realiza un proceso de verificación y limpieza, al final de este proceso resultan 10.904 estudiantes con información consistente.

Para cada estudiante, la información académica se agrupa en 6 segmentos, de acuerdo a su avance en créditos del programa académico, dichos segmentos son primer semestre (segmento 0), entre el 0 y el 25% (segmento 25), entre el 25% y el 50% (segmento 50), entre el 50% y el 75% (segmento 75), entre el 75% y el 100% (segmento 100) y el segmento general para el cual no se hacen divisiones y se tiene en cuenta todo el recorrido académico que lleve el estudiante en el momento de realizar el estudio.

Posteriormente, se definen 5 índices que tienen en cuenta la excelencia, la eficacia y la eficiencia basados en la metodología definida en la Universidad de Los Andes de Venezuela (Garnica Olmos, 1997), estos son el promedio académico, el promedio académico de las asignaturas ganadas, el índice de créditos aprobados ICA, el índice de materias aprobadas IMA y el promedio de créditos inscritos por semestre PCC; calculados para cada segmento de avance en créditos del estudiante.

Con base en los anteriores índices, se calcula el indicador académico de rendimiento IAR, que es el promedio aritmético de los 5 índices luego de estandarizarlos en una escala de 0 a 1. El IAR es categorizado en 3 niveles, Alto Medio y Bajo, con el fin de facilitar su interpretación y permitir la toma de decisiones; esta categorización se hizo basada en el juicio de profesores y el

reglamento estudiantil en sus apartes de pérdida de la calidad de estudiante por bajo rendimiento y créditos máximos para la inscripción entre otros.

El modelo de análisis seleccionado fue la Regresión Logística Multinomial, principalmente porque permite conocer los factores de influencia de un conjunto de variables explicativas ya sean numéricas o categóricas sobre una variable de respuesta de tipo multinomial. Las variables explicativas estaban divididas en 2 grupos, las categóricas de tipo binomial fueron el sexo, la naturaleza del colegio (oficial o privado), el nivel de la ciudad (capital o no capital) y el pago de matrícula académica (exento o no); las de tipo numérico que fueron la edad y los puntajes de las pruebas ICFES posteriores al 2000, matemáticas, lenguaje, exactas (promedio aritmético de física, química y biología) y humanas (promedio aritmético de filosofía, historia y geografía).

Como resumen, se presentan las pruebas de máxima verosimilitud para cada modelo, además se presentan los factores con significancia estadística que influyen para que un estudiante tenga rendimiento académico alto o bajo en relación con la categoría base medio por programa académico y avance en los créditos en dicho programa.

El desarrollo del trabajo siguió la metodología CRISP-DM, un estándar internacional en proyectos de minería de datos, presentando para cada fase las parametrizaciones y resultados obtenidos.

Se utilizó únicamente software de uso libre, entre ellos SQL Express como motor de base de datos para simular el repositorio del Sistema de Información

Académico, RapidMiner para implementar el proceso de ETL y el flujo del proceso de minería de datos y por último el proyecto R para el modelado de la regresión logística.

Finalmente se presentan el análisis de resultados y las conclusiones.

REFERENTE CONTEXTUAL

Descripción del Área Problemática

Colombia presenta altos niveles de deserción -cercanos al 45%-, una de las principales causas es el rendimiento académico (Ministerio de Educación Nacional de Colombia, 2009); en la Universidad de Caldas, los niveles y las causas son similares como lo muestra el estudio Análisis de la Deserción estudiantil en la Universidad de Caldas 1998-2006 (Candamil Calle *et al.,* 2009). Esta problemática aunada a la baja tasa de cobertura, cercana al 25%, disminuye la población que logra obtener un título profesional, y por ende disminuir los niveles de desigualdad presentes en la sociedad.

El desconocimiento de los factores que afectan el rendimiento académico como una de las causas de la deserción, no sólo es un problema en la Universidad de Caldas; instituciones como el Ministerio de Educación Nacional MEN no han estado exentas de esta preocupación y es así, como en el año 2000 incluyó en el Plan de Educación los Exámenes de Estado de Calidad de la Educación Superior (ECAES) con el objetivo de contribuir en el mejoramiento de la calidad y la transparencia en las Instituciones de Educación Superior (IES) en

la presentación del informe, el director del ICFES, Daniel Bogoya hizo énfasis en que la evaluación debe enfocarse en la mejora de las estrategias pedagógicas:

> "Los resultados de un proceso de evaluación educativa permiten reconocer el perfil de competencias de un estudiante, de una institución o de un programa de formación y, de esta manera, favorecen la cualificación de las prácticas pedagógicas, contribuyendo al mejoramiento de la calidad del sistema." (Rocha, Pardo, Bohórquez, & Barrera, 2003)

Adicionalmente, el propio Ministerio ha brindado a las Universidades herramientas que faciliten la gestión académica, entre ellas se encuentra SPADIES, enfocándose en la evaluación del riesgo que tiene un estudiante de desertar (Guzmán, 2007).

Antecedentes

El rendimiento estudiantil: una metodología para su medición

Garnica (1997) plantea que la única forma de lograr una eficiencia educativa es conocer las deficiencias y errores cometidos durante el proceso educativo; indicando que para esto es necesario conocer las causas que influyen en el rendimiento educativo así como su medición, enfatizando que son procesos

separados. Propone la medición del rendimiento educativo a través de la técnica estadística multivariante del Análisis de los Componentes Principales ACP y la comparación de diferencias entre grupos mediante el Análisis Unifactorial de Varianza ANOVA.

Se basa en estudios previos que definen que la medición del rendimiento educativo debe evaluar las siguientes tres variables latentes: Rendimiento general, Consistencia y Motivación en la carrera. Así mismo, incluye el rendimiento académico en el bachillerato como un buen predictor del rendimiento en la Universidad.

Con base en esas variables latentes define cuatro grupos variables a tener en cuenta en el ACP:

Grupo 1. Variables obtenidas directamente de los registros estudiantiles del plantel. Unidades cursadas, unidades aprobadas, materias cursadas, materias aprobadas, promedio global, promedio aprobatorio, semestres cursados, número de semestres intensivos cursados, materias retiradas y número de semestres en que se hace retiros de materias.

Grupo 2. Variables indicadores calculadas por el investigador. Indicador de materias aprobadas ([(número de materias aprobadas) / (número de materias cursadas)] = IMA), indicador de unidades aprobadas ([(unidades aprobadas) / (unidades cursadas)] = IUA) e indicador de materias retiradas ([(número de materias retiradas) / número de materias cursadas)] = IMR).

Grupo 3. Variables promedios calculados por el investigador. Promedio de materias aprobadas ([(número de materias aprobadas) / (número de semestres cursados)] = PMA), promedio de unidades cursadas ([(unidades cursadas) / (número de semestres cursados)] =PUC), promedio de unidades aprobadas ([(unidades aprobadas) / (número de semestres cursados)] = PUA) y promedio de materias retiradas ([(número de materias retiradas) / (número de semestres cursados)] = PMR).

Grupo 4. Variables de tiempo de retardo en los estudios, calculadas por el investigador. Tiempo de retardo. Para los estudiantes que aún no tienen los diez semestres teóricos de duración de la carrera, se calcula su tiempo igual a cero (no se cuentan los semestres intensivos). Para aquellos que han pasado de los diez semestres regulares estudiando la carrera, se calcula la diferencia entre el número de semestres que tiene en la carrera y 10, esta diferencia es la variable bajo estudio: tiempo de retardo en los estudios. Existen otras variables de tiempo, como por ejemplo el tiempo de retardo 2, que es calculado como se mencionó anteriormente, pero restándole el número de semestres en que el alumno estuvo retirado (por reglamentos o por causas personales).

Las variables propuestas por (Garnica, 1997) serán empleadas en el desarrollo del presente proyecto excluyendo las que no están disponibles en el SIA como las materias retiradas.

Data Mining and Knowledge Management in Higher Education - Potential Applications

Luan (2002) introduce el uso de la KDD en el sector educativo, presentando el paralelo de la aplicación de la minería de datos con el sector de los negocios, haciendo la correspondencia entre un conjunto de preguntas, como:

> *"¿Qué clientes podrían irse hacia los competidores? versus ¿Qué cursos atraen más estudiantes" o "Cuáles son mis clientes leales? versus ¿Cuáles son los estudiantes más perseverantes?"*

Luan además presenta una serie de aplicaciones potenciales como la identificación de las tipologías de aprendizaje de los estudiantes, la categorización de los estudiantes para determinar el tiempo esperado de permanencia o deserción y la optimización en el envío de cartas de invitación a inscribirse en la universidad. Las soluciones propuestas siguen el modelo de manejo del conocimiento por niveles propuesto por él (Tiered Knowledge Management Model TKMM).

Minería de datos para descubrir estilos de aprendizaje

Durán & Costaguta (2007), en la Universidad Nacional de Santiago del Estero en Argentina, aplican minería de datos utilizando el aplicativo Weka sobre las respuestas dadas por un conjunto de estudiantes, del programa de Licenciatura

en Sistemas de Información, a la prueba propuesta por Felder y Soloman (citada por Durán & Costaguta, 2007), en la cual se busca clasificar al estudiante en un estilo de aprendizaje de acuerdo al modelo planteado por Felder y Silverman (citada por Durán & Costaguta, 2007). Este modelo está basado en 4 dimensiones; percepción, entrada, procesamiento y comprensión y unos estilos en cada dimensión; Sensorial o intuitivo, Visual o verbal, Activo o reflexivo, Secuencial o global. La investigación tiene como resultado una caracterización de los estilos de aprendizaje de los estudiantes de la licenciatura y proponen unos estilos de enseñanza acordes al conocimiento encontrado.

La minería de datos puede ser empleada no sólo sobre variables fácilmente cuantificables; como las calificaciones, resultados en pruebas de estado, información socioeconómica; sino además, sobre cualidades cognoscitivas, sociológicas o psicológicas, claro, a través de instrumentos cómo la prueba propuesta por Felder y Soloman (citada por Durán & Costaguta, 2007). Para el alcance del presente proyecto no se incluyen este tipo de variables, sin embargo, una siguiente etapa del proyecto podría involucrar este tipo de variables que contextualicen aún más los resultados obtenidos.

Obtención de patrones y reglas en el proceso académico de la Universidad de las Ciencias Informáticas utilizando técnicas de minería de datos.

González, Pérez, Espinosa & Alvarez (2007), definieron un modelo para determinar las calificaciones en ciertas asignaturas clave de los estudiantes de

primer semestre de la Universidad de las Ciencias Informáticas en Cuba, basándose en variables como tipo de colegio, provincia de procedencia, nivel de estudio de los padres utilizando las herramientas de minería de datos de SQL Server y utilizando la metodología para minería de datos CRISP-DM. El estudio identifica las variables que influyen negativamente en las calificaciones de los estudiantes de primer semestre con un alto grado de certeza.

González *et al.* (2007) evidencia que existen factores previos al ingreso a la Universidad que influyen en el rendimiento académico de los estudiantes durante su vida universitaria, por lo que elementos como la información socioeconómica incluyendo la procedencia y el tipo de educación secundaria se incluirán en el desarrollo del presente proyecto.

Educational Data Mining: Collection and Analysis of Score Matrices for Outcomes-Based Assessment

Winters (2006) en su tesis doctoral en Ciencias de la Computación de la Universidad de California Riverside, hace una extensa revisión del estado del arte de la minería de datos educacional, las técnicas de minería y las áreas de la psicometría y la cognición. A partir de ellas encausa su investigación sobre la evaluación del nivel de aprendizaje de los estudiantes en cada objetivo de las asignaturas y el grado en que el plan de estudios cumple con los resultados

esperados del programa académico asociado al departamento de Ciencias de la Computación e Ingeniería. Para esto utiliza el siguiente conjunto de matrices:

1. Matriz de evaluaciones (Score Matriz): calificación obtenida por el estudiante en cada pregunta de las evaluaciones aplicadas para cada curso.
2. Matriz de relevancia (Relevance Matrix): relación lineal entre cada pregunta y el objetivo del curso, indicando su relevancia.
3. Matriz del curso (Course matrix): relación entre los objetivos del curso y los resultados esperados del programa.

Un elemento importante en los sistemas de Inteligencia de Negocios es conocer las causas y los efectos de ciertos comportamientos, de tal forma que orienten a la toma de decisiones, por lo que poder realizar una trazabilidad de las calificaciones y los cursos contra los objetivos de los programas académicos facilitaría la definición de estrategias que mejoren el rendimiento académico.

Aunque por el alcance del presente proyecto y por el tipo de información existente sea inviable la aplicación de las matrices propuestas por Winters, si plantea una necesidad de la Universidad de Caldas de llegar a estos niveles de trazabilidad de sus procesos.

Sin embargo, los cambios en el rendimiento académico del estudiante a través de su recorrido en la Universidad se incorporarán como variables en el desarrollo de los modelos propuestos.

Sistema de Prevención y Análisis de la Deserción en Las Instituciones de Educación Superior (SPADIES)

En Colombia, el Ministerio de Educación Nacional (MEN) junto con el Centro de Estudios Económicos (CEDE) de la Universidad de los Andes construyeron el Sistema de Prevención y Análisis a la Deserción en las Instituciones de Educación Superior (SPADIES) con el fin de hacer seguimiento a la problemática de la deserción en las IES.

Una de las grandes ventajas de este sistema es su alto grado de integración con los sistemas de información de la gran mayoría de instituciones relacionadas con la Educación Superior como el ICFES, Icetex, SNIES y alimentado por los sistemas de Información de todas las IES con carácter de obligatoriedad para estas. Entre algunas de sus características es que presenta información de estadística descriptiva histórica, y adicionalmente incluye análisis de riesgos para determinar que estudiantes son más propensos a desertar; pero deja por fuera elementos para caracterizar y predecir el rendimiento académico de los estudiantes tanto a nivel general como por cursos. (Guzmán, 2007)

SPADIES debe ser una fuente de información adicional a las bases de datos de las Universidades, de tal forma que además de ayudar en la elaboración de estrategias que combatan la deserción, apoye otra clase de decisiones como la mejora del rendimiento académico.

Como se observa a lo largo de los trabajos relacionados, la extracción de conocimiento a partir de grandes volúmenes de datos ha tenido un gran interés en los últimos años; los cuales les ha permitido a las instituciones tomar decisiones fundamentadas en el conocimiento extraído, lo cual es el objetivo último del KDD; en la Universidad de Caldas a pesar de contar con un gran registro de datos relacionados con el desarrollo académico de los estudiantes, este ha sido precariamente utilizado para la definición de estrategias de desarrollo institucional o en la definición de tácticas por parte de los ejecutores académicos (como se evidencia en la falta de estudios o artículos relacionados) limitándose al uso con fines de publicaciones estadísticas, informes a instituciones gubernamentales o el análisis del desarrollo académico de unos cuantos estudiantes de manera individual y manual, de tal forma que es imposible a través de estos medios descubrir los patrones que pueden subyacer en esta gran fuente de datos.

JUSTIFICACIÓN

La Universidad de Caldas cuenta con gran cantidad de información que describe la situación académica de sus estudiantes, así como su estado socioeconómico y su formación previa; sin embargo, esta información es utilizada tangencialmente en la definición de los planes de mejoramiento (ajuste de planes de estudio, redefinición de metodologías de aprendizaje, estrategias de apoyo académico y de bienestar) debido al desconocimiento de los factores que realmente influyen en el rendimiento académico; desperdiciando así recursos de la universidad al no poder focalizarlos adecuadamente.

Los promedios académicos, las tasas de repitencia, la permanencia del estudiante en un programa universitario son factores relevantes en la acreditación de cualquier programa de la institución (Consejo Nacional de Educación Superior, 1995). Teniendo en cuenta lo anterior, el propósito de este trabajo es contribuir con un documento que evidencie los factores que influyen en el rendimiento académico que sirva de apoyo al personal directivo, docente y comunidad educativa en general de la universidad en la mejora de la calidad en la formación de sus profesionales.

Por la capacidad que tiene la minería de datos y en un sentido más amplio la extracción de conocimiento a partir de los datos (KDD por sus siglas en inglés)

de generar modelos que permitan identificar patrones de comportamiento y predecir sucesos (Luan, 2002) han sido vistas con buenos ojos por el sector educativo (Dapozo, Porcel, López, Bogado, & Bargiela, 2006).

A pesar de la existencia de estudios sobre la deserción en el contexto de la Universidad de Caldas, estos se han limitado al enfoque descriptivo, sin buscar la extracción de conocimiento de la fuente primaria de información como es el Sistema de Información Académica, es aquí donde un sistema que integre la minería de datos y el análisis multivariado con los procesos de ETL y generación de informes automatizado, enfoca a los directivos y personal administrativo de la Universidad de Caldas en las labores que realmente generan valor, que les permitan la toma de decisiones de manera ágil y la definición de planes de acción basados en conocimiento real.

Existen aplicativos de uso libre como RapidMiner que permiten la construcción de procesos de minería de datos de punta a punta, sin incurrir en altos costos de licenciamiento o mantenimiento que para instituciones académicas resultan tan difíciles de justificar. No obstante, el que sea una herramienta de uso libre, sus características son similares a cualquier aplicativo propietario, en gran parte debido a que detrás de esta existe una comunidad tanto de programadores como de científicos que le ha dado gran robustez y crecimiento.

FORMULACIÓN DEL PROBLEMA

¿Qué factores influyen más en el rendimiento académico de los estudiantes de la Universidad de Caldas?

La medición del rendimiento académico no debe plantearse como una simple revisión del promedio académico o de la pérdida de asignaturas, este debe evaluarse desde una perspectiva que incluya la excelencia, la eficiencia y la eficacia de los estudios realizados por los estudiantes (Garnica, 1997); algunos de los componentes que reflejan esta perspectiva y que pueden ser fácilmente evaluables son el promedio académico, la cantidad de materias y créditos inscritos por semestre, así como la efectividad en su aprobación.

Es evidente que el rendimiento académico y sus componentes deben estar influenciados por factores socioeconómicos, de calidad de la formación durante el bachillerato (González *et al.* 2007) de tipo vocacional, de elementos sociológicos o sicológicos como las formas de aprendizaje (Durán & Costaguta, 2007); es por esto que el modelo de análisis debe contemplar como variables explicativas dichos factores y fundarse en la estadística multivariada que logre visibilizar sus relaciones (Carvajal, Trejos, & Soto, 2004).

OBJETIVOS

Objetivo General

Describir los factores que inciden en el rendimiento académico de los estudiantes de la Universidad de Caldas por medio de técnicas en minería de datos.

Objetivos Específicos

1. Construir la vista minable de datos a partir de la base de datos académica de la Universidad de Caldas utilizando herramientas que permitan la extracción, transformación y carga (ETL) de manera automatizada.
2. Definir los componentes y el indicador académico de rendimiento.
3. Determinar los factores críticos que afectan el indicador académico de rendimiento utilizando la Regresión Logística Multinomial.
4. Establecer los componentes que más aportan al indicador de rendimiento con un análisis de correlación.
5. Validar los resultados obtenidos.
6. Automatizar la generación de los resultados de acuerdo a los modelos definidos.

RESULTADOS ESPERADOS

1. Vista minable de datos a partir de la base de datos académica de la Universidad de Caldas óptima y consistente.
2. Flujo de trabajo con el proceso de ETL automatizado en una herramienta de uso libre que genere la vista minable.
3. Documento con la estructura de análisis estadístico y los modelos definidos que describan los factores que inciden en el rendimiento académico, que incluya la validación de los resultados.
4. Flujo de trabajo con el proceso de generación de resultados automatizado en una herramienta de uso libre.

ESTRATEGIA METODOLÓGICA

Metodología

Al ser este un proyecto de investigación aplicada, la metodología se basa en la metodología CRISP-DM por su uso generalizado en este tipo de proyectos (Chapman *et al.,* 2007) y que abarca todos los objetivos propuestos. Se limita a las siguientes tareas acordes con el alcance del proyecto.

Comprensión del dominio

Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto y exigencias desde una perspectiva institucional, luego convirtiendo este conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos (Chapman *et. al.,* 2000).

1. Establecimiento de los objetivos del proyecto.
2. Evaluación de la situación.
3. Establecimiento de los objetivos de la minería de datos.

Comprensión de los datos

La fase de entendimiento de datos comienza con la colección de datos inicial y continúa con las actividades que le permiten familiarizarse primero con los datos, identificar los problemas de calidad de datos, descubrir los primeros conocimientos en los datos, y descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta (Chapman *et al.,* 2007).

1. Recopilación inicial de datos.
2. Descripción de los datos.
3. Exploración de los datos.
4. Verificación de calidad de datos.

Preparación de los datos

La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto de datos final (los datos que serán provistos en las herramientas de modelado) de los datos en brutos iniciales. Las tareas de preparación de datos probablemente van a ser realizadas muchas veces y no en cualquier orden prescripto. Las tareas incluyen la selección de tablas, registros, y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan (Chapman *et al.,* 2007).

1. Selección de los datos.
2. Limpieza de datos.
3. Construcción de datos.
4. Integración de datos.

Modelado

En esta fase, varias técnicas de modelado son seleccionadas y aplicadas, y sus parámetros son calibrados a valores óptimos. Típicamente hay varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requerimientos específicos sobre la forma de datos. Por lo tanto, volver a la fase de preparación de datos es a menudo necesario (Chapman *et al.,* 2007).

1. Selección de la técnica de modelado.
2. Diseño de la evaluación.
3. Construcción del modelo.
4. Evaluación del modelo.

Evaluación

En esta etapa del proyecto, se ha construido un modelo (o modelos) que parecen tener la alta calidad desde la perspectiva de análisis de datos.

Antes del proceder al despliegue final del modelo, es importante evaluarlo a fondo y la revisión de los pasos ejecutados para crearlo, para comparar el modelo correctamente obtenido con los objetivos de negocio. Un objetivo clave es determinar si hay alguna cuestión importante de negocio que no ha sido suficientemente considerada. En el final de esta fase, una decisión en el uso de los resultados de minería de datos debería ser obtenida (Chapman *et al.,* 2007).

1. Evaluación de resultados
2. Revisar el proceso
3. Validación de resultados obtenidos

Despliegue

La creación del modelo no es generalmente el final del proyecto. Incluso si el objetivo del modelo es de aumentar el conocimiento de los datos, el conocimiento ganado tendrá que ser organizado y presentado en el modo en el que el cliente pueda usarlo. Ello a menudo implica la aplicación de modelos "vivos" dentro de un proceso de toma de decisiones de una organización, por ejemplo, en tiempo real la personalización de página Web o la repetida obtención de bases de datos de mercadeo. Dependiendo de los requerimientos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización repetida de un proceso cruzado de minería de datos a través de la empresa. En muchos casos, es el cliente, no el analista de

datos, quien lleva el paso de desarrollo. Sin embargo, incluso si el analista realizara el esfuerzo de despliegue, esto es importante para el cliente para entender de frente que acciones necesita para ser ejecutadas en orden para hacer uso de los modelos creados actualmente.

- Generación de informe final

Pruebas

Las pruebas se definen en el apartado “Generar el diseño de la prueba”, en el capítulo Desarrollo siguiendo la metodología CRISP-DM.

