4 Descubrimiento de patrones de desempeño académico en las competencias genéricas

Discovery of Academic Performance Patterns in Generic Skills

Resumen

De acuerdo con los lineamientos Saber Pro del ICFES, todos los estudiantes de programas de pregrado deben presentar los módulos de competencias genéricas sin importar el programa de formación que cursen, el cual incluye competencias de lectura crítica, razonamiento cuantitativo, escritura e inglés. En este capítulo se presentan los resultados de cada una de las fases de la metodología CRISP-DM que se utilizó, con el fin de detectar patrones de desempeño académico en las competencias genéricas de los estudiantes colombianos de programas profesionales que presentaron las pruebas de Estado Saber Pro 2011-B. Con este objetivo, se describe cómo se selecciona, de las bases de datos del ICFES, la información sociodemográfica, económica, académica e institucional de estos estudiantes, que sirve de soporte para las posteriores fases de la metodología. Se detalla además cómo se construye, limpia y transforma un repositorio de datos y cómo a partir de él se descubren patrones asociados al buen o mal desempeño académico de los estudiantes en cada competencia, utilizando un modelo de clasificación basado en árboles de decisión.

Palabras clave: árboles de decisión, competencias genéricas, patrones de desempeño académico, metodología CRISP-DM.

Abstract

According to the ICFES Saber Pro guidelines, all undergraduates must take the generic skills modules, regardless of the undergraduate program they belong to, including critical reading, quantitative reasoning, writing and English skills. This chapter discusses the results of each phase of the CRISP-DM methodology used in order to detect academic performance patterns in generic skills of Colombian undergraduates who took the state exam Saber Pro 2011-B. To this end, it describes how the socio-demographic, economic, academic and institutional information of these students is selected from the ICFES databases to support subsequent phases of the methodology. It is also detailed how data repositories are built, purged and transformed and, based on this, how patterns associated with good or poor academic performance of students in each skill are discovered using the decision tree-based classification model.

Keywords: decision trees, generic skills, academic performance patterns, CRISP-DM methodology.

¿Cómo citar este capítulo?/How to cite this chapter?

Timarán-Pereira, S. R., Hernández-Arteaga, I., Caicedo-Zambrano, S. J., Hidalgo-Troya, A. y Alvarado-Pérez, J. C. (2016). Descubrimiento de patrones de desempeño académico en las competencias genéricas. En *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional* (pp. 101-150). Bogotá: Ediciones Universidad Cooperativa de Colombia. DOI: http://dx.doi.org/10.16925/9789587600490

Introducción

La investigación fue de tipo descriptivo bajo enfoque cuantitativo, aplicando un diseño no experimental. Para el descubrimiento de patrones de desempeño académico, se aplicó la metodología CRISP-DM (*Cross Industry Standard Process for Data Mining*) utilizada, principalmente, en los ambientes académico e industrial; metodología que se ha constituido en la guía de referencia más ampliamente utilizada en el desarrollo de este tipo de proyectos.

Comprensión del negocio o problema

En esta fase se realizaron las actividades que permitieron profundizar y apropiar de una manera completa el problema objeto de estudio, como también los objetivos y requisitos de esta investigación. Esto posibilitó la recolección y producción de los datos necesarios para efectuar una interpretación adecuada de los resultados.

De igual manera, la fase se fundamentó teóricamente sobre temas de calidad de la educación superior, desempeño académico, formación profesional desde el enfoque de competencias y competencias genéricas; además, se abordó el objeto de la investigación, es decir, la concepción y estructura de las pruebas Saber Pro 2011-2; todo lo cual se encuentra de forma detallada y precisa en el capítulo I.

Comprensión de los datos

En esta fase, el grupo de investigación identificó, recopiló y se familiarizó con la información disponible en las bases de datos del ICFES, sobre las competencias genéricas en las pruebas Saber Pro 2011-2 y sobre los datos sociodemográficos, económicos, académicos e institucionales de los estudiantes que presentaron esta prueba, pertenecientes a programas profesionales.

A partir de las bases de datos del ICFES, se construyó un repositorio inicial en el cual se encuentran almacenados los datos de todos los estudiantes de programas técnico profesionales, tecnológicos y profesionales que presentaron las pruebas Saber Pro 2011-2. Este repositorio inicial, que se denomina T153123A94, cuenta con 153 123 registros y 94 atributos. De este repositorio se seleccionaron únicamente los registros de estudiantes de programas profesionales objeto de este

estudio; de esta manera, quedó un repositorio con 97 068 registros y 94 atributos, identificado como T97068A94, el cual sirvió de base para las fases subsiguientes. En la tabla 17 se encuentra el diccionario de datos de este repositorio.

Tabla 17
Diccionario de datos del repositorio inicial T97068A94

N.°	Atributo	Descripción	Valores
1	estu_consecutivo	Código identificador del estudiante	
2	estu_cod_aplicacion	Nombre de la aplicación	EK+año-semestre
3	estu_genero	Género	M - Masculino F - Femenino
4	estu_nacimiento_dia	Día de nacimiento del estudiante	
5	estu_nacimiento_mes	Mes de nacimiento del estudiante	
6	estu_nacimiento_anno	Año de nacimiento del estudiante	
7	estu_pais_reside	País de residencia	
8	estu_estado_civil	Estado civil	Ver tabla 18
9	estu_disc_invidente	Discapacidad invidente	
10	estu_disc_sordo_con_interprete	Discapacidad sordo - requiere intérprete	
11	estu_disc_sordo_sin_interprete	Discapacidad sordo - no requiere intérprete	
12	estu_disc_motriz	Discapacidad motriz	
13	estu_disc_sordo_ceguera	Discapacidad sordo-ceguera	
14	estu_zona	Código identificador de la zona de la ciudad donde reside	
15	estu_anno_egreso	Año en que salió de bachiller	
16	econ_area_vive	Área donde vive	1. Cabecera municipal 2. Área rural
17	estu_reside_codmpio	Código DANE del municipio de residencia	Códigos DANE
18	estu_exam_codmpio_presentacion	Código del municipio donde presentó el examen	Códigos DANE
19	estu_exam_mpio_presentacion	Nombre municipio donde presentó el examen	
20	estu_exam_dpto_presentacion	Nombre del departamento donde presentó el examen	

(continúa)

(viene)