Presupuesto

Para el desarrollo del presente proyecto no se contempla presupuesto pues será desarrollado únicamente por el autor, con la asesoría exclusiva del asesor de Trabajo Final y utilizando aplicaciones informáticas de uso gratuito.

DESARROLLO

Referente Teórico

Acreditación Institucional

El gobierno nacional a través del Consejo Nacional de Educación Superior CESU mediante el acuerdo 6 de 1995 estableció las políticas generales de acreditación, dando las siguientes definiciones: (Consejo Nacional de Educación Superior, 1995)

1. La Acreditación es el acto por el cual el Estado adopta y hace público el reconocimiento que los pares académicos hacen de la comprobación que efectúa una institución sobre la calidad de sus programas académicos, su organización y funcionamiento y el cumplimiento de su función social.
2. La autoevaluación, hecha por las instituciones, para lo cual deben utilizarse guías coherentes con los criterios y características de calidad definidos por el Consejo Nacional de Acreditación. Estas guías podrán ser diferenciadas según el tipo de institución o área del conocimiento, y deberán incluir elementos cuantitativos y cualitativos. Esta autoevaluación deberá tener como punto de partida la definición que haga la institución de su naturaleza, su misión y su

proyecto educativo. Se busca preservar las características propias de cada institución, no de homogeneizarlas.

3. La evaluación externa, hecha por los pares académicos nombrados por el Consejo Nacional de Acreditación, mediante visita a la institución, para comprobar la objetividad y veracidad de la autoevaluación en cuanto a la calidad de sus programas académicos, su organización y funcionamiento y el cumplimiento de su función social. La evaluación externa concluirá con el informe que rindan estos pares sobre los resultados, acompañado de recomendaciones para el mejoramiento institucional, cuando sea necesario.

El Consejo Nacional de Acreditación, define que la calidad de la educación superior es la razón de ser del Sistema Nacional de Acreditación y determina que para determinar la calidad se tienen en cuenta las siguientes características:

Las características universales expresadas en sus notas constitutivas. Estas características sirven como fundamento de la tipología de las instituciones y establecen los denominadores comunes de cada tipo.

Los referentes históricos, es decir, lo que la institución ha pretendido ser, lo que históricamente han sido las instituciones de su tipo y lo que en el momento histórico presente y en la sociedad concreta se reconoce como el tipo al que esta institución pertenece (la normatividad existente y las orientaciones básicas que movilizan el sector educativo, entre otros).

Lo que la institución singularmente considerada define como su especificidad o su vocación primera (la misión institucional y sus propósitos).

La calidad es un elemento primordial de las políticas de educación nacional y uno de los objetivos del Ministerio de Educación Nacional, por lo que el Consejo Nacional para el Aseguramiento de la Calidad (CONACES) definió cinco mecanismos para el aseguramiento de la calidad:

1. Los estándares mínimos de calidad o registro calificado, donde se definen una serie de requisitos mínimos obligatorios para que una Institución de Educación Superior – IES ofrezca programas de calidad en pregrado y postgrado, con el funcionamiento del Consejo Nacional para el Aseguramiento de la Calidad (CONACES).
2. Aplicación obligatoria de los exámenes de calidad (ECAES) a todos los estudiantes de último año para evaluar las competencias requeridas para el ejercicio de la profesión correspondiente.
3. Incremento en el número de programas con acreditación de alta calidad.
4. El diseño e implementación de planes de mejoramiento en aquellas instituciones que reporten debilidades.

Examen de estado de la educación media – ICFES SABER 11°

El examen de estado de la educación media – ICFES SABER 11° realizada por el Instituto Colombiano de Fomento a la Educación Superior a los estudiantes de undécimo grado tiene como objetivo: (ICFES, 2010)

1. Requisito obligatorio para el ingreso a la educación superior
2. Información para los estudiantes sobre sus competencias en las diferentes áreas: apoyo para la orientación sobre su opción profesional
3. Criterio para la autoevaluación de los establecimientos educativos en función de sus proyectos educativos y planes de mejoramiento
4. Criterio para otorgar beneficios educativos (becas, premios).
5. Base para estudios de carácter cultural, social, económico y educativo y retroalimentar el quehacer de la evaluación

Su estructura está dividida en dos componentes, núcleo común y componente flexible, ya para cada uno existen las siguientes temáticas:

Núcleo común

1. Lenguaje (24 preguntas)
2. Matemáticas (24 preguntas)
3. Biología (24 preguntas)
4. Química (24 preguntas)
5. Física (24 preguntas)
6. Ciencias sociales (30 preguntas)

7. Filosofía (24 preguntas)
8. Inglés (45 preguntas)

Componente flexible

1. Profundizaciones (15 preguntas):
    a. Biología
    b. Ciencias sociales
    c. Matemáticas
    d. Lenguaje
2. Interdisciplinar (15 preguntas)
    a. Violencia y sociedad
    b. Medio ambiente

Contiene los siguientes tipos de resultados:

1. Puntaje en cada prueba del núcleo común (lenguaje, matemáticas, ciencias sociales, biología, filosofía, química, física).
2. Puntaje en cada componente y competencia de cada prueba del núcleo común.
3. Niveles de desempeño en cada componente y competencia de cada prueba del núcleo común.
4. Puntaje y nivel de desempeño en inglés.
5. Puntaje y nivel de desempeño en el área de profundización.

6. Puntaje en la prueba indisciplinar.
7. Puesto.

Para cada prueba del núcleo común se obtiene un resultado cuantitativo expresado en una escala que va de 0 a aproximadamente 100 puntos y se interpreta de acuerdo los tres siguientes rangos:

1. 0 a 30,00→ Bajo
2. 30,01 a 70,00→ Medio
3. Más de 70,0→1 Alto

Extracción, transformación y carga ETL

ETL es un conjunto de procesos que recuperan datos desde un conjunto sistemas fuente, los transforman y los cargan en un sistema de destino. La transformación puede ser usada para cambiar los datos de acuerdo al formato y criterio del sistema destino, generando nuevos valores en el sistema destino. La mayoría de sistemas ETL tienen mecanismos para limpieza de los datos antes de agregarlos, basándose en reglas de calidad de los datos. (Rainardi, 2007)

Los sistemas ETL añaden valor a los datos de las siguientes formas: (Kimball & Caserta, 2004)

1. Remueven errores y corrigen datos vacíos.
2. Proveen medidas de la confiabilidad de los datos.

3. Almacenan el flujo de datos transaccionales de manera segura.
4. Ajustan datos de múltiples fuentes para ser usados en conjunto.
5. Estructuran datos para ser utilizados por herramientas de usuario final.

La arquitectura básica de un sistema de ETL se muestra en la Figura 1.

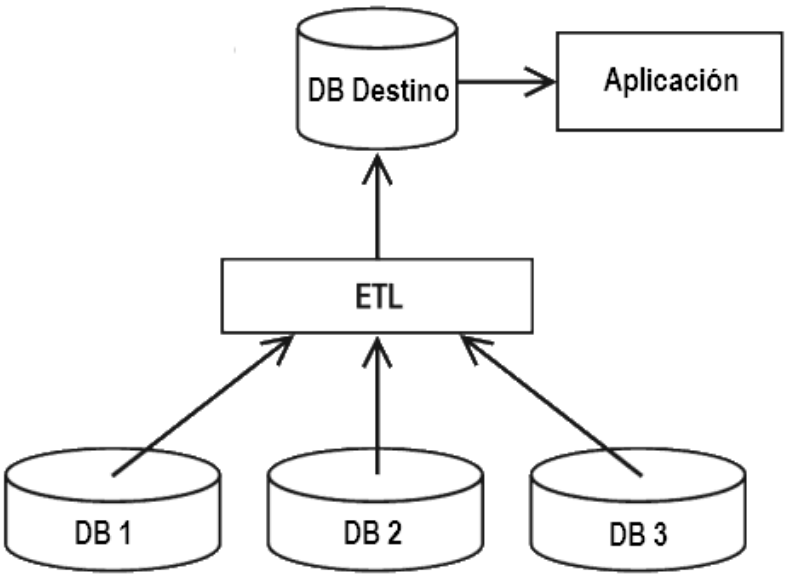

Figura 1. *Arquitectura básica de un sistema ETL*

El proceso de ETL, aunque no es estrictamente necesario para la minería de datos, suele ser necesario cuando se tratan grandes volúmenes de datos, cuando provienen de fuentes heterogéneas, cuando son cambiantes en el tiempo o cuando los algoritmos exigen ciertos formatos en la información de entrada. (Pérez & Santín, 2007)

Minería de Datos y Extracción de conocimiento a partir de datos

Debido al crecimiento en los volúmenes de información que experimentan las empresas y gracias a la disminución del costo en los equipos de computo y los sistemas de almacenamiento, muchas organizaciones cuentan hoy en día con unos vastos conjuntos de datos con información estratégica oculta, y que sin embargo, no puede ser descubierta utilizando los métodos de consulta o estadísticos tradicionales.

La minería de datos utiliza nuevas técnicas para buscar la información, muchas basadas en procesos estadísticos o de inteligencia artificial, que permite identificar patrones y relaciones en la información para crear modelos que permitan caracterizar y hasta predecir el comportamiento de esta; la extracción de conocimiento (KDD, por sus siglas en inglés) abarca la minería de datos, pero además incluye la preparación de la información y la interpretación de los resultados con el fin de dar significado al modelo planteado. (Pujari, 2001)

El conocimiento no se obtiene simplemente por tener un conjunto de datos, estos solamente son la materia prima; que al dársele un significado dentro de un contexto pasa a ser información, y que una vez un experto elabora o identifica un modelo y permite interpretar esa información que aporte un nuevo valor, este pasa a ser conocimiento. (Valhondo, 2003)

Etapas de KDD

Las etapas de KDD inician con un flujo de datos y finalizan con un conocimiento extraído a partir de esos datos. (Pujari, 2001)

1. Selección: seleccionar o segmentar los datos de acuerdo a su relevancia con base en los objetivos planteados para la extracción de conocimiento.
2. Preprocesado: se refiere a la limpieza de la información, eliminando información innecesaria, consolidando información y unificando formatos que provenga de múltiples fuentes.
3. Transformación: consiste en transformar la información recolectada con el objeto de que pueda ser utilizable y navegable (consultar la información desde diferentes puntos de vista) preparándola para la etapa de minería de datos.
4. Minería de datos: extracción de patrones en los datos.
5. Interpretación y evaluación: convertir los patrones encontrados en conocimiento, que sirvan como fundamento para la toma de decisiones.
6. Visualización de datos: busca ofrecer profundidad en el análisis y un entendimiento más intuitivo de los datos y sus patrones definidos en la minería de datos.

La minería de datos es la principal etapa de KDD ya que se enfoca en la búsqueda de las relaciones y patrones ocultos, por lo que se profundiza en el presente documento.

Alcance de la minería de datos

1. Predicción automatizada de tendencias y comportamientos.
2. Descubrimiento automatizado de modelos previamente desconocidos.

La minería de datos produce cinco tipos de información:

1. Asociaciones.
2. Secuencias.
3. Clasificaciones.
4. Agrupamientos.
5. Pronósticos.

Áreas de investigación en la minería de datos

1. Estadísticas: uno de los fundamentos de la minería de datos es la estadística, su amplia difusión y fundamentación teórica la valida, sin embargo, sus resultados deben ser interpretados por expertos debido a su complejidad en la interpretación. Se suelen usar modelos estadísticos como los lineales para llevar a cabo la minería de datos.
2. Máquinas de aprendizaje: es un proceso automatizado de aprendizaje, tomando el aprendizaje como el equivalente a reglas basadas en observaciones, la generalización de comportamientos basada en ejemplos.

3. Aprendizaje supervisado: es el aprendizaje basado en ejemplos, el cual produce una función que establece la correspondencia entre la entrada y la salida deseada (ejemplo).
4. Aprendizaje sin supervisión: es el aprendizaje basado en la observación y el descubrimiento, por lo que no existe un conocimiento a priori, no existe una categorización de la información.
5. Programación matemática

Técnicas de minería de datos

Las técnicas más representativas de la minería de datos son:

1. Descubrimiento de reglas de asociación: utilizado para descubrir hechos que ocurren en común dentro de un contexto, del tipo, cuando un cliente compra un producto X, suele comprar el producto Y.
2. Agrupamiento o Clustering: trata de agrupar datos en diferentes grupos, de tal forma que los datos en cada grupo tengan tendencias o patrones comunes.
3. Descubrimiento de reglas de clasificación o Árboles de decisión: sirven para representar un conjunto de reglas que determinan como se clasifican los elementos, donde cada nueva rama contiene una condición para el conjunto de datos que agrupa.
4. Redes neuronales: forma de aprendizaje basada en la forma en que funciona el sistema nervioso de los animales, basado en la interconexión de neuronas con estímulos ponderados.

5. Regresión lineal: rápida en la búsqueda de relaciones de datos pero limitado su uso a espacios bidimensionales.

Extensiones de la minería de datos

Web Mining: los sitios web y de aprendizaje virtual almacenan conjuntos de información importante que pueden ser sujetos de estudio de la minería de datos, se analizan los logs que registran las acciones de los usuarios, con el fin de encontrar patrones de comportamiento tales como el proceso de navegación de un usuario antes de realizar una compra.

Text mining: gran parte de la información de una organización está almacena en forma de documentos sin ninguna estructura, por lo que la categorización, procesamiento, extracción de información es muy significativa, aún más, teniendo en cuenta que la información esperada no está contenida en un solo documento, sino en su conjunto.

Minería de datos educacional: la minería de datos aplicada al sector educativo, utiliza fuentes de datos como los registros de calificaciones, los logs de los estudiantes en los sistemas de educación virtual y la información descriptiva del estudiante para la búsqueda de patrones o modelos que permitan mejorar el desempeño académico, disminuir la deserción, así como mejorar la utilización de los recursos.

RapidMiner

Es un entorno para procesos de minería de datos y máquinas de aprendizaje, con una arquitectura modular que permite el diseño de análisis mediante el encadenamiento de operadores a través de un entorno gráfico. Hoy en día es el líder en los sistemas de minería de datos de código abierto y usado ampliamente por investigadores y empresas. (Rapid-i). En una encuesta realizada por KDnuggets, un periódico de minería de datos, RapidMiner ocupó el segundo lugar en herramientas de analítica y de minería de datos utilizadas para proyectos reales en 2009 y fue el primero en 2010. (Wikipedia)

La versión inicial fue desarrollada por el departamento de inteligencia artificial de la Universidad de Dortmund en 2001. Se distribuye bajo licencia AGPL y está hospedado en SourceForge.

RapidMiner proporciona más de 500 operadores orientados al análisis de datos, incluyendo los necesarios para realizar operaciones de entrada y salida, preprocesamiento de datos y visualización. También permite utilizar los algoritmos incluidos en Weka y desde octubre de 2010 posee integración con R.

Características principales

1. Integración de datos, ETL analítico, análisis de datos, integrado en un solo aplicativo.
2. Poderosa e intuitiva interface gráfica para el diseño de procesos de análisis.

3. Repositorio de procesos, datos y metadatos.
4. Transformación de metadatos, inspección en tiempo de diseño de los datos y sus metadatos.
5. Reconocimiento de errores y propuestas de ajustes rápidos.
6. Completo y flexible: Cientos de métodos para la carga de datos, su transformación, su modelamiento y su visualización.

Proyecto R para estadística computacional

R es un lenguaje y entorno de programación para análisis estadístico y gráfico. (Proyecto R para estadística computacional)

Se trata de un proyecto de software libre, resultado de la implementación GNU del premiado lenguaje S. R y S-Plus -versión comercial de S- son, probablemente, los dos lenguajes más utilizados en investigación por la comunidad estadística, siendo además muy populares en el campo de la investigación biomédica, la bioinformática y las matemáticas financieras. A esto contribuye la posibilidad de cargar diferentes bibliotecas o paquetes con finalidades específicas de cálculo o gráfico. (Wikipedia)

Características

R proporciona un amplio abanico de herramientas estadísticas (modelos lineales y no lineales, tests estadísticos, análisis de series temporales, algoritmos de clasificación y agrupamiento, etc.) y gráficas.

Al igual que S, se trata de un lenguaje de programación, lo que permite que los usuarios lo extiendan definiendo sus propias funciones. De hecho, gran parte de las funciones de R están escritas en el mismo R, aunque para algoritmos computacionalmente exigentes es posible desarrollar bibliotecas en C, C++ o Fortran que se cargan dinámicamente. Los usuarios más avanzados pueden también manipular los objetos de R directamente desde código desarrollado en C. R también puede extenderse a través de paquetes desarrollados por su comunidad de usuarios.

R hereda de S su orientación a objetos. La tarea de extender R se ve facilitada por su permisiva política de lexical scoping.

Además, R puede integrarse con distintas bases de datos y existen bibliotecas que facilitan su utilización desde lenguajes de programación interpretados como Perl y Python.

Otra de las características de R es su capacidad gráfica, que permite generar gráficos con alta calidad. R posee su propio formato para la documentación basado en LaTeX.

R también puede usarse como herramienta de cálculo numérico, campo en el que puede ser tan eficaz como otras herramientas específicas tales como GNU Octave y su versión comercial, MATLAB.4 Se ha desarrollado una interfaz, RWeka5 para interactuar con Weka que permite leer y escribir ficheros en el formato arff y enriquecer R con los algoritmos de minería de datos de dicha plataforma.

A partir de octubre de 2010, los proyectos R y RapidMiner proporcionan una integración, que permite en los procesos diseñados en RapidMiner incorporar análisis y procesos de los paquetes de R.

Regresión Logística Multinomial

Introducción

El modelo de regresión logística permite estimar la probabilidad de un suceso que depende de los valores de ciertas covariables (Cuadras, 2010).

Supongamos que un suceso (o evento) de interés A puede presentarse o no en cada uno de los individuos de una cierta población. Consideremos una variable binario y que toma los valores:

y = 1 si A se presenta, y = 0 si A no se presenta

Si la probabilidad de A no depende de otras variables, indicando P(A) = p; la verosimilitud de una única observación y es

$$L = p^y(1-p)^{1-y}$$

pues L = p si y = 1; L = 1 p si y = 0:

Si realizamos n pruebas independientes y observamos y1,…, yn, la verosimilitud es

$$L = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} = p^k(1-p)^{n-k}$$

siendo k = Sumatoria(yi) la frecuencia absoluta de A en las n pruebas. Para estimar p resolvemos la ecuación de verosimilitud

$$\frac{\partial}{\partial p} \ln L = 0$$

cuya solución es ^p = k/n; la frecuencia relativa del suceso A: La distribución asintótica de ^p es normal N(p, p(1 - p)/n).

Muy distinta es la estimación cuando esta probabilidad depende de otras variables. La probabilidad de A debe entonces modelarse adecuadamente.

Muy distinta es la estimación cuando esta probabilidad depende de otras variables. La probabilidad de A debe entonces modelarse adecuadamente.

Modelo de regresión logística (Cuadras, 2010)

Supongamos ahora que la probabilidad p depende de los valores de ciertas variables X1;….;Xp: Es decir, si x = (x1;….; xp)' son las observaciones de un cierto individuo w sobre las variables, entonces la probabilidad de acontecer A dado x es p(y = 1|x): Indicaremos esta probabilidad por p(x): La probabilidad contraria de que A no suceda dado x ser· p(y = 0|x) = 1- p(x).

Es fácil darse cuenta que pretender que p(x) sea una función lineal de x no puede funcionar correctamente, pues p(x) está comprendido entre 0 y 1.

Por diversas razones, es muy conveniente suponer un modelo lineal para la llamada transformación logística de la probabilidad.

$$\ln[\frac{p(\mathbf{x})}{1-p(\mathbf{x})}] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \beta_0 + \beta'\mathbf{x}$$

Siendo $\beta = (\beta_1, \cdots, \beta_p)'$ parámetros de regresión: El modelo equivale a suponer las siguientes probabilidades para A y su contrario, ambas en función de x

$$p(\mathbf{x}) = \frac{e^{\beta_0+\beta'\mathbf{x}}}{1+e^{\beta_0+\beta'\mathbf{x}}}, \quad 1-p(\mathbf{x}) = \frac{1}{1+e^{\beta_0+\beta'\mathbf{x}}}.$$

Hagamos ahora una breve comparación con el modelo lineal. El modelo de regresión lineal es

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + e$$

donde se supone que *y* es una variable respuesta cuantitativa y que *e* es un error con media 0 y varianza σ^2. Usando la misma terminología, podemos entender el modelo logístico en el sentido de que

$$y = p(\mathbf{x}) + e$$

donde ahora y sólo toma los valores 0 ó 1. Si *y* = 1 entonces *e* = 1 - p(x) con probabilidad p(x). Si y = 0 entonces *e* = -p(x) con probabilidad 1 - p(x). De este modo, dado x; el error *e* tiene media 0 y varianza p(x)(1 - p(x)).

Dado un individuo *w*, la regla de discriminación logística (suponiendo los parámetros conocidos o estimados) simplemente decide que *w* posee la característica A si p(x) > 0,5; y no la posee si p(x) <= 0,5 Introduciendo la función discriminante

$$L_g(\mathbf{x}) = \ln\left(\frac{p(\mathbf{x})}{1 - p(\mathbf{x})}\right)$$

la regla de decisión logística es

Si $L_g(x) > 0$ entonces *y* = 1; si $L_g(x) <= 0$ entonces *y* = 0.

Estimación de los parámetros. (Cuadras, 2010)

La verosimilitud de una observación y es

$$L = p(\mathbf{x})^y (1 - p(\mathbf{x}))^{1-y}$$

La obtención de n observaciones independientes

$$(y_i, \mathbf{x}_i) = (y_i, x_{i1}, \ldots, x_{ip})$$

se puede tabular matricialmente como

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

Nótese que, para poder tener en cuenta el término constante β_0 en el modelo, la primera columna de X contiene unos.

La verosimilitud de n observaciones independientes es

$$L = \prod_{i=1}^{n} p(\mathbf{x}_i)^{y_i} (1 - p(\mathbf{x}_i))^{1-y_i}$$

Tomando logaritmos

$$\ln L = \sum_{i=1}^{n} y_i \ln p(\mathbf{x}_i)(1 - p(\mathbf{x}))^{1-y_i}$$

A fin de hallar los estimadores máximo verosímiles de los parámetros β deberemos resolver las ecuaciones

$$\frac{\partial}{\partial \beta_j} \ln L = 0, \quad j = 0, 1, \ldots, p$$

Se tiene $\ln p(\mathbf{x}_i) = \beta_0 + \beta_1 \mathbf{x}_i - \ln(1 + e^{\beta_0 + \beta_1 \mathbf{x}_i})$, luego

$$\begin{array}{lll} \frac{\partial}{\partial \beta_0} \ln p(\mathbf{x}_i) & = 1 - \frac{e^{\beta_0 + \beta' \mathbf{x}_i}}{1 + e^{\beta_0 + \beta' \mathbf{x}_i}} & = 1 - p(\mathbf{x}_i) \\ \frac{\partial}{\partial \beta_j} \ln p(\mathbf{x}_i) & = x_{ij} - x_{ij} \frac{e^{\beta_0 + \beta' \mathbf{x}}}{1 + e^{\beta_0 + \beta' \mathbf{x}_i}} & = x_{ij}(1 - p(\mathbf{x}_i)) \end{array}$$

Análogamente derivaríamos

$$\ln(1 - p(\mathbf{x}_i)) = -\ln(1 + e^{\beta_0 + \beta_1 \mathbf{x}_i})$$

Se obtienen entonces las ecuaciones de verosimilitud para estimar los parámetros β,

$$\begin{array}{l} \sum_{i=1}^{n} (y_i - p(\mathbf{x}_i)) = 0, \\ \sum_{i=1}^{n} x_{ij}(y_i - p(\mathbf{x}_i)) = 0, \quad j = 1, \ldots, p \end{array}$$

Utilizando el vector y; la matriz X y el vector de probabilidades

$$\pi(\mathbf{X}) = (p(\mathbf{x}_1) \ldots, p(\mathbf{x}_n))',$$

estas ecuaciones se pueden escribir como $\mathbf{X}'\pi(\mathbf{X}) = \mathbf{X}'\mathbf{y}$,

siendo comparables con las ecuaciones normales $\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{Y}$, para estimar los parámetros β del modelo lineal $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$, salvo que ahora el modelo $\mathbf{X}\beta$ es $\pi(\mathbf{X})$, que depende de β. Sin embargo las ecuaciones no se pueden resolver explícitamente, debiéndose recurrir a procedimientos numéricos iterativos.

Significación de los resultados de la regresión logística

La significación de los coeficientes de regresión logística se prueba en forma análoga a otros estadísticos, en particular a los coeficientes de regresión ordinaria, siendo la hipótesis nula que el correspondiente parámetro de población es igual a cero y las diferencias observadas se han debido al azar del muestreo. Sin embargo, en el caso de la regresión logística las pruebas no son tan seguras como en el de la regresión ordinaria y exigen mayores precauciones, en particular tratándose de muestras pequeñas. En este sentido, los valores de P obtenidos en las pruebas de significación de los respectivos coeficientes, no siempre guardan una relación estrecha con la fuerza o la importancia de una variable en el modelo. Un estadístico de prueba para los

coeficientes de regresión es el de Wald, igual al cuadrado del cociente entre el coeficiente de regresión logística y su error estándar, que tiene una distribución chi-cuadrado con un grado de libertad. Se halla incluido en los paquetes estadísticos corrientes. Existen otros métodos que apuntan a asegurar la validez de los hallazgos, en relación a limitaciones como el tamaño de las muestras (Esper & Machado, 2008).

Como ya se ha dicho, la estimación de los coeficientes de regresión logística se basa en una técnica llamada de máxima verosimilitud, que utiliza para el cálculo una función llamada con el mismo nombre (maximum likelihood function) y que permite no solamente obtener los logits sino también evaluar en forma global el grado de "ajuste" del modelo resultante a los datos muestrales. Esto permite a su vez calibrar la importancia de cada variable en el resultado de la regresión, observando el grado de cambio en el ajuste global de la regresión luego de introducir o retirar la variable en cuestión (el caso análogo en la regresión ordinaria consiste en observar los cambios en el coeficiente de correlación múltiple R^2 con el agregado o retiro de variables). Esta es otra forma de evaluar la significación de los coeficientes de regresión logística. Como se ve, existen numerosas alternativas para hacerlo, lo que indica que no hay un procedimiento único y directo, y analizar los resultados requiere alguna experiencia, así como cierto conocimiento previo del material estudiado. Lo dicho debe hacer tomar precauciones ante la obtención de conclusiones

apresuradas o sin un análisis cuidadoso de su significado en el contexto de la investigación (Esper & Machado, 2008).

Algunas propiedades de la regresión logística.

Como se ha dicho, la regresión logística proporciona ecuaciones con ciertas analogías a las de regresión lineal ordinaria o de mínimos cuadrados tratadas en secciones anteriores, donde la significación estadística de los coeficientes de regresión expresa la importancia de la correspondiente variable independiente en la determinación de la probabilidad de la variable dependiente binaria. En el caso de la regresión logística, las correspondientes ecuaciones consisten en logits, expresiones lineales en función de las variables independientes, que es común convertir en odds o en probabilidades para facilitar su interpretación. Como se ha dicho, en general es preferible trabajar con muestras no demasiado pequeñas (Esper & Machado, 2008).

Algunas utilidades de la regresión logística. Odds ratios

Las aplicaciones de la regresión logística son análogas a las de la regresión ordinaria y algunas han sido esbozadas en lo que antecede. En principio, la determinación del efecto de las variables independientes o predictoras sobre las probabilidades de la variable dependiente, la comparación de la magnitud de

dichos efectos y la evaluación de la independencia estadística de las variables predictoras son aplicaciones importantes. Así, se vio que la ecuación de regresión logística que relaciona la variable dependiente E con el colesterol plasmático, expresada como logits de E, permite pasar a odds de E y a probabilidad de E, para cualquier valor del colesterol. Si la regresión resulta estadísticamente significativa, se puede descartar la hipótesis nula de no asociación entre enfermedad y colesterol, y darle plena utilidad. En este punto puede agregarse otra variable al modelo y evaluar si es significativa como la variable colesterol, y si esta conserva su significación luego de haber introducido la nueva variable. Si es así, se podrá afirmar que ambas variables son predictores independientes de E. También podrá ocurrir que una de las dos variables no sea significativa en presencia de la otra por compartir información, y los conceptos vistos en la Sección 11 son en general aplicables también en regresión logística. La comparación del efecto de distintas variables predictoras, el estudio de sus interrelaciones y la comprobación de hipótesis, son aplicaciones corrientes de la regresión logística. Algunas de las situaciones que pueden dificultar la obtención de resultados válidos son la omisión de variables relevantes, la inclusión de variables que no lo son, la colinearidad entre variables independientes, y otras condiciones de índole formal sobre las que no es posible extenderse, que pueden eventualmente llegar a infringir los supuestos teóricos de la técnica (Esper & Machado, 2008).

Comprensión del Dominio

Objetivos institucionales

Los objetivos institucionales para la Universidad buscan disminuir la problemática descrita en la sección Descripción del Área Problemática y Justificación, sin embargo, los objetivos esenciales perseguidos se presentan a manera de resumen para dar continuidad con la metodología CRISP-DM.

1. Reducir la tasa de deserción.
2. Disminuir la tasa de repitencia.
3. Disminuir el tiempo de permanencia en la Universidad.
4. Mejorar el rendimiento académico.

Evaluación de la Situación

1. Recursos disponibles: El único recurso disponible es el autor del proyecto, el cual ayudó a construir el Sistema de Información Académica desde sus inicios y que utilizó intensivamente, en su cargo como Jefe de la Oficina de Admisiones y Registro Académico de dicha institución.
2. Datos: Sólo se tendrá acceso al Sistema de Información Académica SIA de la Universidad de Caldas autorizado por el Vicerrector Académico Msc. Germán

Gómez. Dicho acceso sólo será de consulta de los datos en un momento inicial, supeditado a la supervisión del Jefe de Registro Académico vigente.

La información con la que se cuenta en dicha base de datos de manera general es:

1. Caracterización poblacional demográfica del estudiante. (edad, sexo, procedencia).
2. Descripción de las materias.
3. Resultados ICFES de cada estudiante.
4. Registro de notas de cada estudiante.
5. Registro de los horarios de los cursos.

Recursos Computacionales

El único recurso computacional para el procesamiento y análisis de la información, será el proporcionado por el autor del proyecto, el cual es un computador portátil Dell 6400 Core Duo 1.66 GHz con 3.24 GB de RAM.

Software: Deberá utilizarse software de uso gratuito, como bases de datos, sistemas de minería de datos, aplicaciones estadísticas y otras que puedan aplicar.

La publicación de los resultados no deberá contener información sensible de los estudiantes.

Antes del año 2000 existía otra escala de evaluación de las pruebas ICFES, a partir del año 2010 estás pruebas no son válidas para presentarse a la Universidad de Caldas.

Terminología

Crédito académico: Un Crédito Académico es la unidad que mide el tiempo estimado de actividad académica del estudiante en función de las competencias profesionales y académicas que se espera que el programa desarrolle. (Altablero. Ministerio de Educación Nacional de Colombia, 2001)

El Crédito Académico equivale a 48 horas totales de trabajo del estudiante, incluidas las horas académicas con acompañamiento docente y las demás horas que deba emplear en actividades independientes de estudio, prácticas, preparación de exámenes u otras que sean necesarias para alcanzar las metas de aprendizaje propuestas, sin incluir las destinadas a la presentación de exámenes finales.

PBM: se refiere al puntaje básico de matrícula, con el cual se establece el valor de los derechos de matrícula de pregrado de los alumnos de la Universidad de Caldas, de acuerdo a la situación socioeconómica del estudiante. El valor de la matrícula se calcula como el producto del factor que le corresponda a cada PBM según la tabla definida en el acuerdo 24 de 2002 del

Consejo Superior, por el salario mínimo legal mensual del año inmediatamente anterior. (Consejo Superior. Universidad de Caldas., 2002)

Está compuesto por dos grupos de variables; el primero denominado variables socioeconómicas que incluye el estrato y los ingresos familiares; el segundo denominado atenuantes que incluye el lugar de residencia, el número de hijos dependientes del ingreso familiar y el carácter del colegio.

SIA, Sistema de Información académica: Es el sistema de información utilizado por la Universidad de Caldas para la gestión de toda su información relacionada con la academia, comprende módulos para aspectos como registro de aspirantes, selección de admitidos, cálculo y generación de matrícula financiera, registro de estudiantes, registro de programas, registro de materias, registro de planes de estudio, registro de calificaciones, registro de egresados entre otros.

Determinación de los objetivos de la minería de datos

1. Definir y clasificar el indicador del rendimiento académico.
2. Determinar los factores de influyen en el indicador de rendimiento académico basados en la información alojada en el Sistema de Información Académica SIA.

Comprensión de los datos

Recolección de datos iniciales

El conjunto de datos se recolectó completamente del Sistema de Información Académica SIA, que a pesar de ser un solo un repositorio de datos, la información se encuentra distribuida en tablas o en formatos no adecuados para su procesamiento.

1. El acceso a los datos se realizó mediante conexión por Terminal Server con el servidor en que se encuentra alojado el SIA.
2. Se utilizó TOAD for Oracle para exportar los datos de la base de datos Oracle a archivos planos.
3. Se exportaron las siguientes tablas que contienen la información demográfica del estudiante, los colegios de graduación de los estudiantes, los puntajes ICFES, el registro de notas, la información de las asignaturas cursadas, los horarios de los cursos.
4. Para estudiantes y notas, sólo se exportaron los registros que tengan relación con estudiantes que hayan ingresado a partir del año 2001
   a. estudiantes: información sobre los estudiantes.
   b. forad (año)(periodo:1 o 2) (20011, 20012, 20021, 20022, 20031, 20032, 20041, 20042, 20051, 20052, 20061, 20062, 20071, 20072, 20081,

20091, 20092, 20101, 20102, , 20111): información de los resultados ICFES.

c. materias: información sobre las materias.
d. clases_encabezado: fechas de la programación de cursos.
e. colegios: información sobre los colegios.
f. egresados: información de los estudiantes que se han graduado.
g. notas: registro de calificaciones de los estudiantes.

5. Se presentaron errores al exportar los puntajes ICFES pues algunos campos contenían comas o saltos de línea que interferían en los archivos planos, se ajustaron estos utilizando un editor de texto que permitiera el reemplazo con expresiones regulares, para el caso Notepad++.

Descripción de los datos

En la Tabla 1 se describen las tablas extraídas del SIA.

Tabla 1. *Tablas extraídas del SIA*

Tabla	Descripción	Registros
estudiantes	Información básica sobre el estudiante. Los estudiantes que hayan ingresado a partir de 2001	23.461
CODIGO (*)	Código del estudiante.	
SEM_INGRES	Semestre de ingreso. Año + período (1 ó 2).	
COD_PENSUM	Código del pensum asignado.	
COD_CARRERA	Código del programa académico.	
NOM_CARRERA	Nombre del programa académico.	
COD_FACULTAD FACULTAD	Código de la facultad a la que pertenece el programa académico.	

Tabla	Descripción	Registros
SEXO	Nombre de la facultad a la que pertenece el programa académico.	
ESTRATO	Sexo del estudiante (F o M).	
DIA_NAC	Estrato socioeconómico.	
MES_NAC	Día de nacimiento.	
ANO_NAC	Mes de nacimiento.	
CIU_CORREO	Año de nacimiento.	
DPTO_CORREO	Ciudad de residencia.	
TIPO_COLEGIO	Departamento de residencia.	
NOM_COLEGIO	Tipo de colegio donde finalizó bachillerato. (1: Oficial, 2: Privado).	
COD_COLEGIO	Nombre del colegio.	
CIU_COLEGIO	Código del colegio.	
DPTO_COLEGIO	Ciudad del colegio.	
SNP	Departamento del colegio.	
PBM	Código del estudiante en las pruebas ICFES.	
TIPO_SNP	Puntaje básico de matrícula.	
FECHA_NAC	Tipo de prueba ICFES (N: A partir de 2000 o V: Antes de 2000).	
	Fecha de nacimiento	
notas	Registro de calificaciones	834.998
CODIGO (*)	Código del estudiante.	
COD_MATERIA (*)	Código de la materia.	
GRUPO	Grupo del curso.	
NOTA_DEF	Nota definitiva (0 a 50)	
NOTA_HAB	Nota de habilitación (0 a 50)	
NOTA_RECUPERA	Nota de recuperación. Sólo aplica para tecnologías (0 a 50)	
HORAS	Número de horas que asistió para materias cualitativas.	
NOTA_FINAL	Nota final. (0 a 50) Es la mayor entre nota definitiva o nota de habilitación. Para materias cualitativas es 50 para Aprobado o 0 para Reprobado.	
FALLAS_S	Número de fallas sin excusa.	
FALLAS_C	Número de fallas con excusa.	
VEZ	Vez en que cursa el estudiante la materia.	
ANO_CURSO (*)	Año en que cursa la materia.	
PERIODO (*)	Período del año en que cursa la materia (1 ó 2).	
materias	Información sobre las materias	15.880

Tabla	Descripción	Registros
COD_MATERIA (*) NOM_MATERIA H_TEORICAS H_PRACTICAS HABILITABLE H_NOPRESEN COD_DEPTO CREDITOS	Código de la materia. Nombre de la materia. Horas teóricas. Horas prácticas, Indica si es habilitable. (S o N) Horas no presenciales. Código del departamento al que pertenece la materia. Número de créditos.	
clases_encabezado	Programación de los cursos para las materias	72.494
COD_MATERIA (*) ANO (*) PERIODO (*) GRUPO (*) FECHA_INICIO FECHA_FINAL	Código de la materia. Año de la programación del curso. Período de la programación del curso. Grupo. Fecha de inicio del curso. Fecha de finalización del curso.	
colegios	Información sobre los colegios	9.575
COL_CODIGO (*) COL_NOMBRE COL_NATURALEZA CIUD_ID CIUD_NOMBRE CIUD_DEPARTAMENTO	Código del colegio. Nombre del colegio. Naturaleza: (Privado, Oficial, Público) Código DANE de la ciudad en la que se encuentra el colegio. Ciudad en la que se encuentra el colegio. Departamento en el que se encuentra el colegio.	
egresados	Información sobre el grado de un estudiante	5.195
CODIGO (*) FECHA_GRADO	Código del estudiante que se graduó. Fecha en la que se graduó.	
forad (año+periodo)	Información sobre el ICFES y la admisión de un aspirante	92.605
ANO (*) PERIODO (*) ADMITIDO PONDERADO PONDERADO_OPC	Año de inscripción del aspirante al proceso de selección en la Universidad de Caldas. Período de inscripción. Indica si fue admitido (P: Primera opción, S: Segunda opción, vacío: No admitido) Ponderado obtenido en primera opción.	

Tabla	Descripción	Registros
TIPO_SNP	Ponderado obtenido en segunda opción.	
CODIGO	Tipo de prueba ICFES (N: A partir de 2000 o V: Antes de	
NBIOLOGIA	2000).	
NFILOSOFIA	Código de estudiante si es admitido.	
NHISTORIA	Puntaje en la prueba de Biología de ICFES a partir de 2000.	
NLENGUAJE	Puntaje en la prueba de Biología de ICFES a partir de 2000.	
NMATEMATICAS	Puntaje en la prueba de Biología de ICFES a partir de 2000.	
NFISICA	Puntaje en la prueba de Biología de ICFES a partir de 2000.	
NQUIMICA	Puntaje en la prueba de Biología de ICFES a partir de 2000.	
NGEOGRAFIA	Puntaje en la prueba de Biología de ICFES a partir de 2000.	
NINTERDISCIPLI	Puntaje en la prueba de Biología de ICFES a partir de 2000.	
NAR	Puntaje en la prueba de Biología de ICFES a partir de 2000.	
COL_CODIGO	Puntaje en la prueba de Biología de ICFES a partir de 2000.	
CIU_CORREO	Código del colegio es que estudiaba el aspirante cuando	
SNP (*)	presentó la prueba.	
	Ciudad de residencia del aspirante	
	Código del aspirante en las pruebas ICFES.	

Exploración y Validación de los datos

Para el proceso de limpieza inicial se utilizó la herramienta Google Refine, que permite explorar de manera ágil los datos y realizar ajustes con múltiples opciones disponibles. (Google)

Se listan las circunstancias encontradas al examinar y validar los datos:

1. Estudiantes sin PBM (309).
2. Estudiantes no tienen ciudad de origen.
3. Estudiantes sin código SNP.
4. Estudiantes con el campo de fecha de nacimiento vacío.
5. Estudiantes con ICFES antes de 2000.

6. Estudiantes con códigos de pensum 999 que no pertenecen a ningún programa académico.
7. Estudiantes sin código SNP.
8. Estudiantes con ciudades y departamentos de procedencia escritos manualmente por lo que se encuentran múltiples variaciones de la misma procedencia.
9. Sólo existen 22 estudiantes de ingreso en 2000 segundo período con ICFES a partir de 2000.
10. Estudiantes sin estrato (57).
11. Estudiantes sin ponderado (63).
12. Colegios sólo con id de ciudad pero sin nombre de ciudad.
13. Notas de estudiantes de Medicina reportados en un sólo período de un año pero que realmente fueron tomados durante los 2 períodos del año.
14. Estudiantes con más de 15 materias por período académico debido a homologaciones por reingresos o transferencias.

Preparación de los datos

Selección de los datos y Limpieza de datos

Para la selección y limpieza de la información se continuó el uso de Google Refine y se creó una base de datos en SQL Server Express en la cual aplicar ajustes a los datos mediante la ejecución de consultas SQL (Anexo 1 Script SQL para el preprocesamiento en motor SQL).

Se aplicaron los siguientes ajustes a los datos antes de aplicar las reglas de exclusión.

1. Asignar las ciudades faltantes de los colegios a partir del ID de ciudad.
2. Asignar las ciudades de origen faltantes de los estudiantes a partir de la ciudad del colegio.
3. Asignar las ciudades de origen faltantes del estudiante a partir de la tabla de admisión FORAD.
4. Asignar el tipo de SNP faltante en estudiantes a partir de la tabla de admisión.
5. Asignar la fecha de nacimiento faltante en estudiantes a partir de los campos día, mes y año de nacimiento.
6. Asignar el departamento faltante en estudiantes a partir de la tabla ciudades.

7. Ajustar la ciudad y el departamento de origen que fueron mal escritos utilizando las funcionalidades (Facet, edición masiva, Cluster Ngram) de Google Refine y ajustándolo al nombre real.
8. Asignar los PBM faltantes calculándolo basado en la liquidación del 2002 y utilizando el salario mínimo vigente en ese año ($309.000).
9. Para estudiantes cuyo semestre de ingreso aparecía 20033 y 20043, se reasignó por 20032 y 20042 respectivamente.
10. Estudiantes que aparecían con múltiples registros en la tabla de admisiones, se tomó sólo una y se creó el campo reingreso acumulando el número de veces que aparece.
11. Cambiar el tipo de colegio Privado Diurno por Privado y Oficial o Nocturno por Oficial.
12. Calcular y asignar el período en que se tomó una asignatura basado en la programación de los cursos (clases_encabezado).
13. Asignar el semestre de ingreso de acuerdo al registro de notas.
14. Calcular la edad de ingreso del estudiante basado en la fecha de nacimiento y el semestre de ingreso.

Se aplicaron las siguientes reglas de exclusión:

1. Estudiantes con ICFES anterior a 2000.
2. Estudiantes con pensum 999.
3. Estudiantes con ingreso anterior al año 2001 (22).
4. Estudiantes sin estrato (57).

5. Estudiantes sin PBM (17).
6. Estudiantes de otros países (1 EE.UU.)
7. Estudiantes sin puntajes ICFES (35).
8. Estudiantes sin ponderado (63).
9. Notas sin estudiantes relacionados.

Construcción de datos e Integración de datos

Esta etapa tuvo como objetivo construir un conjunto de datos o vista minable que tiene la siguiente estructura por estudiante (1 sólo registro por estudiante):

1. Información de identificación académica del estudiante: código, programa, facultad, pensum, semestre de ingreso, estado académico en el período del análisis (estudiante, retirado o graduado). Para el presente análisis es el segundo período de 2010.
2. Información demográfica: sexo, edad, ciudad, departamento, nivel de la ciudad (capital o municipio), tipo de colegio (oficial o privado).
3. Información socioeconómica del estudiante: estrato, PBM.
4. Información de ingreso: puntajes en pruebas ICFES, tipo de admisión (primera opción o segunda opción), ponderado de admisión.
5. Información sobre grado: fecha de graduación y meses transcurridos entre el ingreso y la graduación, cuando aplique.

6. Información sobre los cursos realizados: la información sobre los cursos realizados por un estudiante se agrupa por segmentos en avance de la carrera sin acumularse, para mostrar la evolución del rendimiento a lo largo de su paso por la universidad, así:
   a. Primer semestre.
   b. Entre el 0 y el 25% de los créditos del programa.
   c. Entre el 25% y el 50% de los créditos del programa.
   d. Entre el 50% y el 75% de los créditos del programa.
   e. Entre el 75% y el 100% de los créditos del programa.
   f. General: Entre el 0 y el porcentaje de los créditos del programa en que se encuentre.

   Para cada uno de estos segmentos se calcularon las siguientes variables:

   a. Promedio general.
   b. Promedio de las materias ganadas.
   c. Créditos inscritos.
   d. Créditos ganados.
   e. Créditos perdidos.
   f. Materias inscritas.
   g. Materias ganadas.
   h. Materias perdidas.
   i. Semestres en que realizó inscripción.

7. Índices: basados en el estudio de (Garnica, 1997): se calculan los siguientes índices, para los mismos segmentos definidos para los cursos, que reflejan la excelencia, la velocidad de los estudios y la eficacia en los estudios
   a. IMA, índice de materias aprobadas: número de materias aprobadas sobre el número de materias inscritas. Eficacia.
   b. ICA, índice de créditos aprobados: número de créditos aprobados sobre el número de créditos inscritos. Eficacia.
   c. PCC, Promedio de créditos inscritos por semestre: número de créditos inscritos sobre número de períodos inscritos por segmento. Eficiencia.
   d. Promedios: el promedio general y promedio de las materias ganadas reflejan la excelencia.
8. Indicador académico de rendimiento IAR: este indicador se calculará igualmente por segmento y será el promedio de los diferentes índices normalizados entre 0 y 1.