N.°	Atributo	Descripción	Valores
21	estu_exam_cod	Código identificador del examen/ prueba que está presentando el estudiante	Código 548
22	estu_exam_nombre	Nombre del examen que presentó el estudiante	Examen 2011-2
23	inst_cod_institucion	Código SNP de la institución de educación superior	
24	inst_nombre_institucion	Nombre de la institución de educación superior	
25	inst_origen	Origen institución educativa	Ver tabla 19
26	inst_caracter_academico	Carácter académico	Ver tabla 20
27	estu_prac_id_prgrm_academico	Código identificador programa académico	
28	estu_prgm_academico	Nombre programa académico	
29	estu_nivel_prgm_academico	Nivel del programa académico que cursa el estudiante	Universitario
30	estu_metodo_prgm	Metodología del programa académico bajo la cual cursa el estudiante	Presencial Semi-presencial Distancia (tradicional) A distancia (virtual)
31	dipo_codigomunicipio	Código municipio del programa	Código DANE
32	inst_cod_jornada	Código de la jornada de clases	Ver tabla 21
33	estu_area_conoc	Nombre del área de conocimiento a la que pertenece el programa académico del estudiante	Ver tabla 22
34	estu_nucleo_pregrado	Nombre del núcleo de pregrado al que pertenece el programa académico del estudiante	Ver tabla 23
35	estu_grupo_referencia	Nombre del grupo de referencia al que pertenece el programa académico del estudiante	Ver tabla 24
36	estu_cod_grupo_ref	Código del grupo de referencia al que pertenece el programa académico del estudiante	Ver tabla 24
37	estu_semestre_cursa	Semestre que cursa actualmente el estudiante	1 a 12 semestres 12 significa 12 o más.

(continúa)

(viene)

N.°	Atributo	Descripción	Valores
38	estu_pje_creditos	Porcentaje de créditos cursados y aprobados por el estudiante	Ver tabla 25
39	inst_vlr_matricula_ant	Valor anual de la matrícula el año anterior	Ver tabla 26
40	estu_sn_matricula_propio	Si canceló la matrícula año anterior con recursos propios	Sí (1) No (0)
41	estu_sn_matricula_padres	Si canceló la matrícula año anterior con recursos de los padres o familiares	Sí (1) No (0)
42	estu_sn_matricula_beca	Si canceló la matrícula año anterior con recursos de beca o subsidio	Sí (1) No (0)
43	estu_sn_matricula_credito	Si canceló la matrícula año anterior con recursos de crédito	Sí (1) No (0)
44	estu_titulo_bto	Título de bachiller obtenido por el estudiante	A-Académico N -Normalista T-Técnico
45	estu_exam_semestre_pretacion	Semestre en el que presentó el examen de Estado	1 o 2 semestre
46	estu_exam_anno_presentacion	Año en el que presentó el examen de Estado	
47	estu_hogar_actual	Tipo de hogar actual donde reside el estudiante	1. Es el habitual-permanente 2. Es temporal por razones de estudio u otra razón
48	fami_num_pers_grup_fam	Número total de personas que conforman el grupo familiar	
49	estu_sn_cabeza_fmlia	Si el estudiante es cabeza de familia	Sí (1) No (2)
50	fami_num_pers_cargo	Número de personas a cargo (cuando es cabeza de familia)	0 a 5 personas 5 representa 5 o más
51	fami_cod_educa_padre	Código nivel educativo más alto alcanzado por el padre del estudiante	Ver tabla 27
52	fami_cod_educa_madre	Código nivel educativo más alto alcanzado por la madre del estudiante	Ver Tabla 27
53	fami_sn_lee_escribe_padre	Si puede leer y escribir el padre del estudiante	Sí (1) No (2)

(continúa)

(viene)

N.°	Atributo	Descripción	Valores
54	fami_sn_lee_escribe_madre	Si puede leer y escribir la madre del estudiante	Sí (1) No (2)
55	fami_cod_ocup_padre	Ocupación actual del padre del estudiante (o última si falleció)	Ver Tabla 28
56	fami_cod_ocup_madre	Ocupación actual de la madre del estudiante (o última si falleció)	Ver tabla 28
57	estu_estrato	Estrato socioeconómico de la vivienda donde reside actualmente el estudiante según el recibo del servicio de energía eléctrica	Ver tabla 29
58	fami_num_hermanos	Número de hermanos del estudiante	
59	fami_nivel_hermano	Nivel educativo más alto alcanzado por los hermanos mayores del estudiante	Ver tabla 30
60	econ_cuartos	Número total de cuartos que dispone la residencia del estudiante	
61	fami_nivel_sisben	Nivel de clasificación del Sisben	Ver Tabla 31
62	econ_material_pisos	Material de los pisos que predomina en la vivienda donde reside el hogar habitual o permanente del estudiante	Ver Tabla 32
63	econ_sn_televisor	Si hay televisor en el hogar habitual del estudiante	Sí (1) No (0)
64	econ_sn_motocicleta	Si hay una motocicleta en el hogar habitual del estudiante	Sí (1) No (0)
65	econ_sn_energia	Si posee el servicio de energía eléctrica en el hogar habitual del estudiante	Sí (1) No (0)
66	econ_sn_acueducto	Si posee el servicio de acueducto en el hogar habitual del estudiante	Sí (1) No (0)
67	econ_sn_alcantarillado	Si posee alcantarillado en el hogar habitual del estudiante	Sí (1) No (0)
68	econ_sn_aseo	Si posee el servicio de aseo en el hogar habitual del estudiante	Sí (1) No (0)
69	econ_sn_estufa	Si posee estufa a gas o eléctrica en el hogar habitual del estudiante	Sí (1) No (0)
70	econ_sn_telefonia	Si posee el servicio de telefonía en el hogar habitual del estudiante	(1): Sí (0): no
71	econ_sn_internet	Si posee el servicio de Internet en el hogar habitual del estudiante	(1): Sí (0): no

(continúa)

(viene)

N.°	Atributo	Descripción	Valores
72	econ_sn_servicio_tv	Si posee el servicio de servicio de televisión por cable, satélite o parabólica en el hogar habitual del estudiante	(1): Sí (0): no
73	econ_sn_computador	Si hay un computador en el hogar habitual del estudiante	(1): Sí (0): no
74	econ_sn_celular	Si hay un teléfono celular en el hogar habitual del estudiante	(1): Sí (0): no
75	econ_sn_dvd	Si hay un reproductor DVD en el hogar habitual del estudiante	(1): Sí (0): no
76	econ_sn_lavadora	Si hay una máquina lavadora de ropas en el hogar habitual del estudiante	(1): Sí (0): no
77	econ_sn_microondas	Si hay un horno microondas en el hogar habitual del estudiante	(1): Sí (0): no
78	econ_sn_automovil	Si hay un automóvil particular en el hogar habitual del estudiante	(1): Si (0): no
79	econ_sn_horno	Si hay un horno eléctrico o a gas en el hogar habitual del estudiante	(1): Sí (0): no
80	econ_sn_nevera	Si hay una nevera o enfriador en el hogar habitual del estudiante	(1): Sí (0): no
81	infa_dormitorios	Número de dormitorios en el hogar habitual del estudiante	1 a 10 dormitorios 10 representa 10 o más
82	fami_ing_fmliar_mensual	ingresos mensuales familiares en salarios mínimos	Ver tabla 33
83	estu_trabaja	Si estudiante trabaja o no	Ver tabla 34
84	estu_horas_trabajo	Nro. de horas / semanal si trabaja	
85	estu_otro_idioma_lee	Si lee otro idioma	Ver tabla 35
86	estu_otro_idioma_habla	Si habla otro idioma	Ver tabla 35
87	estu_nivel_postgrado	Si el estudiante posee un postgrado	05 -Especialización 06 -Maestría 07 -Doctorado 08 -Ninguno
88	estu_etnia	Si el estudiante pertenece a una etnia	Ver tabla 36
89	mod_lectura_critica	Puntaje asignado al módulo de lectura crítica	
90	mod_comunica_escrita_punt	Puntaje asignado al módulo de comunicación escrita	

(continúa)

(viene)

N.°	Atributo	Descripción	Valores
91	mod_comunica_escrita_desem	Desempeño asignado al módulo de comunicación escrita	
92	mod_razona_cuantitativo_punt	Puntaje asignado al módulo de razonamiento cuantitativo	
93	mod_ingles_punt	Puntaje asignado al módulo de inglés	
94	mod_ingles_desem	Desempeño asignado al módulo de inglés	

Nota. Tomado de *Examen Saber Pro noviembre de 2011-II. Módulos de competencias genéricas y específicas disponibles. Evaluación de la calidad de la educación superior.* Instituto Colombiano para la Evaluación de la Educación [ICFES], 2011a.

Tabla 18
Estado Civil

Código	Descripción
01	Soltero(a)
02	Casado(a)
03	Viudo(a)
04	Separado(a) o divorciado
05	Unión libre

Nota. Tomado de *Examen Saber Pro noviembre de 2011-II. Módulos de competencias genéricas y específicas disponibles. Evaluación de la calidad de la educación superior.* Instituto Colombiano para la Evaluación de la Educación [ICFES], 2011a.

Tabla 19
Origen institución

Descripción
No oficial - corporación
No oficial - fundación
No oficial
Oficial
Oficial departamental
Oficial municipal
Oficial nacional
Régimen especial

Nota. Tomado de *Examen Saber Pro noviembre de 2011-II. Módulos de competencias genéricas y específicas disponibles. Evaluación de la calidad de la educación superior.* Instituto Colombiano para la Evaluación de la Educación [ICFES], 2011a.

Tabla 20
Carácter académico

Descripción
Académico
Escuela tecnológica
Institución tecnológica
Institución universitaria
Normalista
Técnica profesional
Técnica profesional
Universidad

Nota. Tomado de *Examen Saber Pro noviembre de 2011-II. Módulos de competencias genéricas y específicas disponibles. Evaluación de la calidad de la educación superior.* Instituto Colombiano para la Evaluación de la Educación [ICFES], 2011a.

Tabla 21
Código jornada

Código	Descripción
1	Diurno
2	Nocturno
3	Educación a distancia
4	Mixta
5	Registro calificado
6	Registro calificado - acreditación voluntaria
9	No aplica
10	Registro simple
11	Registro alta calidad
12	Jornada única

Nota. Tomado de *Examen Saber Pro noviembre de 2011-II. Módulos de competencias genéricas y específicas disponibles. Evaluación de la calidad de la educación superior.* Instituto Colombiano para la Evaluación de la Educación [ICFES], 2011a.

Tabla 22
Área del conocimiento

Descripción
Agronomía, veterinaria y afines
Bellas artes
Ciencias de la educación
Ciencias de la salud
Ciencias sociales y humanas
Economía, administración, contaduría y afines
Ingeniería, arquitectura, urbanismo y afines
Matemáticas y ciencias naturales

Nota. Tomado de *Examen Saber Pro noviembre de 2011-II. Módulos de competencias genéricas y específicas disponibles. Evaluación de la calidad de la educación superior.* Instituto Colombiano para la Evaluación de la Educación [ICFES], 2011a.

Tabla 23
Núcleo pregrado

Código	Descripción
1	Administración
2	Agronomía
3	Antropología, artes liberales
4	Arquitectura
5	Artes representativas
6	Artes plásticas, visuales y afines
7	Bacteriología
8	Bibliotecología, otros de ciencias sociales y humanas
9	Biología, microbiología y afines
10	Ciencia política, relaciones internacionales
11	Comunicación social, periodismo y afines
12	Contaduría pública
13	Deportes, educación física y recreación
14	Derecho y afines
15	Diseño
16	Diseño
17	Economía
18	Educación

(continúa)

(viene)

Código	Descripción
19	Enfermería
20	Filosofía, teología y afines
21	Física
22	Formación relacionada con el campo militar o policial
23	Geografía, historia
24	Geología, otros programas de ciencias naturales
25	Ingeniería administrativa y afines
26	Ingeniería agrícola, forestal y afines
27	Ingeniería agroindustrial, alimentos y afines
28	Ingeniería ambiental, sanitaria y afines
29	Ingeniería biomédica y afines
30	Ingeniería civil y afines
31	Ingeniería de minas, metalurgia y afines
32	Ingeniería de sistemas, telemática y afines
33	Ingeniería eléctrica y afines
34	Ingeniería electrónica, telecomunicaciones y afines
35	Ingeniería industrial y afines
36	Ingeniería mecánica y afines
37	Ingeniería química y afines
38	Instrumentación quirúrgica
39	Lenguas modernas, literatura, lingüística y afines
40	Matemáticas, estadística y afines
41	Matemáticas, estadística y afines
42	Medicina
43	Medicina veterinaria
44	Música
45	Nutrición y dietética
46	Odontología
47	Optometría, otros programas de ciencias de la salud
48	Otras ingenierías
49	Otros programas asociados a bellas artes
50	Psicología
51	Publicidad y afines
52	Química y afines

(continúa)

(viene)

Código	Descripción
53	Salud pública
54	Sin clasificar
55	Sin especificar
56	Sociología, trabajo social y afines
57	Terapias
58	Zootecnia

Nota. Tomado de *Examen Saber Pro noviembre de 2011-II. Módulos de competencias genéricas y específicas disponibles. Evaluación de la calidad de la educación superior.* Instituto Colombiano para la Evaluación de la Educación [ICFES], 2011a.