La construcción de los datos se realizó en 2 partes:

1. En motor SQL: el procesamiento inicial se realizó en el motor de base de datos (se utilizó SQL Server Express simulando la base de datos original del SIA en Oracle, pues esta no es gratuito), debido al alto volumen de registros de notas (más de 800.000 registros). En este se harán cálculos básicos y agrupaciones sobre el conjunto de notas.
2. En RapidMiner: validaciones, pivote de datos y generación de nuevas variables.

En motor SQL se realizaron los siguientes procesos:

1. Script SQL para el preprocesamiento en motor SQL el cual puede ser consultado en el

2. Cálculo del segmento de avance al que pertenece cada periodo inscrito por un estudiante, así:

a. Cálculo de créditos necesarios para graduarse por programa académico y semestres de ingreso.

b. Cálculo de los créditos aprobados por período.

c. Cálculo de los créditos aprobados acumulados por período.

d. Como ejemplo, para un estudiante se obtiene una tabla como el ejemplo de la Tabla 2.

3. Cálculo de la información de los cursos realizados por segmento de avance.

a. Promedio general.

b. Promedio de las materias ganadas.

c. Créditos inscritos.

d. Créditos ganados.

e. Créditos perdidos.

f. Materias inscritas.

g. Materias ganadas.

h. Materias perdidas.

i. Semestres en que realizó inscripción.

4. Cálculo del estado del estudiante en el período de análisis.

5. Cálculo del segmento máximo en que se encuentra un estudiante en el período de análisis.
6. Cálculo del número de meses entre en ingreso y el grado, si aplica.
7. Cálculo del número máximo de homologaciones por semestre, con el fin de poder definir un criterio de selección para los estudiantes que reingresaron o hicieron transferencia de otra universidad o programa académico.

Tabla 2. *Ejemplo de asignación de segmentos de avance por período académico*

Código de estudiante	Año	Período	Segmento
1	2005	1	0
1	2005	2	25
1	2006	1	25
1	2006	2	50
1	2007	1	50
1	2007	2	75
1	2008	1	75
1	2008	2	75
1	2009	1	100

Código de estudiante	Año	Período	Segmento
1	2009	2	100

Salida final del motor SQL tiene 2 conjuntos de datos:

Información del estudiante: 1 registro por estudiante con la información de identificación académica, demográfica, socioeconómica, ingreso, grado.

Información sobre los cursos realizados: 1 registro por estudiante y por segmento con la información sobre excelencia, eficiencia y eficacia.

En RapidMiner

Se creó el flujo de procesamiento general (Figura 2), que va desde la carga de los dos conjuntos de datos generados en el motor SQL hasta la ejecución del modelamiento, además se creó el flujo para el cálculo de las correlaciones y las agrupaciones para el informe descriptivo de los datos (Figura 3) (Anexo 2 Proceso RapidMiner).

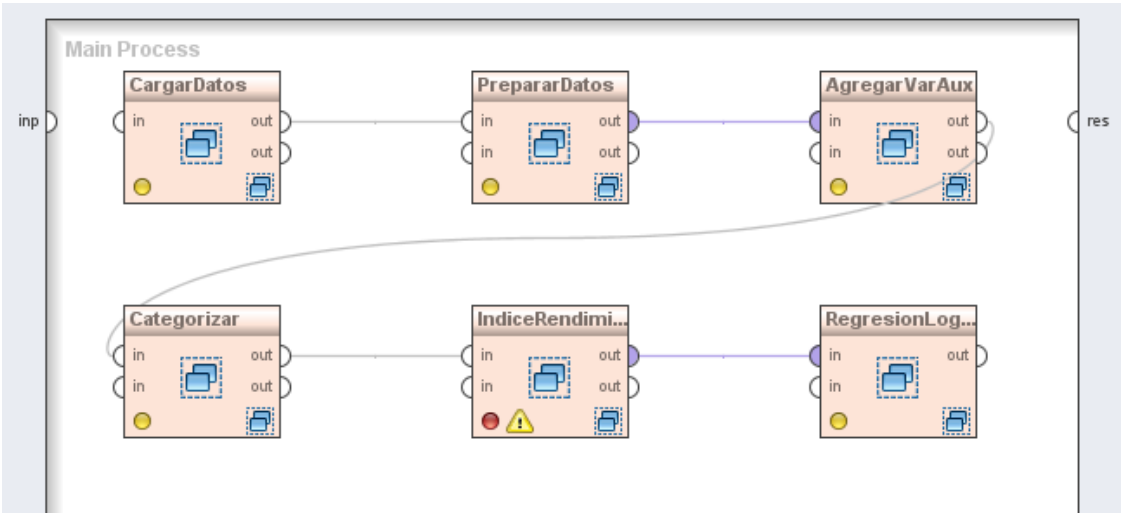

Figura 2. *RapidMiner - Flujo general*

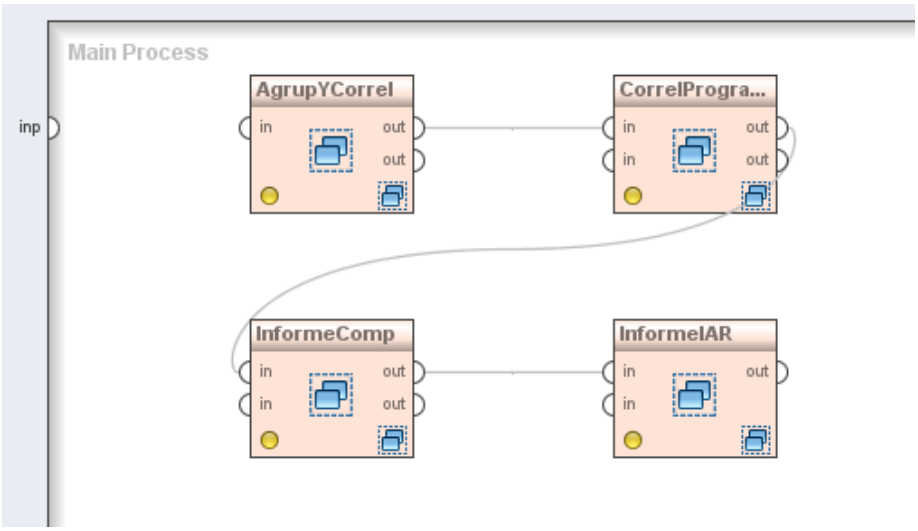

Figura 3 *Flujo para correlaciones y agrupaciones para el informe*

Descripción de los bloques del flujo general:

CargarDatos: carga los dos conjuntos de datos del motor SQL y los almacena en repositorios locales.

PrepararDatos: valida los datos cargados, transforma el conjunto de datos de notas en un sólo registro por estudiante, renombrando los campos de acuerdo al porcentaje de avance en créditos al que pertenezcan, así: variable + "_" + segmento de avance (e.g.: promedio_0, promedio_25, promedio_50, promedio_75, promedio_100, promedio_GER) (Figura 4).

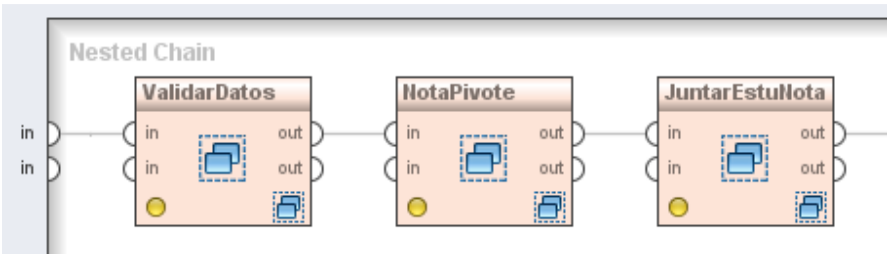

Figura 4. *RapidMiner - Bloque para la preparación de datos*

Adicionalmente, integra los conjuntos de datos de estudiantes y de notas, utilizando como llave el código del estudiante y lo almacena en repositorio local.

AgregarVarAux: calcula los índices IMA (índice de materias aprobadas), ICA (índice de créditos aprobados) y PCC (promedio de créditos cursados por semestre) (Figura 5).

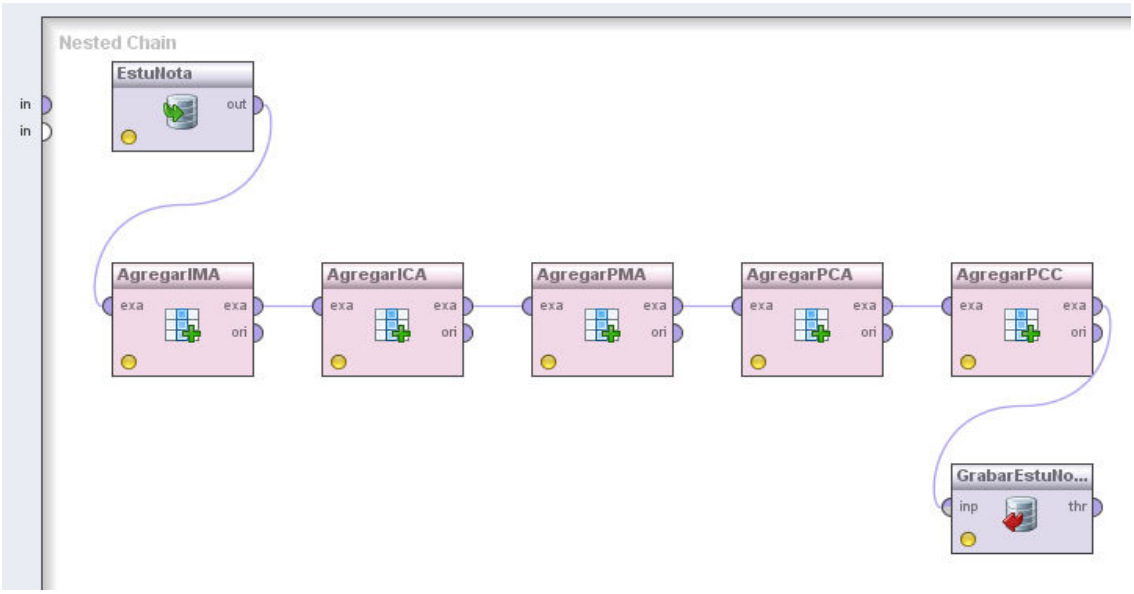

Figura 5.*RapidMiner - Bloque para el cálculo de índices*

Categorizar: Se categorizan las siguientes variables en nuevos atributos, para su utilización en el modelo (Figura 6) de acuerdo a la Tabla 3.

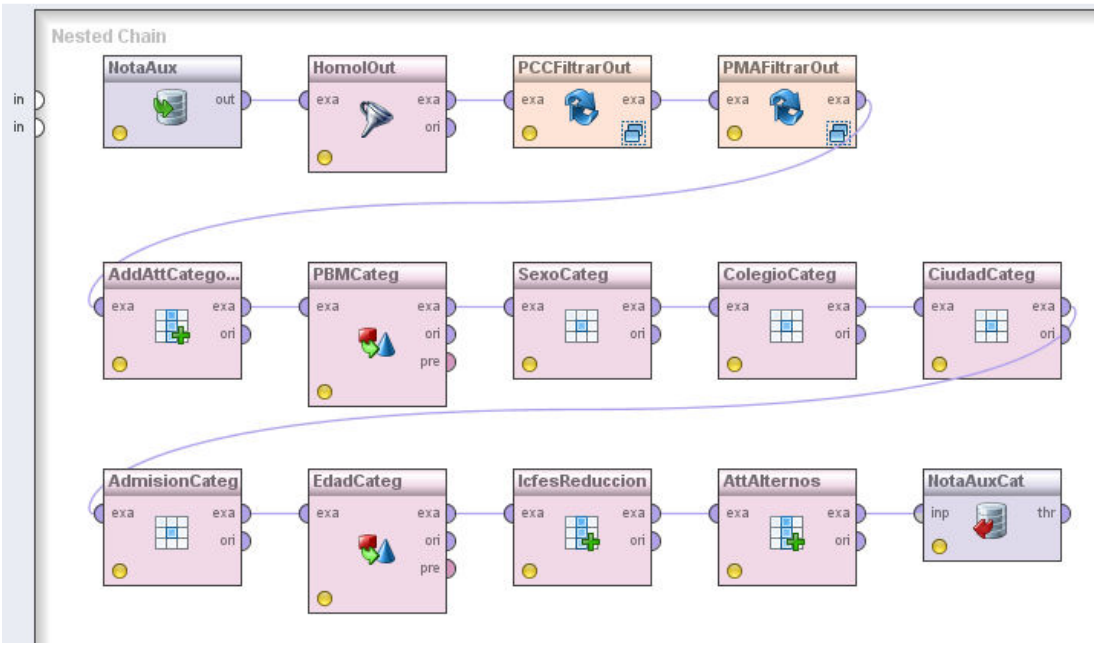

Figura 6 *Bloque para la categorización de atributos*

Tabla 3. *Categorías para variables explicativas*

Atributo	Atributo Categorizado	Categorías
PBM	PBMC PBM inferiores a 18 no pagan derechos de matrícula.	0: menor a 18. 1: a partir de 18.
SEXO	SEXOC	0: Mujer 1: Hombre.
COLTIPO	COLTIPOC	0: Oficial 1: Privado
CIUDAD_NIVEL	CIUDAD_NIVELC	0: NoCapital 1: Capital

ADMISION	ADMISIONC	0: Segunda opción 1: Primera opción
EDAD	EDADC	Antes16: Menor de 16 años E16_18: Entre 16 y 18 años E19_23: Entre 19 y 23 años E24_30: Entre 24 y 30 años Mas31: Mayor de 31 años

Adicionalmente, se agrupan las pruebas del ICFES de acuerdo a su área de conocimiento y facultad asociada en la Universidad como se muestra en la Tabla 4.

Tabla 4. *Agrupamiento de pruebas ICFES*

Atributo	Promedio de las pruebas
MATEMATICAS	MATEMATICAS
LENGUAJE	LENGUAJE
IC_EXACTAS	BIOLOGIA,FISICA,QUIMICA
IC_HUMANAS	FILOSOFIA,HISTORIA,GEOGRAFIA

El conjunto de datos después del procesamiento compuesto por 10.900 registros se resume en las Tabla **5** y Tabla **6**.

Tabla 5 *Resumen del conjunto de datos antes de calcular indicador*

Name	**Type**	**Statistics**	**Range**	**Missings**

ADMISIONC	nominal		1 (9734), 0 (1166)	0
CIUDAD_NIVELC	nominal		0 (3586), 1 (7314)	0
COLTIPOC	nominal		0 (8309), 1 (2591)	0
SEXOC	nominal		1 (5484), 0 (5416)	0
PBMC	nominal		0 (8060), 1 (2840)	0
CIUDAD_NIVEL	nominal		NoCapital (3586), Capital (7314)	0
COD_CARRERA	nominal			0
COD_PENSUM	nominal			0
SEXO	nominal		M (5484), F (5416)	0
ESTRATO	integer	avg = 2.495 +/- 0.917	[1.000 ; 6.000]	0
SEM_INGRES	integer	avg = 20063.860 +/- 24.981	[20011.000 ; 20111.000]	0
PROGRAMA	nominal			0
FACULTAD	nominal			0
PBM	integer	avg = 15.048 +/- 13.506	[1.000 ; 85.000]	0
COLTIPO	nominal		Oficial (8309), Privado (2591)	0
CIUDAD	nominal			0
DEPARTAMENTO	nominal			0
BIOLOGIA	real	avg = 50.443 +/- 6.960	[19.340 ; 87.490]	0
FILOSOFIA	real	avg = 48.089 +/- 7.126	[0.000 ; 86.960]	0
HISTORIA	real	avg = 49.180 +/- 7.020	[16.970 ; 85.340]	0
LENGUAJE	real	avg = 53.398 +/- 7.346	[18.210 ; 94.530]	0
MATEMATICAS	real	avg = 48.401 +/- 9.915	[14.920 ; 117.160]	0
FISICA	real	avg = 47.759 +/- 7.579	[0.000 ; 102.040]	0
QUIMICA	real	avg = 47.942 +/- 6.834	[17.700 ; 94.000]	0
GEOGRAFIA	real	avg = 49.827 +/- 7.388	[16.970 ; 85.340]	0
INTERDISCIPLINAR	real	avg = 53.847 +/- 7.695	[0.000 ; 97.000]	2
PONDERADO	real	avg = 136.574 +/- 71.269	[0.000 ; 429.310]	0
ADMISION	nominal	mode = P (9734), least = S (1166)	P (9734), S (1166)	0
FECHA_GRADO	nominal			9545
EDAD	integer	avg = 18.575 +/- 2.811	[11.000 ; 69.000]	0

ESTADO	nominal		GRADUADO (1355), RETIRADO (4609), ESTUDIANTE (4936)	0
NIVEL_ULT	integer	avg = 48.202 +/- 36.049	[0.000 ; 100.000]	0
GRADO_MESES	integer	avg = 73.646 +/- 11.841	[47.000 ; 121.000]	9545
HOMOLOGACIONES	integer	avg = 0.222 +/- 0.628	[0.000 ; 2.000]	0
creditos_0	integer	avg = 17.713 +/- 2.633	[1.000 ; 21.000]	0
creditos_100	integer	avg = 48.399 +/- 18.041	[6.000 ; 121.000]	8335
creditos_1000	integer	avg = 92.738 +/- 68.842	[2.000 ; 326.000]	0
creditos_25	integer	avg = 40.630 +/- 16.380	[2.000 ; 130.000]	0
creditos_50	integer	avg = 47.582 +/- 16.261	[1.000 ; 149.000]	5222
creditos_75	integer	avg = 45.047 +/- 14.576	[6.000 ; 113.000]	7045
materias_0	integer	avg = 6.698 +/- 1.319	[1.000 ; 11.000]	0
materias_100	integer	avg = 13.122 +/- 6.711	[1.000 ; 42.000]	8335
materias_1000	integer	avg = 32.790 +/- 23.627	[1.000 ; 124.000]	0
materias_25	integer	avg = 15.253 +/- 6.337	[1.000 ; 59.000]	0
materias_50	integer	avg = 17.382 +/- 6.538	[1.000 ; 50.000]	5222
materias_75	integer	avg = 15.251 +/- 6.350	[2.000 ; 51.000]	7045
creditos_perd_0	integer	avg = 4.297 +/- 4.597	[0.000 ; 20.000]	0
creditos_perd_100	integer	avg = 2.637 +/- 5.574	[0.000 ; 54.000]	8335
creditos_perd_1000	integer	avg = 16.078 +/- 14.750	[0.000 ; 103.000]	0
creditos_perd_25	integer	avg = 10.807 +/- 10.593	[0.000 ; 71.000]	0
creditos_perd_50	integer	avg = 6.340 +/- 8.661	[0.000 ; 101.000]	5222
creditos_perd_75	integer	avg = 3.812 +/- 6.559	[0.000 ; 54.000]	7045
materias_perd_0	integer	avg = 1.412 +/- 1.590	[0.000 ; 9.000]	0
materias_perd_100	integer	avg = 0.600 +/- 1.226	[0.000 ; 24.000]	8335
materias_perd_1000	integer	avg = 5.358 +/- 4.898	[0.000 ; 34.000]	0
materias_perd_25	integer	avg = 3.634 +/- 3.602	[0.000 ; 29.000]	0

materias_perd_50	integer	avg = 2.176 +/- 2.969	[0.000 ; 28.000]	5222
materias_perd_75	integer	avg = 1.271 +/- 2.182	[0.000 ; 29.000]	7045
promedio_0	numeric	avg = 3.217 +/- 0.765	[0.000 ; 4.800]	0
promedio_100	numeric	avg = 3.821 +/- 0.408	[0.900 ; 5.000]	8336
promedio_1000	numeric	avg = 3.104 +/- 0.812	[0.000 ; 4.800]	0
promedio_25	numeric	avg = 3.092 +/- 0.803	[0.000 ; 4.700]	0
promedio_50	numeric	avg = 3.508 +/- 0.459	[0.800 ; 4.800]	5222
promedio_75	numeric	avg = 3.659 +/- 0.417	[1.000 ; 4.900]	7045
periodos_0	integer	avg = 1 +/- 0	[1.000 ; 1.000]	0
periodos_100	integer	avg = 3.159 +/- 1.246	[1.000 ; 9.000]	8335
periodos_1000	integer	avg = 5.666 +/- 4.199	[1.000 ; 20.000]	0
periodos_25	integer	avg = 2.490 +/- 1.168	[1.000 ; 11.000]	0
periodos_50	integer	avg = 2.873 +/- 1.088	[1.000 ; 9.000]	5222
periodos_75	integer	avg = 2.645 +/- 0.963	[1.000 ; 9.000]	7045
promedio_gana_0	numeric	avg = 3.690 +/- 0.304	[3.000 ; 5.000]	243
promedio_gana_100	numeric	avg = 3.922 +/- 0.292	[3.000 ; 5.000]	8336
promedio_gana_1000	numeric	avg = 3.683 +/- 0.262	[3.000 ; 5.000]	235
promedio_gana_25	numeric	avg = 3.668 +/- 0.264	[3.000 ; 5.000]	235
promedio_gana_50	numeric	avg = 3.713 +/- 0.255	[3.000 ; 4.800]	5222
promedio_gana_75	numeric	avg = 3.789 +/- 0.266	[3.000 ; 5.000]	7045
creditos_gana_0	integer	avg = 13.416 +/- 5.022	[1.000 ; 21.000]	0
creditos_gana_100	integer	avg = 45.762 +/- 17.221	[3.000 ; 119.000]	8335
creditos_gana_1000	integer	avg = 76.659 +/- 66.322	[1.000 ; 281.000]	0
creditos_gana_25	integer	avg = 29.823 +/- 13.971	[1.000 ; 65.000]	0
creditos_gana_50	integer	avg = 41.241 +/- 13.703	[1.000 ; 79.000]	5222
creditos_gana_75	integer	avg = 41.236 +/- 13.023	[3.000 ; 90.000]	7045
materias_gana_0	integer	avg = 5.285 +/- 1.956	[1.000 ; 9.000]	0

materias_gana_100	integer	avg = 12.522 +/- 6.502	[1.000 ; 37.000]	8335
materias_gana_1000	integer	avg = 27.432 +/- 22.902	[1.000 ; 111.000]	0
materias_gana_25	integer	avg = 11.619 +/- 5.663	[1.000 ; 33.000]	0
materias_gana_50	integer	avg = 15.207 +/- 5.729	[1.000 ; 36.000]	5222
materias_gana_75	integer	avg = 13.980 +/- 5.799	[1.000 ; 37.000]	7045
IMA0	real	avg = 0.786 +/- 0.234	[0.100 ; 1.000]	0
IMA25	real	avg = 0.747 +/- 0.238	[0.062 ; 1.000]	0
IMA50	real	avg = 0.882 +/- 0.144	[0.154 ; 1.000]	5222
IMA75	real	avg = 0.924 +/- 0.116	[0.143 ; 1.000]	7045
IMA100	real	avg = 0.954 +/- 0.088	[0.222 ; 1.000]	8335
IMAGER	real	avg = 0.749 +/- 0.237	[0.062 ; 1.000]	0
ICA0	real	avg = 0.757 +/- 0.257	[0.048 ; 1.000]	0
ICA25	real	avg = 0.724 +/- 0.252	[0.034 ; 1.000]	0
ICA50	real	avg = 0.877 +/- 0.151	[0.086 ; 1.000]	5222
ICA75	real	avg = 0.922 +/- 0.121	[0.118 ; 1.000]	7045
ICA100	real	avg = 0.949 +/- 0.101	[0.167 ; 1.000]	8335
ICAGER	real	avg = 0.729 +/- 0.252	[0.034 ; 1.000]	0
PMA0	real	avg = 5.285 +/- 1.956	[1.000 ; 9.000]	0
PMA25	real	avg = 4.817 +/- 1.999	[0.333 ; 9.667]	0
PMA50	real	avg = 5.499 +/- 1.660	[0.500 ; 9.667]	5222
PMA75	real	avg = 5.436 +/- 1.780	[0.667 ; 9.667]	7045
PMA100	real	avg = 4.197 +/- 1.944	[0.667 ; 9.667]	8335
PMAGER	real	avg = 4.564 +/- 1.834	[0.333 ; 9.500]	0
PCA0	real	avg = 13.416 +/- 5.022	[1.000 ; 21.000]	0
PCA25	real	avg = 12.315 +/- 4.965	[0.500 ; 21.500]	0
PCA50	real	avg = 14.903 +/- 3.729	[1.000 ; 21.667]	5222
PCA75	real	avg = 16.108 +/- 3.469	[1.750 ; 21.667]	7045

PCA100	real	avg = 14.912 +/- 3.237	[3.000 ; 21.750]	8335
PCAGER	real	avg = 12.149 +/- 4.716	[0.500 ; 21.000]	0
PCC0	real	avg = 17.713 +/- 2.633	[1.000 ; 21.000]	0
PCC25	real	avg = 16.849 +/- 2.562	[2.000 ; 21.667]	0
PCC50	real	avg = 16.864 +/- 2.550	[1.000 ; 21.750]	5222
PCC75	real	avg = 17.387 +/- 2.677	[3.750 ; 21.667]	7045
PCC100	real	avg = 15.695 +/- 2.891	[4.000 ; 21.750]	8335
PCCGER	real	avg = 16.581 +/- 2.351	[2.000 ; 21.500]	0
IC_EXACTAS	real	avg = 48.714 +/- 5.692	[25.483 ; 85.703]	0
IC_HUMANAS	real	avg = 49.032 +/- 5.684	[25.150 ; 76.313]	0

Tabla 6 *Distribución de estudiantes por programa y avance en créditos*

Cod	Programa	N0	N25	N50	N75	N100
001	ARTES PLASTICAS	234	234	117	65	41
007	LICENCIATURA EN MUSICA	306	306	160	96	56
010	DISEÑO VISUAL	277	277	207	183	124
019	LICENCIATURA EN EDUCACION FISICA Y RECREACION	34	34	33	30	26
020	LICENCIATURA EN BIOLOGIA Y QUIMICA	585	585	225	129	76
021	LICENCIATURA EN CIENCIAS SOCIALES	483	483	213	133	70
022	LICENCIATURA EN LENGUAS MODERNAS	423	423	288	206	146
024	ENFERMERIA	660	660	458	343	276
025	LICENCIATURA EN FILOSOFIA Y LETRAS	400	400	145	85	60
026	TRABAJO SOCIAL	585	585	399	285	193
027	DESARROLLO FAMILIAR	471	471	231	168	128
050	INGENIERÍA AGRONÓMICA	677	677	354	207	143
051	DERECHO	314	314	233	174	127
052	MEDICINA	493	493	215	90	51
053	MEDICINA VETERINARIA Y ZOOTECNIA	631	631	378	307	183
060	GEOLOGIA	657	657	247	154	90
080	INGENIERIA DE ALIMENTOS	656	656	394	312	236
170	INGENIERÍA EN SISTEMAS Y COMPUTACIÓN	506	506	248	165	114
171	BIOLOGIA	500	500	227	154	88
172	LICENCIATURA EN ARTES ESCENICAS CON ENFASIS EN TEATRO	340	340	145	99	69
205	ANTROPOLOGIA	401	401	137	79	34
206	SOCIOLOGIA	466	466	125	65	31
207	LICENCIATURA EN EDUCACIÓN BÁSICA CON ÉNFASIS EN EDUCACIÓN FÍSICA, RECREACIÓN Y DEPORTES	735	735	436	274	162

213	PROFESIONAL EN FILOSOFÍA Y LETRAS	66	66	63	52	41
	Total	10900	10900	5678	3855	2565

IndicadorRendimiento: en este bloque se calcula el Indicador Académico de Rendimiento por segmento de avance, adicionalmente se seleccionan sólo los estudiantes que cuyo índice PCC (promedio de créditos inscritos por semestre) sea máximo 21 y su índice PMA (promedio de materias aprobadas por semestre) sea máximo 9, pues son los límites establecidos por la Universidad y excluyen los casos atípicos como reingresos o transferencias de otras universidades o programas académicos y que por el mecanismo de registro de las evaluaciones son registradas todas las notas homologadas en un solo semestre (Figura 7).

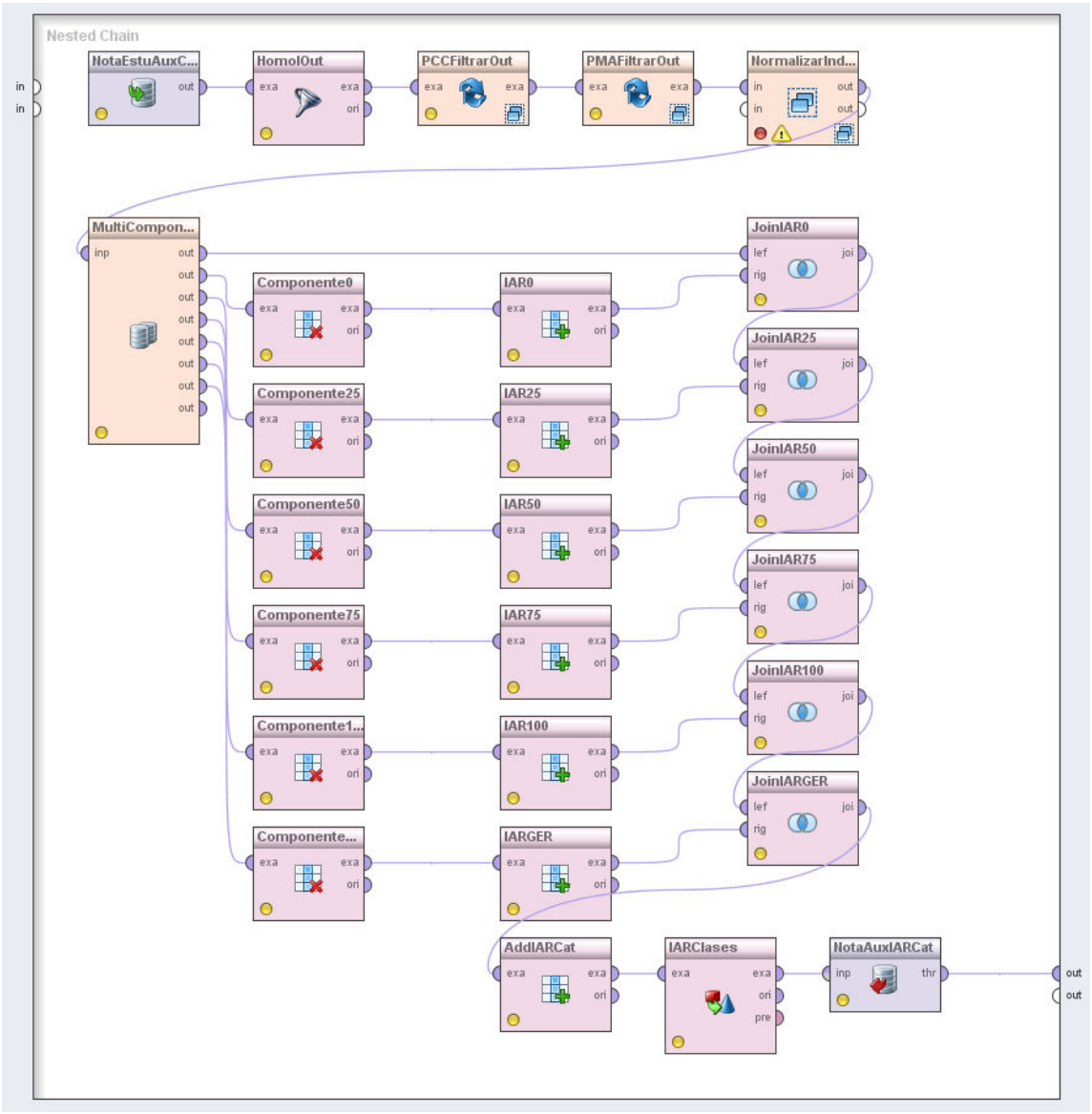

Figura 7. *RapidMiner – Bloque para el cálculo del Índice Académico*

En este mismo bloque se categoriza el Indicador Académico de Rendimiento IAR en 3 categorías, estas se fijan basándose en el reglamento estudiantil y el

juicio de varios profesores consultados. Los valores fijados para categorizar se encuentran en la Tabla 7.

Tabla 7. *Categorización del Indicador Académico de Rendimiento IAR*

Categoría	IAR máximo	Rangos en los índices componentes
Bajo	0.525714285	Promedio [0-3.0) Promedio materias ganadas [3.0- 3.5) IMA [0-0.5) ICA [0-0.5) PCC [0-0.5)
Medio	0.86342857	Promedio [3.0-3.7) Promedio materias ganadas [3.5-3.8) IMA [0.5-1) ICA [0.5-1) PCC [0.5-1)
Alto	1	Promedio [3.7-5.0] Promedio materias ganadas [3.8-5.0] IMA [1] ICA [1] PCC [1]

En la Figura 8 se muestra la distribución del indicador académico de rendimiento IAR para el segmento del 25% en avance en créditos por programa académico, para el mismo segmento en la Figura 9 se muestra la distribución por sexo y en la Figura 10 se muestra la distribución por naturaleza del colegio; este tipo de gráficos son generados directamente por RapidMiner a partir de los conjuntos de datos generados.

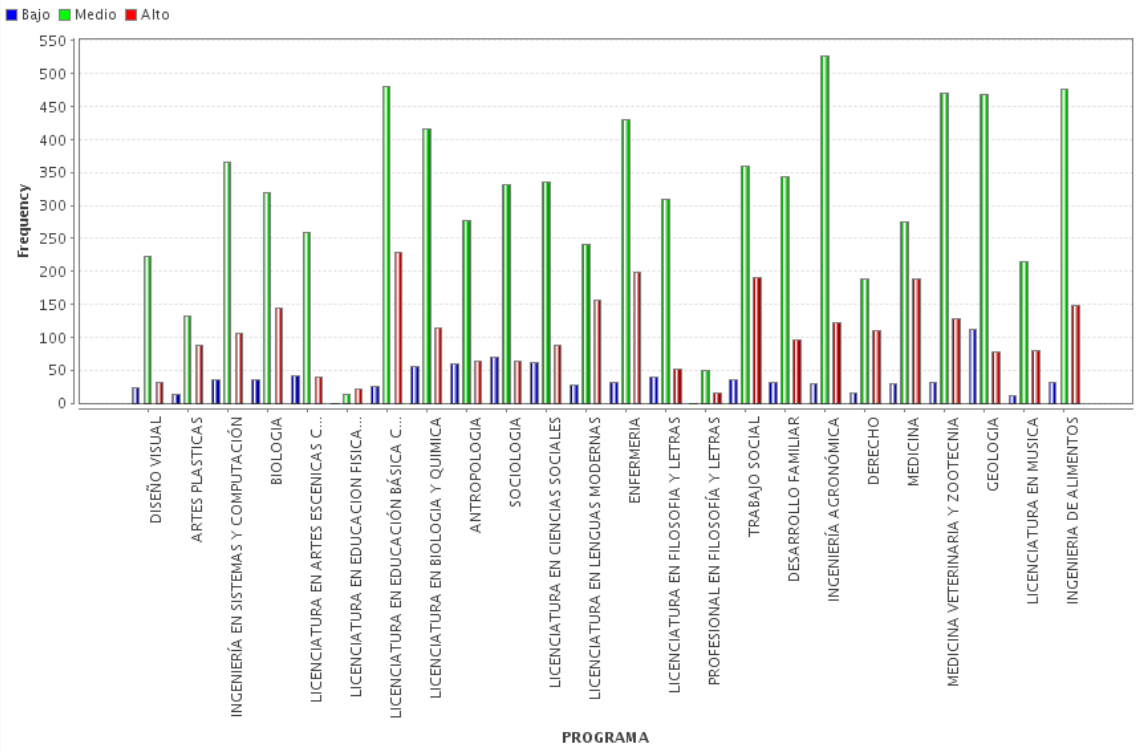

Figura 8 *Distribución por categorías IAR25 según programa*

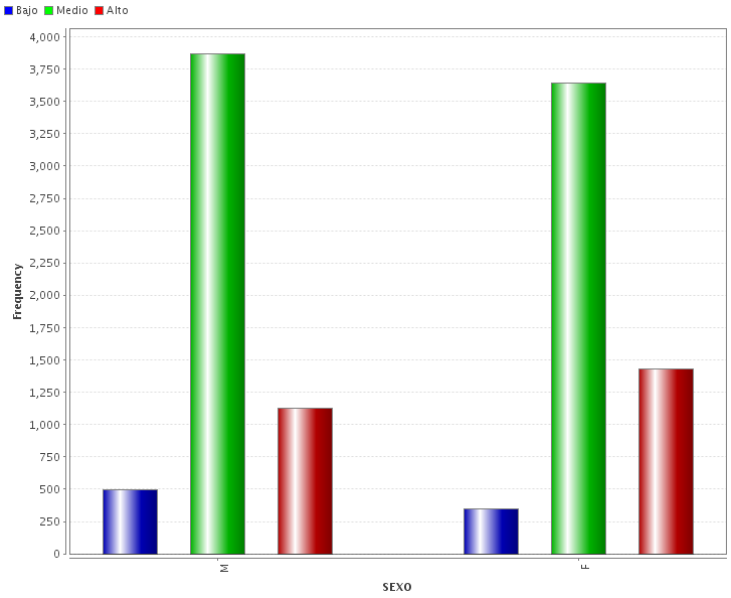

Figura 9 *Distribución por categorías IAR25 según sexo*

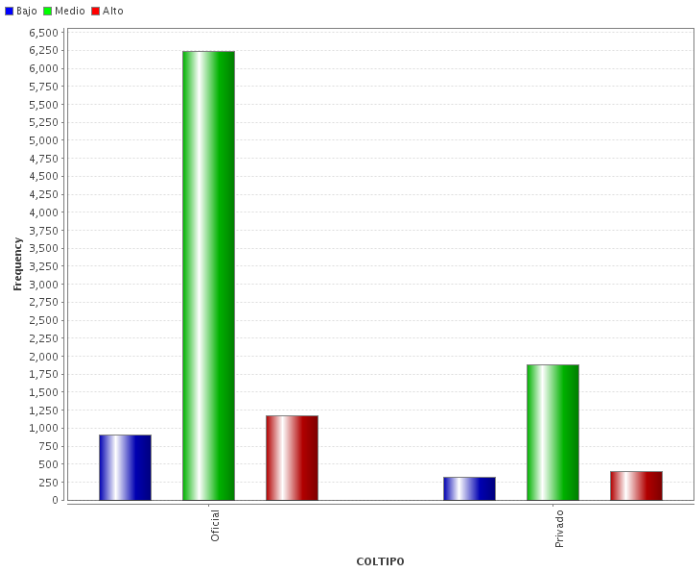

Figura 10 *Distribución por categorías IAR25 según naturaleza del colegio*

RegresionLogistica: bloque en que se modela la regresión logística por programa académico. Se amplía en la sección Modelado.

En el flujo de correlaciones y agrupaciones para el informe (Figura 3), en el bloque Agrupaciones (Figura 11) se realizan las agrupaciones para generar el informe descriptivo de los datos.

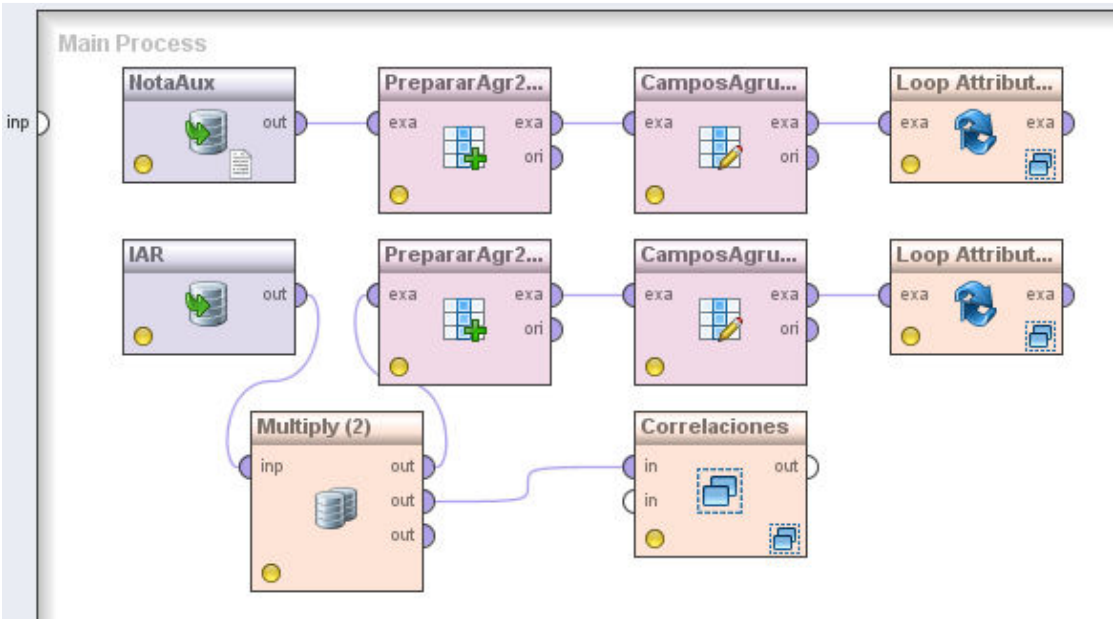

Figura 11. *RapidMiner - Bloque para el cálculo de agrupaciones y correlaciones*

Los atributos para las agrupaciones son las siguientes:

1. Programa
2. Estrato
3. PBMC:
4. Facultad
5. ColTipo
6. CiudadNivel

7. Admision
8. EdadC
9. Programa-Sexo
10. Programa-Estrato
11. Facultad-Sexo
12. Programa-PBMC

Se generan los siguientes resúmenes por cada agrupación, en la Figura 12 se muestra la configuración en RapidMiner:

1. Para cada uno de los índices por segmento de avance:
a. Conteo
b. Mínimo
c. Máximo
d. Promedio
e. Mediana
f. Desviación estándar

2. Para cada indicador académico de rendimiento por segmento de avance
a. Conteo
b. Mínimo
c. Máximo
d. Promedio
e. Mediana
f. Desviación estándar

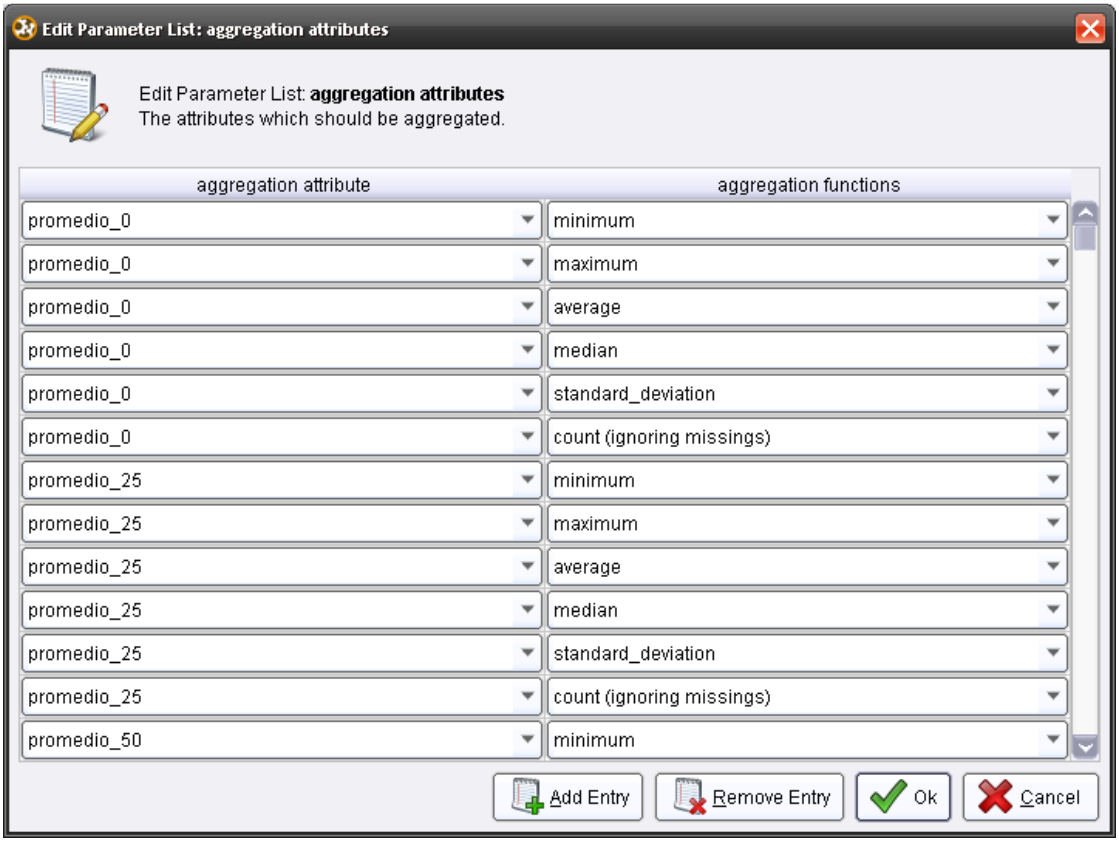

Figura 12. *RapidMiner - Ejemplo de cálculo de agrupamientos*

Modelado

Seleccionar técnicas de modelado

Se seleccionó la Regresión Logística Multinomial como técnica de modelado por las siguientes características:

1. La variable dependiente es categórica multinomial. Aunque el indicador académico de rendimiento IAR es un valor numérico, se categorizó en las categorías Bajo, Medio y Alto para facilitar su interpretación y su uso en la toma de decisiones.
2. Una de sus principales características es identificar los factores críticos que afectan la variable dependiente, puesto que los parámetros de un modelo logístico se interpretan como tasas de disparidad.
3. Las variables explicativas son de tipo numérico y categórico.