Tabla 24
Grupo de Referencia

Código	Detalle
1	Arquitectura y urbanismo
2	Artes, diseño, comunicación
3	Bellas artes y diseño
4	Ciencias agropecuarias
5	Ciencias económicas y administrativas
6	Ciencias militares y navales
7	Ciencias naturales y exactas
8	Ciencias sociales
9	Comunicación, periodismo y publicidad
10	Derecho
11	Educación
12	Grupo referencia nacional
13	Humanidades
14	Ingeniería
15	Medicina
16	Militar y policial
17	Recreación y deportes
18	Salud

Nota. Tomado de *Examen Saber Pro noviembre de 2011-II. Módulos de competencias genéricas y específicas disponibles. Evaluación de la calidad de la educación superior.* Instituto Colombiano para la Evaluación de la Educación [ICFES], 2011a.

Tabla 25
Porcentaje de créditos cursados y aprobados

Código	Descripción
0	No sigue el sistema de créditos
1	Menos del 40% (años 2008-1, 2008-2 y 2009-2)
1	Menos del 75% (año 2010-2)
2	Entre el 40% y el 59% (años 2008-1, 2008-2 y 2009-2)
2	Entre el 75% y el 80% (año 2010-2)
3	Entre el 60% y el 79% (años 2008-1, 2008-2 y 2009-2)
3	Entre el 81% y el 90% (año 2010-2)
4	El 80% o más (años 2008-1, 2008-2 y 2009-2)
4	Más del 90% (año 2010-2)

Nota. Tomado de *Examen Saber Pro noviembre de 2011-II. Módulos de competencias genéricas y específicas disponibles. Evaluación de la calidad de la educación superior.* Instituto Colombiano para la Evaluación de la Educación [ICFES], 2011a.

Tabla 26
Valor anual de la matrícula año anterior

Código	Descripción
1	No pagó matrícula
2	Menos de 500 mil
3	Entre 500 mil y menos de 1 millón
4	Entre 1 millón y menos de 3 millones
5	Entre 3 millones y menos de 5 millones
6	Más de 5 millones

Nota. Tomado de *Examen Saber Pro noviembre de 2011-II. Módulos de competencias genéricas y específicas disponibles. Evaluación de la calidad de la educación superior.* Instituto Colombiano para la Evaluación de la Educación [ICFES], 2011a.

Tabla 27
Educación del padre o madre

Código	Descripción
0	Ninguno
9	Primaria incompleta
10	Primaria completa

(continúa)

(viene)

Código	Descripción
11	Bachillerato incompleto
12	Bachillerato completo
13	Educación técnica o tecnológica sin título
14	Educación técnica o tecnológica con título
15	Educación profesional sin título
16	Educación profesional con título
17	Postgrado

Nota. Tomado de *Examen Saber Pro noviembre de 2011-II. Módulos de competencias genéricas y específicas disponibles. Evaluación de la calidad de la educación superior.* Instituto Colombiano para la Evaluación de la Educación [ICFES], 2011a.

Tabla 28
Ocupación del padre o madre

Código	Descripción
13	Empresario(a) dueño(a) de empresa industrial, comercial, agropecuaria o de servicios de más de 10 trabajadores.
14	Pequeño empresario(a) dueño(a) de microempresa o pequeño negocio familiar, de finca-parcela, que vive de su explotación.
15	Empleado con cargo como director(a) o gerente general de empresa privada -entidad pública.
16	Empleado(a) de nivel directivo (con personas a cargo) en empresa privada -entidad pública.
17	Empleado(a) de nivel técnico/profesional de empresa privada -entidad pública.
18	Empleado(a) de nivel auxiliar -administrativo de empresa privada -entidad pública.
19	Obrero u operario empleado(a) de empresa privada -entidad pública.
20	Profesional independiente ejerce su profesión sin vinculación laboral permanente en empresa privada -entidad pública.
21	Trabajador por cuenta propia ejerce un oficio sin vinculación laboral: comerciantes por cuenta propia; obreros u operarios independientes; jornaleros; trabajadores independientes que prestan servicios.
22	Hogar personas dedicadas principalmente a las labores del hogar.
23	Pensionado(a) persona que vive de una pensión por concepto de jubilación.
24	Rentista persona que vive de ingresos —rentas— generados de su patrimonio.

(continúa)

(viene)

Código	Descripción
25	Estudiante persona que la mayor parte del tiempo se dedica al estudio.
26	Otra actividad u ocupación.
99	No sabe.

Nota. Tomado de *Examen Saber Pro noviembre de 2011-II. Módulos de competencias genéricas y específicas disponibles. Evaluación de la calidad de la educación superior.* Instituto Colombiano para la Evaluación de la Educación [ICFES], 2011a.

Tabla 29
Estrato socioeconómico

Código	Descripción
1	Estrato 1
2	Estrato 2
3	Estrato 3
4	Estrato 4
5	Estrato 5
6	Estrato 6
8	Vive en una zona rural donde no hay estratificación socioeconómica

Nota. Tomado de *Examen Saber Pro noviembre de 2011-II. Módulos de competencias genéricas y específicas disponibles. Evaluación de la calidad de la educación superior.* Instituto Colombiano para la Evaluación de la Educación [ICFES], 2011a.

Tabla 30
Educación de los hermanos mayores

Código	Descripción
0	Ninguno
9	Primaria incompleta
10	Primaria completa
11	Bachillerato incompleto
12	Bachillerato completo
13	Educación técnica o tecnológica sin título
14	Educación técnica o tecnológica con título

(continúa)

(viene)

Código	Descripción
15	Educación profesional sin título
16	Educación profesional con título
17	Postgrado
88	No tiene hermanos mayores
99	No sabe

Nota. Tomado de *Examen Saber Pro noviembre de 2011-II. Módulos de competencias genéricas y específicas disponibles. Evaluación de la calidad de la educación superior.* Instituto Colombiano para la Evaluación de la Educación [ICFES], 2011a.