La regresión logística se aplicará a los conjuntos de estudiantes por cada programa académico y por el avance en créditos en este, con el fin de comparar estudiantes en situaciones similares y facilitar la utilización de los resultados en la definición de estrategias para la mejora del rendimiento académico.

Generar el diseño de prueba

Las estimaciones de los oddsratio de todas las regresiones logísticas son analizadas inferencialmente mediante la prueba estadística de Wald, el cual juega el mismo rol que la prueba t-student en el análisis de regresión lineal múltiple. Se asumirá que un factor es estadísticamente significativo cuando el p-valor de la prueba sea menor que 0.1 (es decir, el error tipo I asumido es del 10%).

La significancia estadística del modelo logístico completo se hará con la prueba de razón de verosimilitud asociada a la distribución chi-cuadrada. Se asumirá que el modelo es estadísticamente significativo si el p-valor de la prueba es menor que 0.1 (es decir, el error tipo I asumido es del 10%).

De la misma forma, la significancia estadística de los coeficientes de correlación se determinó con la prueba t-student. Se asumirá que la correlación es estadísticamente significativa cuando el p-valor de la prueba sea menor que 0.1 (es decir, el error tipo I asumido es del 10%).

Construcción del modelo

Para la aplicación del modelo de Regresión Logística Multinomial, se utilizó la integración de RapidMiner con el proyecto R y se aplicó por programa académico y segmento de avance en los créditos de la carrera.

Los parámetros de la regresión logística son los siguientes:

Variable dependiente: IAR (índice académico de rendimiento) con categoría de referencia “Medio” por lo que el modelo nos dará información acerca de los factores de riesgo o de protección que hacen que la proporción estudiantes de nivel alto o estudiantes de nivel bajo sobre estudiantes de nivel medio aumente o disminuya.

Variables explicativas:

1. Binomiales:

a. CIUDAD_NIVELC: Nivel de la ciudad. 0 municipio, 1 Capital
b. COLTIPOC: Naturaleza del colegio. 0 Oficial 1 Privado
c. PBMC: Puntaje básico de matrícula. 0 Menor a 18, 1 A partir de 18.
d. SEXOC: Sexo: 0 Mujer, 1 Hombre.

2. Numéricas:

a. EDAD: Edad de ingreso del estudiante a la universidad.
b. MATEMATICAS: Puntaje en la prueba ICFES de matemáticas.
c. LENGUAJE: Puntaje en la prueba ICFES de lenguaje.
d. IC_EXACTAS: Puntaje en la agrupación de las pruebas ICFES IC_EXACTAS.
e. IC_HUMANAS: Puntaje en la agrupación de las pruebas ICFES IC_ HUMANAS.

El proceso implementado en RapidMiner para el cálculo de la regresión logística (Figura 13) se presenta a continuación con la explicación paso a paso:

1. Primer nivel: Cargar información

a. Se toma el conjunto de datos general que contiene todos los atributos, incluyendo las variables explicativas y los IAR.

b. Se tiene un filtro que puede ser empleado para seleccionar un solo programa o todos los programas,

2. Por último se tiene un ciclo que itera por cada programa académico.

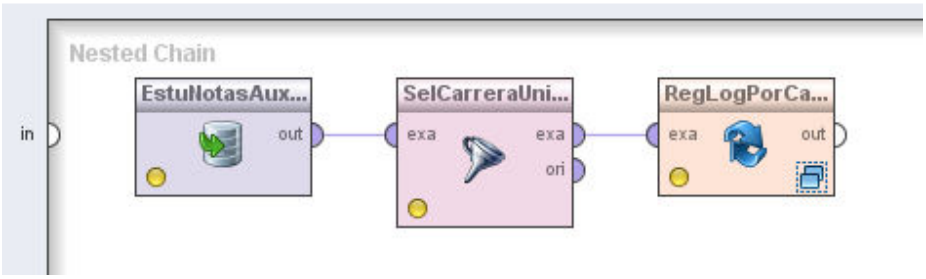

Figura 13. *RapidMiner - Bloque para el análisis de la regresión logística. Nivel 1*

2. Segundo nivel: Ciclo por programa (Figura 14).

a. Se filtran los registros por el programa que defina la iteración (Figura 10).

b. Se seleccionan sólo las variables explicativas del modelo y todos los IAR (0, 25, 50, 75, 100 y General).

c. Se hace un ciclo por cada IAR.

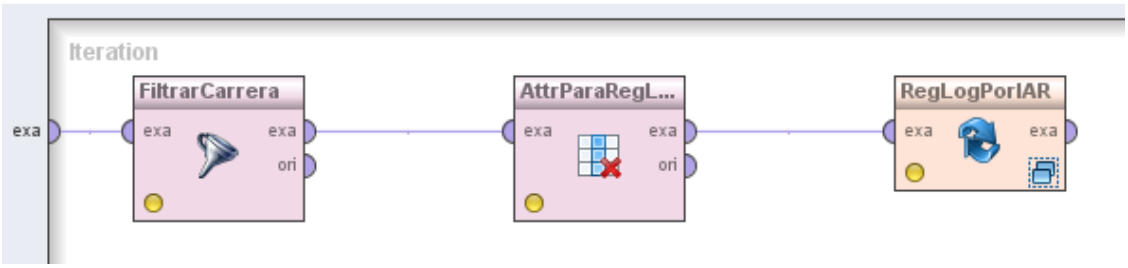

Figura 14. *RapidMiner - Bloque para el análisis de la regresión logística. Nivel 2 Ciclo por programa*

3. Tercer nivel: Ciclo por IAR de acuerdo al porcentaje de créditos de avance en el programa (Figura 15).

a. Se crea una copia del IAR de la iteración y se renombra por IAR(iteración)CN.
b. Se eliminan todos los IAR excepto la copia realizada.
c. Se asigna el rol “Label” a la copia del IAR, ya que la variable dependiente para el modelo de Regresión Logística debe tener este rol.
d. Se eliminan los registros cuyo atributo IAR (rol “Label“) esté vacío.
e. Se eliminan los registros que contengan variables explicativas vacías.
f. Se almacenan en un repositorio local los datos a los cuales se les aplicará la Regresión Logística.
g. Aplicar la Regresión Logística al conjunto de datos. Se maneja dentro de un bloque de excepción para continuar el análisis así se presente un error.

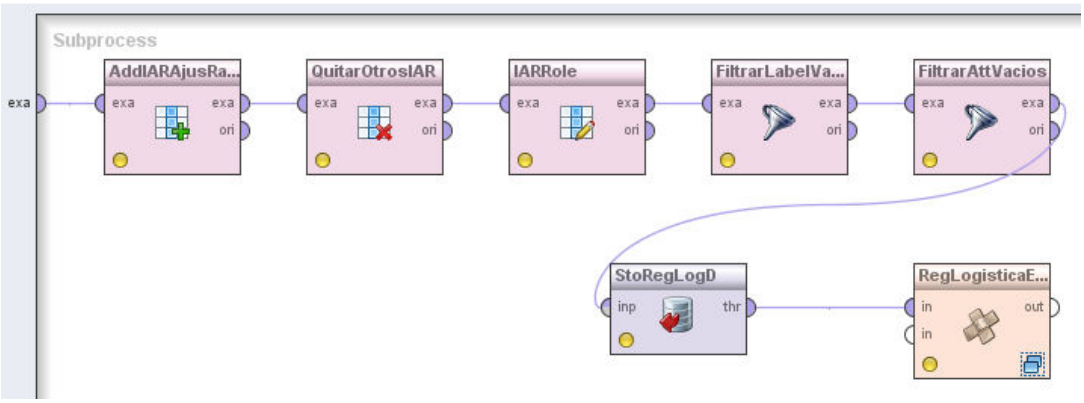

Figura 15. *RapidMiner - Bloque para el análisis de la regresión logística. Nivel 3 Ciclo por IAR*

4. Cuarto Nivel: Parametrización en R (Figura 16)

Se utilizó el paquete mlogit, definido en RapidMiner. El script para la parametrización de la regresión logística en R se lista en la Tabla 8.

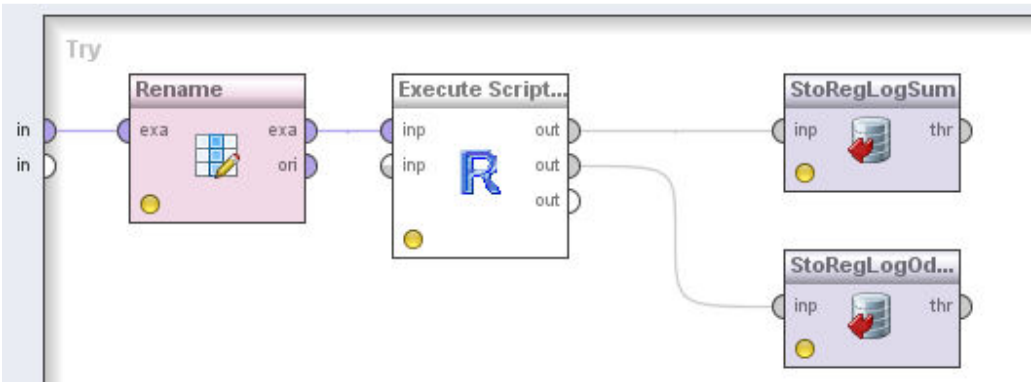

Figura 16. *RapidMiner - Bloque para el análisis de la regresión logística. Nivel 4 Parametrización R*

Tabla 8. *Script R para el cálculo de la Regresión Logística*

```
library(mlogit)
RL <- mlogit.data(IAR, choice="IARCN", shape="wide")
RL.train <- mlogit(formula = IARCN ~ 0 | CIUDAD_NIVELC + COLTIPOC +
SEXOC + PBMC + LENGUAJE + MATEMATICAS + EDAD + IC_EXACTAS +
IC_HUMANAS, data = RL, method = "nr", reflevel="Medio", estimate=TRUE)
summary <- summary(RL.train)
oddsratios <- exp(RL.train$coefficients)
```

Resultados

Los resultados de la Regresión Logística aplicados por programa y segmento de avance en los créditos son retornados por R en el formato mostrado en la Tabla 9.

Tabla 9. *Ejemplo de resultado de la Regresión Logística en R - summary*

```
mlogit(formula = IARCN ~ 0 | CIUDAD_NIVELC + COLTIPOC + SEXOC +
    PBMC + LENGUAJE + MATEMATICAS + EDAD + IC_EXACTAS +
IC_HUMANAS,
    data = RL, reflevel = "Medio", estimate = TRUE, method = "nr",
    print.level = 0)

Frequencies of alternatives:
   Medio     Alto     Bajo
0.776957 0.180207 0.042836

nr method
6 iterations, 0h:0m:0s
g'(-H)^-1g = 0.000298
successive fonction values within tolerance limits

Coefficients :
                      Estimate Std. Error t-value  Pr(>|t|)
altAlto               -9.4212704  1.7665602 -5.3331 9.654e-08 ***
altBajo               -8.1273401  3.0640543 -2.6525  0.007990 **
altAlto:CIUDAD_NIVELC1 -0.4695436  0.2201909 -2.1324  0.032971 *
altBajo:CIUDAD_NIVELC1 -0.5851631  0.4061380 -1.4408  0.149642
altAlto:COLTIPOC1      -0.0074555  0.2700657 -0.0276  0.977976
altBajo:COLTIPOC1      -0.2484522  0.5338065 -0.4654  0.641620
altAlto:SEXOC0          0.3972246  0.2232803  1.7790  0.075233 .
altBajo:SEXOC0         -0.7018911  0.5142363 -1.3649  0.172278
altAlto:PBMC1          -0.0613030  0.2493075 -0.2459  0.805765
altBajo:PBMC1          -0.0397583  0.4813998 -0.0826  0.934178
altAlto:LENGUAJE        0.0038141  0.0182871  0.2086  0.834785
altBajo:LENGUAJE       -0.0331205  0.0343082 -0.9654  0.334353
altAlto:MATEMATICAS     0.0197355  0.0152862  1.2911  0.196682
altBajo:MATEMATICAS     0.0470925  0.0272090  1.7308  0.083494 .
altAlto:EDAD            0.1074305  0.0367175  2.9259  0.003435 **
altBajo:EDAD            0.0244569  0.0790489  0.3094  0.757025
altAlto:IC_EXACTAS      0.0680066  0.0297928  2.2826  0.022451 *
altBajo:IC_EXACTAS      0.0634783  0.0534194  1.1883  0.234715
altAlto:IC_HUMANAS      0.0324642  0.0236750  1.3712  0.170298
altBajo:IC_HUMANAS      0.0314092  0.0434937  0.7222  0.470199
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -413.1
```

```
McFadden R^2:  0.046348
Likelihood ratio test : chisq = 40.153 (p.value=0.0019894)
```

Los resultados de la Regresión Logística por Programa y Porcentaje de avance en créditos se incluyen en el Anexo 3 Resultados de la Regresión Logística por Programa y Porcentaje de avance en créditos.

El resumen (Tabla 7) contiene los coeficientes -estimate- de la Regresión Logística para las categorías de IAR “Bajo” y “Alto” en relación con la categoría base “Medio”, así como el nivel de significancia -Pr(>|t|)- por cada covariable.

Los oddsratios son presentados por R, debajo de cada par categoría-covariable (Tabla 10).

Tabla 10. *Ejemplo de resultado de la Regresión Logística en R – oddsratios*

```
            altAlto             altBajo altAlto:CIUDAD_NIVELC1
       8.098307e-05        2.953528e-04        6.252876e-01
altBajo:CIUDAD_NIVELC1     altAlto:COLTIPOC1     altBajo:COLTIPOC1
       5.570150e-01        9.925723e-01        7.800072e-01
       altAlto:SEXOC0       altBajo:SEXOC0        altAlto:PBMC1
       1.487690e+00        4.956471e-01        9.405382e-01
       altBajo:PBMC1      altAlto:LENGUAJE      altBajo:LENGUAJE
       9.610217e-01        1.003821e+00        9.674219e-01
  altAlto:MATEMATICAS   altBajo:MATEMATICAS         altAlto:EDAD
       1.019931e+00        1.048219e+00        1.113413e+00
       altBajo:EDAD    altAlto:IC_EXACTAS    altBajo:IC_EXACTAS
       1.024758e+00        1.070372e+00        1.065536e+00
   altAlto:IC_HUMANAS    altBajo:IC_HUMANAS
       1.032997e+00        1.031908e+00
```

Ejemplos de interpretación de los resultados

La relación Alto/Medio (122 alto / 526 medio = 23.2 altos por cada 100 medios) es 1.59 veces mayor en municipios que en capitales (37.1 altos por cada 100 medios) (Tabla 11)

Tabla 11. *Ejemplo de resultado.Relación Alto/Medio - Ciudad.*

```
altAlto:CIUDAD_NIVELC1 -0.4695436  0.2201909 -2.1324  0.032971 *

altAlto:CIUDAD_NIVELC1
6.252876e-01 (inverso= 1.59928)
```

La relación Alto/Medio (122 alto / 526 medio = 23.2 bajos por cada 100 medios) aumenta 1.11 veces por cada año que tenga de más el estudiante al momento de ingresar a la universidad (Tabla 12).

Tabla 12. *Ejemplo de resultado.Relación Alto/Medio - Edad.*

```
altAlto:EDAD          0.1074305  0.0367175  2.9259  0.003435 **

altAlto:EDAD
1.113413e+00
```

La relación Alto/Medio (122 alto / 526 medio = 23.2 altos por cada 100 medios) aumenta 1.07 veces por unidad que tenga de más en el ICFES en el promedio de las ciencias exactas (Tabla 13).

Tabla 13. *Ejemplo de resultado.Relación Alto/Medio - ICFES.*

altAlto:IC_EXACTAS	0.0680066	0.0297928	2.2826	0.022451 *
altAlto:IC_EXACTAS 1.070372e+00				

Evaluación del modelo

Se presenta la prueba de razón de verosimilitud asociada a la distribución chi-cuadrada, se resaltan los p-valores que son menores de 0.1, de acuerdo al diseño de la prueba.

Para cada posición se presenta el parámetro chi-cuadrada y el p-valor debajo, las posiciones en rosa presentaron errores al generar el modelo (Tabla 14).

De manera posterior se presenta un resumen de la prueba en la Tabla 15.

Tabla 14. *Prueba de razón de verosimilitud por Programa Académico*

Programa	0	25	50	75	100	General
Agronomía	40.153	37.898	36.989		2.1459	37.982

Programa	0	25	50	75	100	General
	0.0019894	0.003996	0.0052576		0.98894	0.0038948
Medicina	85.992 7.5102e-11	99.714 2.4993e-13	50.101 1.031e-07	29.295 0.0005777	14.927 0.092948	102.03 9.4102e-14
Artes Plásticas	25.141 0.12108	18.312 0.4353	28.632 0.053065	19.494 0.36199	9.0803 0.4299	22.799 0.19844
Diseño Visual	40.661 0.0016951	35.763 0.0075669	30.016 0.037294	21.69 0.24606	17.777 0.037851	56.926 6.3591e-06
Lenguas Modernas	84.491 1.3875e-10	63.132 6.3188e-07	46.265 0.00027138		21.843 0.23902	70.067 4.4027e-08
Enfermería	21.618 0.24937	30.299 0.03464	30.27 0.034905		22.456 0.0075404	40.315 0.0018906
Filosofía	50.638 6.0408e-05	30.871 0.029794	31.247 0.026949		38.825 0.003008	28.735 0.051712
Trabajo Social	9.937 0.0021289	38.32 0.0035125		24.327 0.003813	10.927 0.28074	30.661 0.031496
Derecho	62.959 6.7463e-07	40.318 0.0018885	34.156 0.012048	30.709 0.031096		54.253 1.6759e-05
Veterinaria	38.507 0.0033175	24.969 0.12578	47.401 0.00018473	22.705 0.0068932		32.198 0.020828
Ing.Alimentos	62.848 7.034e-07	65.905 2.1992e-07	30.933 0.00030392	18.087 0.034173	12.271 0.19846	60.543 1.6709e-06
Ing.Computación	34.807 0.0099965	31.008 0.028729	21.317 0.26373		11.644 0.23414	31.91 0.022532
Lic. Música	46.528	66.741			12.506	57.351

Programa	0	25	50	75	100	General
	0.00024833	1.5957e-07			0.18629	5.4432e-06
Lic. Biología y Química	78.984 1.2894e-09	74.427 7.9401e-09	24.945 0.12645	43.583 0.00066162	20.726 0.013925	90.936 9.8017e-12
Lic. Sociales	40.566 0.0017466	42.974 0.00080701		16.28 0.061259	16.312 0.060636	42.532 0.00093142)
Desarrollo Familiar	62.163 9.1061e-07	48.244 0.00013847	35.648 0.0078253	29.118 0.046955	17.47 0.041845	58.038 4.2292e-06
Geología	101.03 1.4311e-13	93.879 2.8854e-12	24.848 0.12916	15.317 0.082583	17.194 0.045767	90.509 1.1698e-11
Biología	60.699 1.5769e-06	70.929 3.1453e-08	27.448 0.070969	27.188 0.07553	23.267 0.0056246	59.674 2.3099e-06
Lic. Artes Escénicas	22.767 0.1997	45.036 0.00040952	24.466 0.14033		7.1573 0.62075	37.623 0.0043439
Antropología	96.07 1.1555e-12	75.917 4.3951e-09	22.687 0.20291		14.501 0.10557	69.6 5.2805e-08
Sociología	39.899 0.0021542	33.038 0.016516	28.423 0.05591	52.784 2.8332e-05	9.7281 0.37294	52.632 2.9902e-05
Lic. EduBásica EduFísica	74.212 8.6448e-09	76.091 4.1008e-09	36.567 0.0059652	25.598 0.10931		61.567 1.1391e-06

Tabla 15. *Resumen de la prueba de razón de verosimilitud*

Modelos	Cantidad	Porcentaje
Total de modelos generados (programa-avance en créditos)	132	100%
Modelos no generados por error en RapidMiner	14	10%
Modelos que no presentaron significancia estadística (p-valor inferior a 0.1)	24	18%

Evaluación

Evaluación de los resultados

Los factores que influyen de manera positiva o negativa para la inclusión de un estudiante en la categoría Alto o Bajo en relación con la categoría base Medio, se resumen por programa académico entre la Tabla 16 a la Tabla 37.

Tabla 16. *Resumen de OddsRatio para el programa de Ingeniería Agronómica.*

Nivel	0	25	50	75	100	General
Alto	NoCapital (1.59)** Hombre (1.48). Edad (1.11)** Exactas (1.07)*	Exactas (1.10)* Humanas (1.12)**	Privado(2.64)* NoPaga (3.86)* Exactas (1.10).	Error	No significativos	Matemáticas (1.083)**
Bajo	Matemáticas (1.04).	Mujer (2.48)*	Edad (1.14).			Mujer (2.54)*
Likelihood ratio test chisq = p.value=	40.153 0.0019894	37.898 0.003996	36.989 0.0052576	Error	2.1459 0.98894	37.982 0.0038948

Para formatear la Tabla 16 se utilizó el rosa para indicar que se presentó un error en RapidMiner, amarillo para indicar que el modelo no fue significativo estadísticamente y verde para mostrar que el modelo fue significativo; adicionalmente, para los oddsratio se utilizó la convención (**) para un nivel de significación de 0.01, (*) para un nivel de 0.05 y (.) para un nivel de 0.1

Interpretación de resultados.

Se presenta la interpretación de resultados completa para el programa de Ingeniería Agronómica a manera de ejemplo, tanto para variables categóricas como numéricas.

1. La relación estudiantes con IAR Alto sobre Medio en primer semestre:

a. Es 1.59 veces mayor en estudiantes provenientes de municipios que no son capital que en provenientes de capitales.

b. Es 1.48 veces mayor en hombres que en mujeres.

c. Aumenta 1.11 veces por cada año que tenga de más el estudiante al momento de ingresar a la universidad.

d. Aumenta 1.07 veces por punto que tenga de más el estudiante en el promedio de las áreas de ciencias exactas (biología, química y física) de las pruebas ICFES.

2. La relación estudiantes con IAR Bajo sobre Medio en primer semestre:

a. Aumenta 1.04 veces por punto que tenga de más el estudiante en el área de matemáticas de las pruebas ICFES.

3. La relación estudiantes con IAR Alto sobre Medio en el primer 25 por ciento del avance en créditos del programa académico:

a. Aumenta 1.10 veces por punto que tenga de más el estudiante en el promedio de las áreas de ciencias exactas (biología, química y física) de las pruebas ICFES.
b. Aumenta 1.12 veces por punto que tenga de más el estudiante en el promedio de las áreas de ciencias humanas (filosofía, historia y geografía) de las pruebas ICFES.

4. La relación estudiantes con IAR Bajo sobre Medio en el primer 25 por ciento del avance en créditos del programa académico:
   Es 2.48 veces mayor en mujeres que en hombres.

5. La relación estudiantes con IAR Alto sobre Medio entre el 25 por ciento y el 50 por ciento del avance en créditos del programa académico:

a. Es 2.64 veces mayor en estudiantes que provienen de colegios privados que de los que provienen de colegios públicos.
b. Es 3.86 veces mayor en estudiantes que no se les hace excepción de matrícula académica que a los estudiantes que se les hace.
c. Aumenta 1.10 veces por punto que tenga de más el estudiante en el promedio de las áreas de ciencias exactas de las pruebas ICFES.

6. La relación estudiantes con IAR Bajo sobre Medio entre el 25 por ciento y el 50 por ciento del avance en créditos del programa académico:

Aumenta 1.14 veces por cada año que tenga de más el estudiante al momento de ingresar a la universidad.

7. La relación estudiantes con IAR Alto o Bajo sobre Medio entre el 50 por ciento y el 75 por ciento del avance en créditos del programa académico se presentó un error en el sistema de minería de datos.

8. La relación estudiantes con IAR Alto o Bajo sobre Medio entre el 75 por ciento y el 100 por ciento del avance en créditos del programa académico no se presentaron factores con significancia estadística (p-valores inferiores a 0.1), adicionalmente el modelo no es estadísticamente significativo pues el p-valor en la prueba de máxima verosimilitud es 0.98.

9. La relación estudiantes con IAR Alto sobre Medio sin tener en cuenta el avance en créditos en que se encuentre el estudiante:
   Aumenta 1.083 veces por punto que tenga de más el estudiante en el área de matemáticas de las pruebas ICFES.

10. La relación estudiantes con IAR Bajo sobre Medio sin tener en cuenta el avance en créditos en que se encuentre el estudiante:
    Es 2.54 veces mayor en mujeres que en hombres.

Tabla 17. *Resumen de OddsRatio para el programa de Antropología*

Nivel	0	25	50	75	100	General
Alto	Exactas (1.173)** Humanas (1.065).	Hombre (2.884)* Exactas (1.139)* Humanas (1.097)*	Paga (4.981)* Exactas (1.225)*		Lenguaje (1.314). Matemáticas (0.801).	Mujer (4.207)* Exactas (1.193)*
Bajo	NoCapital (2.504)** Privado (2.557)** Edad (1.143)**	NoCapital (1.859)* Privado (2.323)* Matemáticas (0.947)* Edad (1.123)*		Error		NoCapital (1.846)* Privado (2.252)* Matemáticas (0.948)* Edad (1.122)*
Likelihood ratio test chisq = p.value=	96.07 1.1555e-12	75.917 4.3951e-09	22.687 0.20291	Error	14.501 0.10557	69.6 5.2805e-08

Tabla 18. *Resumen de OddsRatio para el programa de Artes Plásticas*

Nivel	0	25	50	75	100	General
Alto	Hombre (1.85)* Paga (2.31)*	Paga (2.15).	Edad (1.12). Humanas (1.10).	Matemáticas (1.10).	No significativos	Edad (1.096)*
Bajo	Paga (3.05).					
Likelihood ratio test chisq = p.value=	25.141 0.12108	18.312 0.4353	28.632 0.053065	19.494 0.36199	9.0803 0.4299	22.799 0.19844

Tabla 19. *Resumen de OddsRatio para el programa de Biología*

Nivel	0	25	50	75	100	General
Alto	Hombre (1.569)* Exactas (1.091)** Paga (1.539).	Privado (2.365)* Edad (0.753)* Exactas (1.171)***	Exactas (1.151)** Privado (2.222).	Edad (0.723).	Matemáticas (1.179)*	Matemáticas (1.062)* Exactas (1.203)***
Bajo	Privado (2.406)* Humanas (0.865)** Lenguaje (1.054). Edad (1.097).	Privado (2.129)* Humanas (0.912)*				Privado (1.984)* Humanas (0.903)*
Likelihood ratio test chisq = p.value=	60.699 1.5769e-06	70.929 3.1453e-08	27.448 0.070969	27.188 0.07553	23.267 0.0056246	59.674 2.3099e-06

Tabla 20. *Resumen de OddsRatio para el programa de Derecho*

Nivel	0	25	50	75	100	General
Alto	Exactas (1.08)** Matemáticas (1.02).	Hombre (2.15)** Exactas (1.08)**	Hombre (1.80). Edad (0.85).	Hombre (1.96). Exactas (1.07).		NoCapital (1.70). Edad (0.86)*

Nivel	0	25	50	75	100	General
	Hombre (3.08)***					Matemáticas (1.04) Hombre (2.61)***
Bajo	Capital (11.07)* Público (8.18)* Humanas (1.13). Hombre (4.91)*	Humanas (1.11).	Lenguaje (0.72)*			Exactas (0.91). Humanas (1.11).
Likelihood ratio test chisq = p.value=	62.959 6.7463e-07	40.318 0.0018885	34.156 0.012048	30.709 0.031096	Error	54.253 1.6759e-05

Tabla 21. *Resumen de OddsRatio para el programa de Desarrollo Familiar*

Nivel	0	25	50	75	100	General
Alto	Lenguaje (1.057)* Humanas (1.076)* Exactas (1.111)**	Exactas (1.104)* Humanas (1.115)* (8.024)*** NoCapital (1.944). Lenguaje (1.057).	NoCapital (2.819)* Humanas (1.143)* Exactas (1.127).	NoCapital (3.332)**	No significativos	Exactas (1.122)* Humanas (1.127)* NoCapital (2.720)**
Bajo	Edad (1.079)*	Mujer (2.558)*				Mujer (2.692)*

Nivel	0	25	50	75	100	General
		Edad (1.059).				
Likelihood ratio test chisq = p.value=	62.163 9.1061e-07	48.244 0.00013847	35.648 0.0078253	29.118 0.046955	17.47 0.041845	58.038 4.2292e-06

Tabla 22. *Resumen de OddsRatio para el programa de Diseño Visual*

Nivel	0	25	50	75	100	General
Alto	Hombre (2.08). Paga (2.51)*		Hombre (2.64)* Paga (2.20)* Exactas (1.07).	Hombre (1.86). Matemáticas (1.04).	NoCapital (3.03). Matemáticas (1.10)* Edad (0.70)*	Privado (0.35)* Hombre (3.21)** Paga (2.92)*
Bajo	Mujer (3.77)* Paga (3.75)**	Público (2.52). Mujer (4.76)** Paga (2.68)*				Privado (0.35)* Mujer (3.94)* Paga (2.97)*
Likelihood ratio test chisq = p.value=	40.661 0.0016951	35.763 0.0075669	30.016 0.037294	21.69 0.24606	17.777 0.037851	56.926 6.3591e-06

Tabla 23. *Resumen de OddsRatio para el programa de Enfermería*

Nivel	0	25	50	75	100	General
Alto	Exactas (1.05)*	Hombre (1.58)*	Capital (1.45).		Hombre (1.88).	Exactas (1.05)*

Nivel	0	25	50	75	100	General
			Matemáticas (0.96)* Hombre (2.08)**		Lenguaje (1.04). Edad (0.88). Exactas (0.92)*	Matemáticas (0.96)** Hombre (1.65)*
Bajo	Edad (1.12).	Mujer (2.01). Matemáticas (1.03)*	Lenguaje (1.29).			Matemáticas (1.03). Mujer (2.01).
Likelihood ratio test chisq = p.value=	21.618 0.24937	30.299 0.03464	30.27 0.034905	Error	22.456 0.0075404	40.315 0.0018906

Tabla 24. *Resumen de OddsRatio para el programa de Filosofía*

Nivel	0	25	50	75	100	General
Alto	Humanas (1.09)* Lenguaje (1.05). Matemáticas (1.04)*	Edad (1.16)* Humanas (1.08). Lenguaje (1.07)*	Exactas (1.11).		Edad (1.83). Exactas (1.98)** Humanas (0.66)* Hombre (198.48)*	Lenguaje (1.06)* Hombre (1.92).
Bajo	Exactas (0.88)* Matemáticas (1.04). Mujer (2.18)*	Exactas (0.92)*	Paga (10.84)* Lenguaje (1.19)*			Exactas (0.91)*

Nivel	0	25	50	75	100	General
Likelihood ratio test	50.638	30.871	31.247	Error	38.825	28.735
chisq = p.value=	6.0408e-05	0.029794	0.026949		0.003008	0.051712

Tabla 25. *Resumen de OddsRatio para el programa de Geología*

Nivel	0	25	50	75	100	General
Alto	Lenguaje (1.073)** Exactas (1.139)*** NoCapital (1.588). Paga (1.746).	NoCapital (2.044)* Lenguaje (1.075)* Exactas (1.107)* Matemáticas (1.043).		Paga (3.016)* Capital (2.794).	Privado (3.762)* Lenguaje (1.085). Matemáticas (0.925).	Matemáticas (1.060)* Exactas (1.098)* Humanas (1.082).
Bajo	Mujer (1.683)* Humanas (0.939)* Capital (2.101)**	Edad (1.105)* Humanas (0.944)* Capital (2.063)***	Privado (5.140).			Edad (1.115)* Humanas (0.940)** Capital (2.084)***
Likelihood ratio	101.03	93.879	24.848	15.317	17.194	90.509
chisq = p.value=	1.4311e-13	2.8854e-12	0.12916	0.082583	0.045767	1.1698e-11

Tabla 26. *Resumen de OddsRatio para el programa de Ingeniería de Alimentos*

Nivel	0	25	50	75	100	General
Alto	Exactas (1.15)*** Humanas (1.05)* Hombre (1.60)*	Exactas (1.14)*** Humanas (1.07)** Hombre (1.71)*	NoCapital (1.64). Lenguaje (1.06)* Exactas (1.13)***	Lenguaje (1.05)* Exactas (1.07)* Humanas (0.92)*	NoCapital (2.69)*	Exactas (1.16)*** Matemáticas (1.03).
Bajo	Mujer (1.95).	Humanas(0.91)* Mujer(1.93)				Humanas (0.90)**

Nivel	0	25	50	75	100	General
						Mujer (1.97)*
Likelihood ratio test chisq = p.value=	62.848 7.034e-07	65.905 2.1992e-07	30.933 0.00030392	18.087 0.034173	12.271 0.19846	60.543 1.6709e-06

Tabla 27 *Resumen de OddsRatio para el programa de Ingeniería de Computación*

Nivel	0	25	50	75	100	General
Alto	Público (1.69). Exactas (1.11)*** Matemáticas (1.02). Hombre (1.60).	Exactas (1.10)** Hombre (1.98)*			No significativos	Exactas (1.09). Hombre (2.18)*
Bajo		Capital (1.93). Edad (1.12).	Paga (5.19). Exactas (1.30).			Capital (2.02).
Likelihood ratio test chisq = p.value=	34.807 0.0099965	31.008 0.028729	21.317 0.26373	Error	11.644 0.23414	31.91 0.022532

Tabla 28 *Resumen de OddsRatio para el programa de Lic. Artes Escénicas*

Nivel	0	25	50	75	100	General
Alto		Exactas (1.171)** Lenguaje (1.063).	No significativos		No significativos	Exactas (1.123)** Lenguaje (1.047).

Nivel	0	25	50	75	100	General
Bajo	Privado(2.586)*	Privado (1.971). Mujer (1.855).		Error		Mujer (1.828).
Likelihood ratio test chisq = p.value=	22.767 0.1997	45.036 0.00040952	24.466 0.14033	Error	7.1573 0.62075	37.623 0.0043439

Tabla 29 *Resumen de OddsRatio para el programa de Lic. Biología y Química*

Nivel	0	25	50	75	100	General
Alto	NoCapital (2.07)** Edad (0.86) Humanas (1.09)***	NoCapital (1.85)* Exactas (1.09)* Humanas (1.09)** Matemáticas (1.04).	NoCapital (2.14). Exactas (1.12)*	NoCapital (2.63)* Edad (0.62)* Matemáticas (1.06).	Exactas (0.78)* Lenguaje (1.12)*	NoCapital (2.25)** Exactas (1.08)* Humanas (1.10)** Matemáticas (1.05)*
Bajo	Nocapital (1.73). Mujer (1.75).	Privado (2.23)* Humanas (0.93)* NoPaga (2.17). Mujer(1.58).				Privado (2.34)** Humanas (0.93)* NoPaga (2.61)*
Likelihood ratio test chisq = p.value=	78.984 1.2894e-09	74.427 7.9401e-09	24.945 0.12645	43.583 0.00066162	20.726 0.013925	90.936 9.8017e-12

Tabla 30.*Resumen de OddsRatio para el programa de Licenciatura en Lenguas Modernas*

Nivel	0	25	50	75	100	General
Alto	NoCapital (1.54). Edad (0.92). Exactas (1.09)** Humanas (1.10)** Lenguaje (1.03). Hombre (1.51).	Edad (0.83)** Humanas (1.11)** Lenguaje (1.03).	Edad (0.79)** Exactas (1.11)* Humanas (1.10)* Hombre (1.87).		Exactas (1.10). Humanas (1.10).	NoCapital (2.18)** Privado (2.19)* Edad (0.87)* Exactas (1.09)* Humanas (1.12)** Hombre (2.41)**
Bajo	Edad (1.13)* Lenguaje (1.08)*	Edad (1.12)* Humanas (0.91). Lenguaje (1.05). Mujer (1.90).	NoCapital (7.29). Lenguaje (1.20).			Edad (1.13)**
Likelihood ratio test chisq = p.value=	84.491 1.3875e-10	63.132 6.3188e-07	46.265 0.00027138	Error	21.843 0.23902	70.067 4.4027e-08

Tabla 31. *Resumen de OddsRatio para el programa de Lic. en Ciencias Sociales*

Nivel	0	25	50	75	100	General
Alto	Nocapital (1.62). Oficial (2.08). Matemáticas (1.04)** Mujer (1.88)*	NoCapital (2.31)* Exactas (1.09). Matemáticas (1.04).		Humanas (1.168)**	NoCapital (7.608)* Lenguaje (0.860)*	Matemáticas (1.050)* NoCapital (1.668). Exactas (1.073).
Bajo		Privado (1.89)*	Error			Privado (1.803). NoPaga (2.091).
Likelihood ratio test chisq = p.value=	40.566 0.0017466	42.974 0.00080701	Error	16.28 0.061259	16.312 0.060636	42.532 0.00093142)

Tabla 32. *Resumen de OddsRatio para el programa de Lic. en Educación Básica con énfasis en Educación Física*

Nivel	0	25	50	75	100	General
Alto	Exactas (1.067)* Edad (1.094)** Humanas (1.067)** Hombre (2.934)*** NoCapital	Hombre (3.169)*** Edad (1.116)*** Humanas (1.094)***	Hombre (2.051)* Exactas (1.067). Humanas (1.051).	Hombre (2.481)**		Paga (1.720)* Edad (1.080)* Humanas (1.063)* Hombre (3.111)***

Nivel	0	25	50	75	100	General
	(1.367).					
Bajo		Lenguaje (1.049).	Paga (14.80)**		Error	Lenguaje (1.052).
Likelihood ratio test chisq = p.value=	74.212 8.6448e-09	76.091 4.1008e-09	36.567 0.0059652	25.598 0.10931	Error	61.567 1.1391e-06

Tabla 33. *Resumen de OddsRatio para el programa de Lic. en Música*

Nivel	0	25	50	75	100	General
Alto	Exactas (1.07)* Hombre (2.16)*	Exactas (1.18)*** Humanas (1.07). Hombre (2.90)**			Hombre (4.97)* Humanas (0.82).	Exactas (1.14)** Hombre (2.62)*
Bajo	Privado (3.88). Lenguaje (0.89). NoPaga (5.75).	Privado (3.23). Edad (0.78). NoPaga (3.78)*	Error	Error		Privado (3.22). Edad (0.78)* NoPaga (3.68)*
Likelihood ratio test chisq = p.value=	46.528 0.00024833	66.741 1.5957e-07	Error	Error	12.506 0.18629	57.351 5.4432e-06

Tabla 34. *Resumen de OddsRatio para el programa de Medicina*

Nivel	0	25	50	75	100	General
Alto	Privado (1.6). Edad (0.86)* Exactas (0.95)** Humanas (1.06)** Lenguaje (0.95)** Mujer (1.85)**	Privado (1.82)* Edad (0.75)** Humanas (1.05)* Lenguaje (0.95)** Matemáticas (0.97)*** Hombre (1.91)**	Hombre (2.13)* Lenguaje (0.94)* Edad (0.51)*** Exactas (0.93)* Humanas (1.08)*	Privado (3.65). Hombre (3.54). Lenguaje (0.89)* Edad (0.73). Exactas (0.82)** Humanas (1.21)*	Lenguaje (0.88)* Matemáticas (1.15).	Capital (1.57). Edad (0.71)*** Lenguaje (0.96)* Matemáticas (0.98)* Hombre (1.78)*
Bajo	Edad (1.15)* Matemáticas (1.03)* Paga (2.7)*	Edad (1.17)** Humanas (0.89)*** Matemáticas (1.02).				Edad (1.17)** Humanas (0.89)*** Matemáticas (1.021).
Likelihood ratio test chisq = p.value=	85.992 7.5102e-11	99.714 2.4993e-13	50.101 1.031e-07	29.295 0.0005777	14.927 0.092948	102.03 9.4102e-14

Tabla 35. *Resumen de OddsRatio para el programa de Sociología*

Nivel	0	25	50	75	100	General
Alto	NoCapital (2.136)* NoPaga (2.294)* Lenguaje (1.047).	Matemáticas (1.064)* NoPaga (3.274). Exactas (1.110).	Exactas (0.778)* Matemáticas (1.360)** Capital (24.61). Humanas	No significativos	No significativos	Lenguaje (1.150)* Matemáticas (1.160)***

Nivel	0	25	50	75	100	General
			(1.271).			
Bajo	Privado (2.192)* Edad (1.125)**	Edad (1.078)*				Edad (1.069).
Likelihood ratiovalue=	39.899 0.0021542	33.038 0.016516	28.423 0.05591	52.784 2.8332e-05	9.7281 0.37294	52.632 2.9902e-05

Tabla 36. *Resumen de OddsRatio para el programa de Trabajo Social*

Nivel	0	25	50	75	100	General
Alto	Exactas (1.10)*** Humanas (1.07)** Lenguaje (1.03).	NoCapital (1.54)* Exactas (1.08)** Humanas (1.07)** Lenguaje (1.03)*		NoCapital (2.06)** Humanas (1.11)**	Edad (1.11).	NoCapital (1.57)* Exactas (1.07)* Lenguaje (1.03).
Bajo	Exactas (1.10). Mujer (2.78).	Edad (1.10). Mujer (2.66)*				Edad (1.09). Mujer (2.30).
Likelihood ratio test chisq = p.value=	9.937 0.0021289	38.32 0.0035125	Error	24.327 0.003813	10.927 0.28074	30.661 0.031496

Tabla 37. *Resumen de OddsRatio para el programa de Medicina Veterinaria y Zootecnia*

Nivel	0	25	50	75	100	General
Alto	Exactas (1.07)* Lenguaje (1.03)* Hombre (2.04)***	Hombre (2.70)**	Humanas (1.14)*** Hombre (1.65).	NoPaga (1.69). Exactas (1.13)**		Hombre (2.11)* Edad (0.80). Exactas (1.13)*
Bajo			Matemáticas (0.81).			
Likelihood ratio test	38.507	24.969	47.401	22.705	Error	32.198
chisq = p.value=	0.0033175	0.12578	0.00018473	0.0068932		0.020828

En la Tabla 38 se resumen los factores de influencia –positiva o negativa- indicando la cantidad de veces que un factor que aparece en los diferentes programas, dividiéndolo por su nivel de significancia estadística. Se resaltan en una escala de color el número de veces que el factor aparece en los diferentes programas (verde más veces, rojo menos veces).

Tabla 38 Resumen de los factores de influencia y sus apariciones por categoría de IAR y nivel de avance en créditos

	Alto							Bajo				
Factor/Signif.	0	25	50	75	100	Gen	T	0	25	50	Gen	T
Mujer	2					1	3	6	9		7	22
*	1					1	2	3	4		4	11
**	1						1		1			1
.								3	4		3	10
Hombre	10	8	7	2	2	10	39	1				1
*	3	4	3		1	5	16	1				1
**		3	1			2	6					

	Alto							Bajo				
Factor/Signif.	0	25	50	75	100	Gen	T	0	25	50	Gen	T
***	3	1				2	6					
.	4		3	2	1	1	11					
Capital			2	1		1	4	2	2		2	6
*								1				1
**								1				1
***									1		1	2
.			2	1		1	4		1		1	2
Nocapital	7	5	2	3	2	6	25	2	1	1	1	5
*	1	4	1	1	1	1	9		1		1	2
**	2			2		3	7	1				1
.	4	1	1		1	2	9	1		1		2
Paga	3		1	1		2	7	2	1	2	1	6
*	1		1	1		2	5	1	1	1	1	4
**								1		1		2
.	2						2					
NoPaga	1	1	1	1			4	1	2		3	6
*	1		1				2		1		2	3
.		1		1			2	1	1		1	3
Oficial	2					1	3	1	1		1	3
*						1		1			1	2
.	2						2		1			1
Privado	1	2	2	1	1	1	8	4	6		5	15
*		2	1		1	1	5	2	4		2	8
**								1			1	2
.	1		1	1			3	1	2		2	5
Edad	5	5	4	3	3	5	25	6	9	1	7	23
*	2	2		1	1	3	9	3	4		3	10
**	2	2	1				5	2	1		2	5
***		1	1			1	3					
.	1		2	2	2	1	8	1	4	1	2	8
Exactas	14	14	10	4	3	14	59	2	1		2	5
*	4	5	3	1	2	8	23	1	1		1	3
**	5	4	1	2	1	2	15					
***	5	3	1			2	11					
.		2	5	1		2	10	1			1	2
Humanas	9	11	7	4	1	5	37	3	7		6	16
*	3	3	3	2	1	2	14	1	4		2	7
**	4	5		2		2	13	1			2	3

	Alto							Bajo				
Factor/Signif.	0	25	50	75	100	Gen	T	0	25	50	Gen	T
***	1	1	1				3		1		1	2
.	1	2	3			1	7	1	2		1	4
Lenguaje	8	7	2	2	5	5	29	3	2	4	1	10
*	2	3	2	2	3	3	15	1		2		3
**	2	1					3					
.	4	3			2	2	11	2	2	2	1	7
Matemáticas	4	5	2	1	4	10	26	3	3	1	3	10
*	1	1	1		2	5	10	1	2		1	4
**	1		1			3	5					
***		1				1	2					
.	2	3		1	2	1	9	2	1	1	2	6

Adicionalmente, se presentan el tipo de influencia (positiva o negativa) por factor en la Tabla 39 para la categoría Alto y en la Tabla 40 para la categoría Bajo.

Tabla 39 Tipo de influencia por factor para la categoría IAR Alto

	Alto								
	Positiva				Negativa				
Factor	Cont	Pro	Máx	Mín	Cont	Pro	Máx	Mín	Tot
Mujer	3	2.65	4.21	1.85					3
Hombre	39	7.23	198	1.48					39
Capital	4	7.61	24.6	1.45					4
NoCapital	25	2.27	7.61	1.37					25
Paga	7	2.24	3.02	1.54					7
NoPaga	4	2.78	3.86	1.69					4
Oficial	3	2.2	2.84	1.69					3
Privado	8	2.53	3.76	1.6					8
Edad	7	1.22	1.83	1.08	18	0.78	0.92	0.51	25
Exactas	53	1.13	1.98	1.05	6	0.86	0.95	0.78	59
Humanas	35	1.1	1.27	1.05	2	0.79	0.92	0.66	37
Lenguaje	22	1.06	1.15	1.03	7	0.92	0.96	0.86	29
Matemáticas	21	1.08	1.36	1.02	5	0.96	0.98	0.93	26

Oddsratio Cont: conteo, Pro: promedio, Máx: máximo, Mín: mínimo, Tot: conteo por categoría IAR.