Tabla 31
Nivel de Sisben

Código	Descripción
1	Nivel 1
2	Nivel 2
3	Nivel 3
4	Está clasificada en otro nivel
5	No está clasificada por el Sisben

Nota. Tomado de *Examen Saber Pro noviembre de 2011-II. Módulos de competencias genéricas y específicas disponibles. Evaluación de la calidad de la educación superior.* Instituto Colombiano para la Evaluación de la Educación [ICFES], 2011a.

Tabla 32
Material de los pisos

Código	Descripción
1	Tierra, arena
2	Cemento, gravilla, ladrillo
3	Madera burda, tabla -tablón
4	Madera pulida, baldosa, tableta, mármol, alfombra
5	Madera pulida, mármol, alfombra -tapete de pared a pared

Nota. Tomado de *Examen Saber Pro noviembre de 2011-II. Módulos de competencias genéricas y específicas disponibles. Evaluación de la calidad de la educación superior.* Instituto Colombiano para la Evaluación de la Educación [ICFES], 2011a.

Tabla 33

Ingresos familiares mensuales

Código	Descripción
1	Menos de 1 SM
2	Entre 1 y menos de 2 SM
3	Entre 2 y menos de 3 SM
4	Entre 3 y menos de 5 S SM
5	Entre 5 y menos de 7 SM
6	Entre 7 y menos de 10 SM
7	10 o más SM

Nota. Tomado de *Examen Saber Pro noviembre de 2011-II. Módulos de competencias genéricas y específicas disponibles. Evaluación de la calidad de la educación superior.* Instituto Colombiano para la Evaluación de la Educación [ICFES], 2011a.

Tabla 34

Trabajo del estudiante

Código	Descripción
0	No
3	Sí, para contribuir a pagar su matrícula y/o los gastos del hogar
4	Sí, por ser práctica obligatoria del programa de estudios
5	Sí, para adquirir experiencia y/o recursos para sus gastos personales

Nota. Tomado de *Examen Saber Pro noviembre de 2011-II. Módulos de competencias genéricas y específicas disponibles. Evaluación de la calidad de la educación superior.* Instituto Colombiano para la Evaluación de la Educación [ICFES], 2011a.

Tabla 35

Lee o habla otro idioma

Código	Descripción
01	Inglés
02	Francés
03	Alemán
04	Italiano
05	Portugués
06	Japonés
99	Otro

Nota. Tomado de *Examen Saber Pro noviembre de 2011-II. Módulos de competencias genéricas y específicas disponibles. Evaluación de la calidad de la educación superior.* Instituto Colombiano para la Evaluación de la Educación [ICFES], 2011a.

Tabla 36
Etnia estudiante

Código	Descripción
01	Comunidades negras
02	Raizal (isleño)
03	Paez
04	Sikuani
05	Arhuaco
06	Emberá
07	Guambiano
08	Pijao
09	Wayúu
10	Zenú
11	Pasto
12	Cancuamo
13	Inga
14	Tucano
15	Huitoto
16	Cubeo
17	Comunidad Rom (gitana)
99	Otro

Nota. Tomado de *Examen Saber Pro noviembre de 2011-II. Módulos de competencias genéricas y específicas disponibles. Evaluación de la calidad de la educación superior.* Instituto Colombiano para la Evaluación de la Educación [ICFES], 2011a.

Con base en la conceptualización de desempeño académico y los antecedentes teóricos sobre los factores que intervienen en él —que se asocian y colindan unos con otros—, los 94 atributos del conjunto de datos T97068A94 se clasificaron en cuatro dimensiones: sociodemográfica, económica, académica e institucional. La tabla 37 presenta esta clasificación.

Tabla 37

Clasificación de atributos en dimensiones

Dimensión	Atributo
Sociodemográfica	estu_consecutivo, estu_cod_aplicación, estu_genero, estu_nacimiento_dia, estu_nacimiento_mes, estu_nacimiento_ anno, estu_pais_ reside, estu_estado_civil, estu_disc_invidente, estu_disc_sordo_con_ interprete, estu_disc_sordo_sin_interprete, estu_disc_motriz, estu_disc_sordo_ ceguera, estu_zona, econ_area_vive, estu_reside_codmpio, estu_hogar_actual, fami_num_pers_grup_fam, estu_sn_cabeza_fmlia, fami_num_pers_cargo, fami_cod_educa_padre, fami_cod_educa_madre, fami_sn_lee_escribe_padre, fami_sn_lee_escribe_madre, fami_cod_ocup_padre, fami_cod_ocup_madre, fami_num_hermanos, fami_nivel _hermano, estu _etnia
Económica	inst_vlr_matricula_ant, estu_sn_matricula_propio, estu_sn_matricula_padres, estu_sn_matricula_beca, estu_sn_matricula_credito, estu_estrato, econ_cuartos, fami_nivel_sisben, econ_material_pisos, econ_sn_televisor, econ_sn_motocicleta, econ_sn_energía, econ_sn_acueducto, econ_sn_alcantarillado, econ_sn_aseo, econ_sn_estufa, econ_sn_telefonia econ_sn_internet, econ_sn_servicio_tv, econ_sn _computador, econ_sn _celular , econ_ sn _dvd, econ_sn_lavadora, econ_sn _microondas, econ_sn_automovil, econ_sn_horno, econ_sn _nevera, infa_dormitorios, fami_ing _fmliar_mensual, estu_trabaja, estu_horas_trabajo
Académica	estu_anno_egreso, estu_exam_codmpio_presentacion, ,estu_exam_mpio_ presentacion , estu_exam_dpto_presentacion, estu_exam_cod, estu_exam_nombre, estu_prac_ id_ prgrm_academico, estu_prgm_academico, estu_nivel_prgm_ academico, estu_metodo_progr, dipo_codigomunicipio, inst_cod_jornada, estu_area_conoc, estu_nucleo_pregrado, estu_grupo _referencia, estu_cod_grupo_ref, estu_semestre_cursa, estu_pje_ creditos, estu_titulo_bto, estu_exam_semestre_pretacion, estu_exam_anno_ presentación, estu_otro_idioma _lee, estu_otro _idioma _habla, estu_nivel_postgrado
Institucional	inst_cod_institucion, inst_nombre_institucion, inst_origen, inst_caracter_academico, dipo_codigo_municipio
Clases	mod_lectura _critica, mod_comunica_escrita_punt, mod_comunica _escrita_desem, mod_razona_cuantitativo_punt, mod_ingles_punt, mod_ingles_desem

Nota. Elaboración propia.

Además de lo descrito, se realizó una exploración de los datos del repositorio T97068A94 y, mediante un análisis preliminar, se determinó qué variables son potencialmente importantes para el estudio. Este proceso se describe en el capítulo IV.

Preparación de los datos

En esta fase, los 94 atributos del repositorio base T97068A94, considerados por el ICFES como los más importantes para capturar la información de las pruebas Saber Pro 2011-2, fueron depurados teniendo en cuenta la calidad de los datos y las técnicas de minería de datos por aplicar; además, se limpiaron (eliminación de datos nulos y valores constantes) e integraron los datos; se generaron atributos adicionales a partir de los existentes por ganancia de información y se realizaron transformaciones o cambios de formato a los valores de los atributos que se consideraron necesarios. Como resultado de esta fase se obtuvo un repositorio de datos limpio y transformado, listo para aplicarle las técnicas de minería de datos.

En la tabla 38 se describen los atributos del repositorio que presentaron un porcentaje alto de valores nulos y en la tabla 39 se muestran los atributos con valores constantes y los atributos identificadores.

Tabla 38
Atributos con porcentaje alto de datos nulos

Atributo	Nulos (%)
estu_disc_ sordo_ ceguera, econ_area_vive, fami_sn_lee_escribe_padre, fami_sn_lee_escribe_madre, fami_num_hermanos, fami_nivel_hermano, econ_cuartos, econ_sn_televisor, econ_sn_motocicleta, econ_sn_energia, econ_sn_acueducto, econ_sn_alcantarillado, econ_sn_aseo, econ_sn_estufa, estu_otro_idioma_lee, estu_otro_idioma_habla, estu_nivel_postgrado, estu_etnia	100
estu_disc_invidente, estu_disc_sordo_con_interprete, estu_disc_sordo_sin_interprete, estu_zona	99,99
estu_disc motriz	99,92
estu_anno egreso	97,77

Nota. Elaboración propia.

Tabla 39
Atributos con valores constantes y atributos identificadores

Atributo	Valor
estu_consecutivo	SBPRO+identificador
estu_cod_aplicacion	EK20112
estu_pais_reside	Colombia
estu_exam_nombre	Examen específico 2011-2
estu_exam_cod	548
estu_nivel_prgm_academico	Universitario

Nota. Elaboración propia.

Se efectuó una primera selección de atributos y se descartaron aquellos que presentaron un alto porcentaje de valores nulos dada la imposibilidad de encontrar sus valores a través de fuentes externas de datos (véase tabla 12). De igual manera, se descartaron los atributos con valores constantes y aquellos que solo servían de identificadores de cada estudiante (véase tabla 13).

Se sabe que la alta dimensionalidad es un problema para el descubrimiento de patrones con minería de datos (Hernández et al., 2005). Uno de los criterios utilizados para resolver este problema es la reducción del número de atributos por analizar, a través de su transformación en nuevos atributos que generalicen los datos y que ofrezcan mayor información. Teniendo en cuenta este criterio, en el repositorio de datos T97068A94 se seleccionaron aquellos atributos que por sí mismos no tenían mayor significado, pero que al integrarlos en uno nuevo, adquirían mayor semántica. En este sentido, se crearon nuevos atributos y se redujo el número de variables para la investigación. En la tabla 40 se describe el proceso de construcción de nuevos atributos y se muestran aquellos que se reemplazan. Los atributos reemplazados fueron eliminados del repositorio.

Tabla 40
Nuevos atributos del repositorio de datos

Atributo	Descripción	Acción	Valores
estu_nacimiento_fecha	Fecha de nacimiento del estudiante.	Reemplaza a los atributos: estu _nacimiento_día estu_nacimiento_mes estu_nacimiento_año	Valores con formato DD/MM/YYYY

(continúa)

(viene)

Atributo	Descripción	Acción	Valores
estu_edad	Edad del estudiante en el momento de presentar la prueba.	Reemplaza al atributo: estu_nacimiento_fecha	Valores numéricos
estu_discapacidad	Tipo de discapacidad del estudiante.	Reemplaza a los atributos: estu_disc_invidente estu_disc_sordo_con_interprete estu_disc_sordo_sin_interprete estu_disc_motriz estu_disc_sordo_ceguera	I= invidente R= sordo con interprete S= sordo sin interprete M=discapacidad motriz C= sordo ceguera N = sin discapacidad
estu_reside_coddpto	Departamento de residencia del estudiante.	Reemplaza al atributo: estu_reside_codmpio	Códigos DANE
estu_pers_cargo	Si el estudiante tiene personas a cargo.	Reemplaza al atributo: fami_num_pers_cargo	Sí No
estu_financiacion_matricula	Forma de financiar el pago de la matrícula.	Reemplaza a los atributos: estu_sn_matricula_propio estu_sn_matricula_padres estu_sn_matricula_beca estu_sn_matricula_credito	Combinaciones en código binario de 4 Bits
fami_nivel_educa_padres	Máximo nivel educativo completo entre el padre y la madre.	Reemplaza a los atributos: fami_cod_educa_padre fami_cod_educa_madre	Primaria secundaria Técnico/ tecnológico Profesional Postgrado Ninguno
fami_ocup_padre	Ocupación del padre.	Reemplaza al atributo: fami_cod_ocup_padre	Directivo Empleado Empresario Hogar Independiente Otra Pensionado Profesional
fami_ocup_madre	Ocupación de la madre.	Reemplaza al atributo: fami_cod_ocup_madre	Los mismos valores de la ocupación del padre

(continúa)

(viene)