Tabla 40 Tipo de influencia por factor para la categoría IAR Bajo

	Bajo								
	Positiva				**Negativa**				
Factor	**Cont**	**Pro**	**Máx**	**Mín**	**Cont**	**Pro**	**Máx**	**Mín**	**Tot**
Mujer	22	2.41	4.76	1.58					22
Hombre	1	4.91	4.91	4.91					1
Capital	6	3.54	11.1	1.93					6
NoCapital	5	3.05	7.29	1.73					5
Paga	6	6.29	14.8	2.68					6
NoPaga	6	3.35	5.75	2.09					6
Oficial	3	4.5	8.18	2.52					3
Privado	15	2.43	3.88	1.8					16
Edad	21	1.12	1.17	1.06	2	0.78	0.78	0.78	23
Exactas	1	1.1	1.1	1.1	4	0.91	0.92	0.88	5
Humanas	3	1.12	1.13	1.11	13	0.91	0.94	0.87	16
Lenguaje	8	1.12	1.29	1.05	2	0.81	0.89	0.72	10
Matemáticas	7	1.03	1.04	1.02	3	0.9	0.95	0.81	10

La influencia positiva para la categoría Alto define un factor de predisposición para la excelencia académica, en cambio, para la categoría Bajo una influencia positiva define un factor de riesgo para el bajo rendimiento.

Correlaciones

Para determinar los componentes que tienen mayor influencia en el Indicador de Rendimiento Académico IAR, se presentan las matrices de correlación de Pearson para cada segmento de avance en créditos (Tabla 41 a

Tabla 46). Se resaltan en escala de grises las correlaciones (gris oscuro mayor correlación, gris claro menor correlación)

Tabla 41 *Matriz de correlación IAR0 y componentes*

Attributes	promedio_0	promedio_gana_0	IMA0	ICA0	PCC0	IAR0
promedio_0	1	0.498**	0.881**	0.848**	0.02*	0.931**
sig.		0.000	0.000	0.000	0.036	0.000
promedio_gana_0	0.498**	1	0.225**	0.201**	0.059**	0.376**
sig.	0.000		0.000	0.000	0.000	0.000
IMA0	0.881**	0.225**	1	0.977**	0.024*	0.961**
sig.	0.000	0.000		0.000	0.012	0.000
ICA0	0.848**	0.201**	0.977**	1	0.016.	0.948**
sig.	0.000	0.000	0.000		0.091	0.000
PCC0	0.02*	0.059**	0.024*	0.016.	1	0.206**
sig.	0.036	0.000	0.012	0.091		0.000
IAR0	0.931**	0.376**	0.961**	0.948**	0.206**	1
sig.	0.000	0.000	0.000	0.000	0.000	

(**) Significancia nivel 0.01 (*)Significancia nivel 0.05 (.)Significancia nivel 0.1

Tabla 42 Matriz de correlación IAR25 y componentes

Attributes	promedio_25	promedio_gana_25	IMA25	ICA25	PCC25	IAR25
promedio_25	1	0.489**	0.924**	0.902**	0.142**	0.95**
sig.		0.000	0.000	0.000	0.000	0.000
promedio_gana_25	0.489**	1	0.345**	0.337**	0.131**	0.452**
sig.	0.000		0.000	0.000	0.000	0.000
IMA25	0.924**	0.345**	1	0.985**	0.177**	0.973**
sig.	0.000	0.000		0.000	0.000	0.000
ICA25	0.902**	0.337**	0.985**	1	0.172**	0.966**
sig.	0.000	0.000	0.000		0.000	0.000
PCC25	0.142**	0.131**	0.177**	0.172**	1	0.323**
sig.	0.000	0.000	0.000	0.000		0.000
IAR25	0.95**	0.452**	0.973**	0.966**	0.323**	1
sig.	0.000	0.000	0.000	0.000	0.000	

Tabla 43 Matriz de correlación IAR50 y componentes

Attributes	promedio_50	promedio_gana_50	IMA50	ICA50	PCC50	IAR50
promedio_50	1	0.734**	0.875**	0.858**	0.298**	0.92**
sig.		0.000	0.000	0.000	0.000	0.000
promedio_gana_50	0.734**	1	0.448**	0.446**	0.16**	0.59**
sig.	0.000		0.000	0.000	0.000	0.000

IMA50	0.875**	0.448**	1	0.983**	0.314**	0.948**
sig.	0.000	0.000		0.000	0.000	0.000
ICA50	0.858**	0.446**	0.983**	1	0.309**	0.943**
sig.	0.000	0.000	0.000		0.000	0.000
PCC50	0.298**	0.16**	0.314**	0.309**	1	0.536**
sig.	0.000	0.000	0.000	0.000		0.000
IAR50	0.92**	0.59**	0.948**	0.943**	0.536**	1
sig.	0.000	0.000	0.000	0.000	0.000	

Tabla 44 Matriz de correlación IAR75 y componentes

Attributes	promedio_75	promedio_gana_75	IMA75	ICA75	PCC75	IAR75
promedio_75	1	0.784**	0.824**	0.799**	0.211**	0.89**
sig.		0.000	0.000	0.000	0.000	0.000
promedio_gana_75	0.784**	1	0.414**	0.409**	0.086**	0.587**
sig.	0.000		0.000	0.000	0.000	0.000
IMA75	0.824**	0.414**	1	0.973**	0.241**	0.918**
sig.	0.000	0.000		0.000	0.000	0.000
ICA75	0.799**	0.409**	0.973**	1	0.227**	0.907**
sig.	0.000	0.000	0.000		0.000	0.000
PCC75	0.211**	0.086**	0.241**	0.227**	1	0.521**
sig.	0.000	0.000	0.000	0.000		0.000
IAR75	0.89**	0.587**	0.918**	0.907**	0.521**	1
sig.	0.000	0.000	0.000	0.000	0.000	

Tabla 45 Matriz de correlación IAR100 y componentes

Attributes	promedio_100	promedio_gana_100	IMA100	ICA100	PCC100	IAR100
promedio_100	1	0.816**	0.731**	0.624**	0.149**	0.851**
sig.		0.000	0.000	0.000	0.000	0.000
promedio_gana_100	0.816**	1	0.323**	0.276**	0.13**	0.612**
sig.	0.000		0.000	0.000	0.000	0.000
IMA100	0.731**	0.323**	1	0.851**	0.117**	0.825**
sig.	0.000	0.000		0.000	0.000	0.000
ICA100	0.624**	0.276**	0.851**	1	0.061**	0.772**
sig.	0.000	0.000	0.000		0.002	0.000
PCC100	0.149**	0.13**	0.117**	0.061**	1	0.52**
sig.	0.000	0.000	0.000	0.002		0.000
IAR100	0.851**	0.612**	0.825**	0.772**	0.52**	1
sig.	0.000	0.000	0.000	0.000	0.000	

Tabla 46 Matriz de correlación IARGER y componentes

Attributes	prome_gen	prome_gana_gen	IMAGEN	ICAGEN	PCCGEN	IARGEN
promedio_gen	1	0.518**	0.932**	0.913**	0.089**	0.956**

sig.		0.000	0.000	0.000	0.000	0.000
promedio_gana_gen	0.518**	1	0.398**	0.391**	0.126**	0.498**
sig.	0.000		0.000	0.000	0.000	0.000
IMAGER	0.932**	0.398**	1	0.987**	0.116**	0.976**
sig.	0.000	0.000		0.000	0.000	0.000
ICAGER	0.913**	0.391**	0.987**	1	0.108**	0.968**
sig.	0.000	0.000	0.000		0.000	0.000
PCCGER	0.089**	0.126**	0.116**	0.108**	1	0.255**
sig.	0.000	0.000	0.000	0.000		0.000
IARGER	0.956**	0.498**	0.976**	0.968**	0.255**	1
sig.	0.000	0.000	0.000	0.000	0.000	

Las correlaciones entre el Indicador de Rendimiento Académico IAR y sus componentes, conserva la misma estructura para los diferentes segmentos de avance en créditos, una alta correlación entre el IAR y el promedio, el índice de créditos aprobados ICA y el índice de materias aprobadas IMA; y una baja correlación con el promedio de materias ganadas y el promedio de créditos cursados por semestre PCC.

Despliegue

Una de las ventajas de la implementación de la minería de datos con aplicaciones como RapidMiner, es que comprenden el proceso de punta a punta, iniciando desde el procedo de Extracción, Transformación y Carga, procesos de validación, de modelamiento y validación de los modelos generados, hasta la generación de informes con los resultados obtenidos en cada uno de los procesos previos.

El sistema implementado abarca desde la toma de la información del Sistema de Información Académica SIA hasta la entrega de los informes de resultados.

La definición del plan de ejecución del sistema debe ser una decisión de las directivas de acuerdo a la conveniencia administrativa, sin embargo, es conveniente ejecutar el proceso al finalizar cada semestre, de tal forma que se actualice el modelo a las nuevas condiciones y se determine si existe cambio en los factores de riesgo o de protección.

Análisis de Resultados

Los puntajes en las pruebas ICFES se presentan como factores de influencia para el IAR Alto en un conjunto importante de programas académicos (Exactas 59, Humanas 37, Lenguaje 29 y Matemáticas 26), similar a lo encontrado por (Pereira, Hernández, & Gómez), sin embargo, a pesar de haber encontrado significancia estadística en el valor predictivo de las pruebas, los coeficientes eran bajos. Adicionalmente, también fue evidente la menor influencia del área de Matemáticas.

Para el IAR Bajo, los puntajes ICFES disminuyen su presencia y se modifican las relaciones (Humanas 16, Lenguaje 10, Matemáticas 10 y Exactas 5), aunque permanecen, como se muestra en los resultados presentados por SPADIES en el cual se identifica que la deserción acumulada para estudiantes con bajos resultados en las pruebas ICFES llega al 57.5% frente al 35.8% en estudiantes con altos resultados (Angulo, 2009). El resultado en Exactas se muestra como un factor positivo para la excelencia académica, sin tener mayor influencia en el bajo rendimiento.

También, se observa que no siempre existe una influencia positiva de los puntajes de las pruebas ICFES como lo había identificado en la Universidad Tecnológica

de Pereira que un estudiante con el máximo puntaje ICFES en la prueba de matemática obtuvo bajo promedio en la materia de Matemáticas I (Carvajal Olaya, Trejos Carpintero, & Soto Mejía, 2004); sin embargo, ellos no encontraron relación significativa entre las pruebas ICFES y la nota en Matemáticas I. Además, encontraron que la influencia de los factores varía entre los programas académicos tanto por elementos vocacionales como por el nivel de exigencia académico que suele presentar en las universidades.

Otro de los factores que aparece con alta frecuencia en el IAR Alto con una influencia positiva es NoCapital que indica que el estudiante proviene de un municipio que no capital y que según lo explica Pautsch se debe al nivel de compromiso en este tipo de estudiantes debido a que deben convencer a su familia y obtener recursos para sus estudios (Pautsch, la Red Martínez, & Cutro, 2009).

El sexo del estudiante se muestra con un alto grado de influencia, para las dos categorías de IAR, el ser hombre se presenta como un factor de influencia para la excelencia (39 ocurrencias) y el ser mujer como un factor de riesgo para tener bajo rendimiento (22 ocurrencias), que según cita de Candamil en el estudio UN-Icfes 2004 existía una mayoría femenina en la deserción (Candamil Calle, Palomá Parra, & Sánchez Buitrago, 2009). Esta observación entra en oposición a lo encontrado por el Ministerio de Educación Nacional MEN, en la cual encuentre

mayores niveles de deserción para los hombres (Ministerio de Educación Nacional de Colombia, 2009).

La edad se presenta en ambas categorías del IAR con números importantes, presentando influencias concordantes para las categorías, la frecuencia en que aparece como influencia negativa para el IAR Alto (un aumento de la edad disminuye la relación rendimiento Alto/Medio, 18 ocurrencias) es similar a la frecuencia en que aparece como factor de riesgo para la categoría IAR Bajo (un aumento de la edad aumenta la relación rendimiento Bajo/Medio, 21 ocurrencias), de acuerdo a lo descrito por el MEN en el que indican que las personas que ingresan a una mayor edad acumulan tasas de deserción 17% más altas que las personas más jóvenes.

Los factores asociados al puntaje básico de matrícula (si paga o no paga matrícula académica) aparecen con unos oddsratios altos aunque con baja frecuencia; sin embargo, al contrario a lo hallado por SPADIES que muestra que a bajos ingresos familiares la deserción es mayor; en los resultados del estudio se observa que el riesgo es mayor para estudiantes que pagan matrícula académica. El informe de SPADIES no indica si la deserción por este factor se debió a elementos académicos.

CONCLUSIONES

El modelo de regresión logística logró determinar para la mayoría de casos los factores que influyen en que un estudiante de un programa esté en la categoría Alto o Bajo en relación con la categoría base Medio con niveles de significación estadística mínima del 10%.

Los factores de influencia en el indicador de rendimiento académico varían por programa académico y según el avance en créditos del estudiante, no se presentan factores comunes que permitan hacer generalizaciones, por lo que los programas de apoyo académico deberán hacer estas distinciones.

En la mayoría de casos, los factores que influyen en que la relación de Alto sobre Medio para un programa y porcentaje de avance en créditos, son diferentes a los que influyen en la relación Bajo sobre Medio para el mismo programa y porcentaje de avance en créditos, lo que muestra la diferencia entre lo requerido para lograr la excelencia y lo requerido para mantenerse en la universidad.

Todas las variables explicativas aparecen por lo menos en un modelo, por lo que se descartó su eliminación.

En algunos programas, la muestra de estudiantes en bajo rendimiento no es significativa.

Para la mayoría de programas, la muestra de estudiantes que se encontraban entre el 75 y el 100 por ciento de los créditos del programa no permitió darle significancia al modelo.

RapidMiner permite la construcción de complicados procesos de minería de datos en cada una de sus etapas, sin embargo, la documentación disponible es precaria y la construcción de estos se basa más en el ensayo y error, lo intuitivo del aplicativo, o los ejemplos y preguntas que la comunidad hace sobre este.

Desde finales de 2010, RapidMiner se integra con el proyecto R de estadística computacional lo que le ha brindado gran potencia desde el punto de vista estadístico, al incorporar los paquetes por defecto y los que su comunidad desarrolla; empero, sus resultados no siempre pueden ser retornados como un conjunto de datos de fácil manipulación, limitándose a la generación de un flujo de información sin una estructura fija que obliga a un complicado procesamiento posterior.

La construcción de la vista minable fue la actividad que mayor tiempo tomó en la realización del estudio (alrededor del 40%), es por esto que los proyectos de minería de datos deben tener futuras etapas que aprovechen este conjunto de datos para la aplicación de otras técnicas de minería o agreguen nuevas variables

que fortalezcan los modelos ya definidos, de tal forma que hagan rentable el esfuerzo inicial.

Es normal encontrar en sitios en internet y en anuncios de compañías de software, ambos relacionados con la minería de datos, que esta es un proceso simple, que sólo requiere de un buen conjunto de datos y una aplicación que realice el proceso, sin embargo, la elección de la técnica, la validación de los resultados y su análisis requieren de una fuerte fundamentación en estadística y de conocimiento del dominio; lo que puede dar lugar, si no se cuenta con ellas, a modelos sin significancia estadística o sin sentido para el dominio en estudio.

La metodología CRISP-DM permite la ejecución de un proyecto de minería de datos exitoso, pues además de definir los pasos a seguir de una manera clara y poder decidir cuales actividades incluir de acuerdo al alcance del proyecto, además indica claramente las salidas de cada actividad y hace recomendaciones puntuales para estas.

RECOMENDACIONES

El sistema propuesto se presenta como una plataforma de análisis, más que como un sistema cerrado, al que podrían incorporarse nuevas variables y nuevos análisis que generen otros modelos.

Estas nuevas variables o modelos deberían incorporar algunos de los elementos enunciados en los antecedentes, como son:

1. Relacionar las competencias de cada asignatura con los perfiles profesionales, así como los niveles de competencia en el ICFES en cada área buscando modelar el perfil vocacional del estudiante.
2. Integración con SPADIES.
3. Agregar al sistema los aspirantes no admitidos, con el fin de ofrecerles posibilidades acordes a su situación.
4. Analizar la relación de los indicadores académicos de rendimiento a través de los diferentes segmentos de avance en créditos de los estudiantes.
5. Los estilos de aprendizaje de los estudiantes.
6. Comportamiento en los sistemas de aprendizaje virtual.
7. Incorporar el desempeño laboral del egresado a través del Observatorio Laboral.

El Sistema de Información Académica es la principal fuente de información acerca de los estudiantes, la información recolectada desde el momento de la admisión debería fortalecerse para lograr recaudar información veraz sobre muchos de los factores que estudios previos han definido como relevantes en el rendimiento académico y la deserción; lo que permitiría contextualizar esos estudios a la realidad de la Universidad de Caldas.

BIBLIOGRAFÍA

Altablero. Ministerio de Educación Nacional de Colombia. (Noviembre de 2001). Sistema de Créditos Académicos. Recuperado el 30 de octubre de 2011, de http://www.mineducacion.gov.co/1621/article-87727.html

Candamil, M. D., Palomá, L. L., & Sánchez, J. O. (2009). Análisis de la deserción estudiantil en la Universidad de Caldas 1998-2006. Manizales.

Carvajal, P., Trejos, A. A., & Soto, J. (2004). Búsqueda de la relación entre áreas Icfes en matemáticas, física, lenguaje y rendimiento en matemáticas I y matemáticas II a través del análisis de componentes principales. *Scientia et Technica , Año X (26)* . Pereira.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., y otros. (2000). CRISP-DM 1.0 Step-by-step data mining guide. CRISPDM.

Consejo Nacional de Educación Superior. (1995). Acuerdo No. 06 - Políticas generales de acreditación.

Consejo Superior. Universidad de Caldas. (5 de diciembre de 2002). *Acuerdo No. 24* . Manizales, Caldas, Colombia.

Cuadras, C. M. (2010). *Nuevos métodos de análisis multivariante.* Barcelona: CMC Editions.

Dapozo, G., Porcel, E., López, M. V., Bogado, V., & Bargiela, R. (2006). Aplicación de minería de datos con una herramienta de software libre en la evaluación del rendimiento académico de los alumnos de la carrera de Sistemas de la FACENA-UNNE. *Anales del VII Workshop de Investigadores en Ciencias de la Computación* .

Durán, E., & Costaguta, R. (2007). Minería de datos para descubrir estilos de aprendizaje. *2 (42)* . Revista Iberoamericana de Educación.

Esper, R. J., & Machado, R. A. (2008). *La Investigación en Medicina: Bases teóricas y prácticas. Elementos de Bioestadística.* Buenos Aires: La Prensa Médica Argentina.

Garnica, E. (1997). El rendimiento estudiantil: una metodología para su medición. *(13)* , 7-26. Venezuela: Instituto de Investigaciones Económicas y Sociales. Universidad de Los Andes.

González Díaz, E., Pérez Hernández, Z., Espinosa Conde, I., & Alvarez Reyes, S. (2007). Obtención de patrones y reglas en el proceso académico de la Universidad de las Ciencias Informáticas utilizando técnicas de minería de datos. La Habana: Universidad de las Ciencias Informáticas.

González, E., Pérez, Z., Espinosa, I., & Alvarez, S. (2007). Obtención de patrones y reglas en el proceso académico de la Universidad de las Ciencias Informáticas utilizando técnicas de minería de datos. La Habana: Universidad de las Ciencias Informáticas.

Google. (s.f.). *Google Refine*. Recuperado el 17 de octubre de 2011, de http://code.google.com/p/google-refine/

Guzmán, C. (Enero de 2007). *Sistema de Prevención y Análisis de la Deserción en Las Instituciones de Educación Superior (SPADIES).* Recuperado el 20 de julio de 2010, de Proyecto "Estrategias para disminuir la deserción en Educación Superior": http://spadies.uniandes.edu.co/spadies2/recursos/MEN_ProyectoDesercion.pdf

ICFES. (2010). *Guia sobre que se evalua e interpretacion de resultados individuales SABER.* Recuperado el 16 de Octubre de 2011, de Icfes: http://www2.icfes.gov.co/index.php?option=com_docman&task=doc_view&gid=3348&Itemid=59

Kimball, R., & Caserta, J. (2004). *The Data WarehouseETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data.* Indianapolis: Wiley Publishing, Inc.

Luan, J. (2002). Data Mining and Knowledge Management in Higher Education - Potential Applications. Toronto: Annual Forum for the Association for Institutional Research.

Ministerio de Educación Nacional de Colombia. (2009). Deserción Estudiantil en la Educación Superior Colombiana. Metodología de seguimiento, diagnóstico y elementos para su prevención. Bogotá.

Pérez, C., & Santín, D. (2007). *Mineria de Datos. Técnicas y Herramientas.* Madrid: Paraninfo.

Proyecto R para estadística computacional. (s.f.). Recuperado el 16 de octubre de 2011, de http://www.r-project.org/

Pujari, A. K. (2001). *Data Mining Techniques.* Universities Press.

Rainardi, V. (2007). *Building a data warehouse with examples in SQL Server.* Nueva York: Apress.

Rapid-i. (s.f.). *RapidMiner*. Recuperado el 16 de octubre de 2011, de http://rapid-i.com/

Rocha, M., Pardo, C. A., Bohórquez, S. E., & Barrera, L. (2003). *Informe Nacional De Resultados. Exámenes De Estado De Calidad De La Educación Superior - Ecaes.* Bogotá: ICFES.

Valhondo, D. (2003). *Gestión del conocimiento: Del mito a la realidad.* Madrid: Ediciones Díaz de Santos.

Wikipedia. (s.f.). *R Lenguaje de programación*. Recuperado el 16 de octubre de 2011, de http://es.wikipedia.org/wiki/R_(lenguaje_de_programaci%C3%B3n)

Wikipedia. (s.f.). *RapidMiner.* Recuperado el 16 de octubre de 2011, de http://es.wikipedia.org/wiki/RapidMiner

Winters, T. d. (2006). Educational Data Mining: Collection and Analysis of Score Matrices for Outcomes-Based Assessment. *2006* . Riverside, California, Estados Unidos: University of California, Riverside.

Anexos

Anexo 1 Script SQL para el preprocesamiento en motor SQL

Anexo 2 Proceso RapidMiner

Anexo 3 Resultados de la Regresión Logística por Programa y Porcentaje de avance en créditos

Anexo 1 Script SQL para el preprocesamiento en motor SQL

Anexo 2 Proceso RapidMiner

Anexo 3 Resultados de la Regresión Logística por Programa y Porcentaje de avance en créditos