Atributo	Descripción	Acción	Valores
inst_tipo	Tipo de institución del estudiante.	Reemplaza a los atributos: inst_origen	Oficial Privada Régimen especial
inst_acreditada	Determina si la institución a la que pertenece el estudiante es acreditada o no según CNA.	Reemplaza a los atributos: inst_cód_institución inst_nombre_institucion	Acreditada No acreditada
inst_prog_acreditado	Determina si el programa al cual pertenece el estudiante es acreditado o no según CNA.	Reemplaza a los atributos: estu_prac_id_progr_ académico estu_progr_académico	Acreditado No acreditado
area_grupo_referencia	Áreas de los grupos de referencia de los programas.	Reemplaza al atributo: estu_grupo_referencia	Ciencias humanas Ciencias sociales Ciencias naturales y técnicas
inst_cod_dpto	Código DANE del departamento donde se ofrece el programa.	Reemplaza al atributo: dipo_cod_municipio	Códigos departamentos según DANE
inst_dpto_programa	Nombre del departamento donde se ofrece el programa académico.	Reemplaza al atributo: inst_cod_dpto	Nombres departamentos de Colombia según DANE
inst_programa_zona	Zona geográfica donde se ofrece el programa.	Reemplaza al atributo: inst_dpto_programa	Bogotá Eje Cafetero Caribe Centro Oriente Pacífico Centro Sur Llano
num_estudiantes_zona	Número de estudiantes por zona geográfica.	Nuevo atributo generado a partir del atributo: inst_programa_zona	Alto Medio Bajo
num_instituciones_ zona	Número de IES por zona geográfica.	Nuevo atributo generado a partir del atributo: inst_programa_zona	Alto Medio Bajo

(continúa)

(viene)

Atributo	Descripción	Acción	Valores
eco_condicion_ vivienda	Condición de la vivienda del estudiante	Reemplaza al atributo: econ_material_pisos	Buena Mala Regular
eco_condicion_hogar	Condición del hogar del estudiante	Reemplaza a los atributos: econ_ sn _dvd econ_sn_lavadora econ_sn _microondas econ_sn_horno econ_sn _nevera	Buena Regular Mala
eco_condicion_ transporte	Condición de transporte en el hogar del estudiante	Reemplaza al atributo: econ_sn_automovil	Particular Público
eco_condicion_tic	Condición de uso de TIC en el hogar del estudiante	Reemplaza a los atributos: econ_sn_telefonia econ_sn _celular econ_sn_internet econ_sn_servicio_tv econ_sn _computador	Buena Regular Mala
eco_condicion_vive	Condición de vida del estudiante	Reemplaza a los atributos: infa_dormitorios fami_num_pers_fam	Sin hacinamiento Hacinamiento medio Hacinamiento crítico
mod_lectura_critica_ desemp	Desempeño del estudiante en el módulo de lectura crítica.	Reemplaza al atributo: mod_lectura_critica	Sobre la media (>= la media) Bajo la media (< la media)
mod_comunica_ escrita_desemp	Desempeño del estudiante en el módulo de comunicación escrita.	Reemplaza a los atributos: mod_comunica_escrita_ punt mod_comunica_escrita_ desem	Sobre la media (>= la media) Bajo la media (< la media)
mod_razona_ cuantitativo_desemp	Desempeño del estudiante en el módulo de razonamiento cuantitativo.	Reemplaza al atributo: mod_razona_cuantitativo_ punt	Sobre la media (>= la media) Bajo la media (< la media)
mod_ingles_desemp	Desempeño del estudiante en el módulo de inglés	Reemplaza a los atributos: mod_ingles_punt mod_ingles_desem	Sobre la media (>= la media) Bajo la media (< la media)

Nota. Elaboración propia.

Con el fin de facilitar la detección de patrones de rendimiento académico, se discretizaron los valores numéricos de ciertos atributos, para lo cual se tuvo en cuenta un rango de valores y la proporcionalidad de las frecuencias por cada valor, a fin de evitar sesgos en la construcción de los modelos de minería de datos. En la tabla 41 se detalla este proceso para el atributo estu_edad.

Tabla 41
Valores discretizados del atributo estu_edad_examen

Valor	Número de estudiantes
Edad hasta 21	17 785
Edad 22	15 602
Edad 23	12 133
Edad de 25 a 26	12 022
Edad de 27 a 29	10 766
Edad de 30 a 34	9915
Edad mayor igual que 35	9863
Edad 24	8900
Sin dato	82
Total	97 068

Nota. Elaboración propia.

En esta investigación se consideró conveniente crear un nuevo atributo denominado inst_programa_zona, con el cual se reorganizarían las IES en zonas geográficas, de manera que la información suministrada aportara al descubrimiento de patrones de desempeño académico, ya que el atributo estu_zona de la base de datos del ICFES (véase tabla 1), que ubica la zona cardinal de la ciudad de residencia del estudiante, no aporta información relevante para este estudio.

Durante el proceso se consideraron varias opciones para la clasificación de dichas zonas, entre las cuales están la reorganización por zonas geográficas como tal, zonas político-administrativas, zonas culturales y zonas de los Órganos Colegiados de Administración y Decisión (OCAD); estas últimas son creadas y utilizadas en el marco del Sistema General de Regalías (SGR), que busca beneficiar a todos los departamentos y municipios del país, no únicamente a aquellos en los cuales se lleva a cabo la explotación de recursos naturales no renovables; por esta razón, se constituyen en una propuesta que se fundamenta en argumentos

de justicia social y que tiene la finalidad de distribuir de manera equitativa los recursos de regalías en todas las zonas del país. Todos los recursos del SGR financian proyectos de inversión presentados a los OCAD por las entidades territoriales, según lo reglamentado en el Decreto 1075 del 2012. Los OCAD se constituyen a nivel municipal, departamental, regional, nacional; además, está el OCAD de Ciencia Tecnología e Innovación, al cual pertenecen las universidades.

Estas zonas OCAD, en la actualidad, constituyen un reordenamiento territorial del país frente a las regiones geográficas que ha manejado históricamente y que ha permitido marcar diferencias significativas entre el centro y la periferia de Colombia. La ventaja de esta organización es que cada zona OCAD agrupa departamentos completos, con sus respectivos municipios. Gracias a esto es posible predecir cuál sería el rendimiento académico en las competencias genéricas en estas zonas con base en la información de las pruebas Saber Pro y, en futuros estudios, se puede analizar cómo han influido las regalías en el desarrollo de las IES y en el rendimiento académico de los estudiantes de estas zonas.

Los OCAD se encuentran distribuidos en seis zonas que cubren la totalidad del país. Teniendo en cuenta que Bogotá es la ciudad donde se concentran la gran mayoría de universidades, para esta investigación se consideró conveniente tomarla como una nueva zona. Así pues, para este estudio las zonas geográficas en las que se agrupan las IES son siete, como se detallan en la tabla 42.

Tabla 42
Valores del atributo inst_programa_zona

Zonas	Departamentos
Caribe	Atlántico, Bolívar, Cesar, Córdoba, Guajira, Magdalena, San Andrés, Providencia y Santa Catalina y Sucre.
Centro Sur	Amazonas, Caquetá, Huila, Putumayo y Tolima.
Centro Oriente	Boyacá, Cundinamarca, Norte de Santander y Santander.
Eje Cafetero	Antioquia, Caldas, Risaralda y Quindío.
Llano	Arauca, Casanare, Guainía, Guaviare, Meta, Vaupés y Vichada.
Pacífico	Cauca, Chocó, Nariño y Valle del Cauca.
Bogotá	Distrito Capital de Bogotá

Nota. Adaptado de *Colombia. Departamento Nacional de Planeación. Sistema General de Regalías SGR. (2012). Bogotá: DNP.*

Clasificadas las IES por zonas, se procedió a contar el número de estudiantes e IES por zonas y a crear dos nuevos atributos, a saber: num_estudiantes_zona y num_instituciones_zona, los cuales fueron discretizados.

Para los valores del atributo num_estudiantes_zona, se considera alto si el número de estudiantes es mayor que 20 000; medio si el número de estudiantes está entre 10 000 y 20 000, y bajo si es menor que 10 000. En la tabla 43 se muestran los datos en referencia.

Tabla 43

Valores discretizados del atributo num_estudiantes_zona

Zona	N.° de estudiantes	Valor
Bogotá	33 544	Alto
Eje Cafetero	16 500	Medio
Caribe	15 312	Medio
Centro Oriente	13 519	Medio
Pacífico	11 766	Medio
Centro Sur	5283	Bajo
Llano	1144	Bajo
Total de estudiantes	97 068	

Nota. Elaboración propia.

Para los valores del atributo num_instituciones_zona, se considera alto si el número de IES es mayor que 70; medio si el número de IES está entre 60 y 70, y bajo si es menor que 60. En la tabla 44 se muestran los resultados de tal discretización.

Tabla 44

Valores discretizados del atributo num_instituciones_zona

Zona	Número de Instituciones	Valor
Bogotá	73	Alto
Eje Cafetero	78	Alto
Caribe	64	Medio
Centro oriente	56	Bajo
Pacífico	64	Medio
Centro sur	21	Bajo
Llano	12	Bajo
	368	

Nota. Elaboración propia.

Para la nueva variable eco_condicion_hogar se tuvieron en cuenta los materiales de los pisos de la vivienda. En la tabla 45 se muestran estos valores y el material de los pisos.

Tabla 45
Valores del atributo eco_condicion_hogar

Material de los Pisos de la Vivienda	Condición Vivienda
Tierra, arena	Mala
Cemento, gravilla, ladrillo	Mala
Madera burda, tabla – tablón	Regular
Madera pulida, baldosa, tableta, mármol, alfombra	Buena
Madera pulida, mármol, alfombra - tapete de pared a pared	Buena

Nota. Elaboración propia.

Para asignar los valores a la nueva variable eco_condicion_hogar se creó un índice con los valores de los atributos de los electrodomésticos que esta variable reemplazó. El índice es el resultado de la sumatoria de los valores de la presencia (1) o ausencia (0) de cada electrodoméstico en la vivienda. Si el índice está entre 4 y 5, se asume que la condición del hogar es buena; si está entre 2 y 3 la condición del hogar es media, y si está entre 0 y 1 la condición del hogar es mala. En la tabla 46 se muestran las variables que intervienen en el cálculo del índice condición_hogar.

Tabla 46
Cálculo del índice de condición_hogar

Electrodomésticos	Sí	No
econ_sn_dvd	1	0
econ_sn_lavadora	1	0
econ_sn_microondas	1	0
econ_sn_horno	1	0
econ_sn_nevera	1	0

Nota. Elaboración propia.

De manera semejante se procedió para la discretización de los valores de la nueva variable eco_condicion_tic para que tome los valores buena, media y mala, con base en los servicios que dispone el estudiante. En la tabla 47 se muestran los atributos que intervienen en el cálculo del índice de condicion_tic.

Tabla 47
Cálculo del índice de condición_tic

Servicios	Sí	No
econ_sn_internet	1	0
econ_sn_servicio_tv	1	0
econ_sn_telefonia	1	0
econ_sn_celular	1	0
econ_sn_computador	1	0

Nota. Elaboración propia.

Para los valores de la variable eco_condicion_transporte se consideraron dos opciones: particular si el estudiante dispone de automóvil y público en caso contrario, para lo cual se tiene en cuenta la variable econ_sn_automovil que reemplaza.

Para asignar los valores a la nueva variable eco_condicion_vive se calculó el índice de hacinamiento. El hacinamiento refiere la relación entre el número de personas que habitan una vivienda o casa y el espacio o número de cuartos disponibles (Spicker, Alvarez y Gordon, s.f.).

Generalmente se aceptan los siguientes valores: sin hacinamiento, hasta 2,4; hacinamiento medio, de 2,5 a 4,9, y hacinamiento crítico, de 5,0 o más.

Teniendo en cuenta estos conceptos, el índice de hacinamiento para cada estudiante se obtiene dividiendo los valores de los atributos fami_num_pers_fam entre infa_dormitorios, variables que fueron reemplazadas por el atributo eco_condicion_vive. Los valores de este nuevo atributo se asignaron teniendo en cuenta los valores aceptados para el hacinamiento.

Para obtener los valores de la nueva variable area_grupo_referencia, se agruparon los valores del atributo estu_grupo_referencia en ciencias humanas, ciencias sociales y ciencias naturales y técnicas. Los datos se muestran en la tabla 48.

Tabla 48
Asignación de valores del atributo área_grupo_referencia

Grupo de referencia programas	Área grupo referencia programas
Ciencias económicas y administrativas	Ciencias sociales
Ingeniería	Ciencias naturales y técnicas
Educación	Ciencias sociales
Derecho	Ciencias humanas

(continúa)

(viene)

Grupo de referencia programas	Área grupo referencia programas
Salud	Ciencias humanas
Ciencias sociales	Ciencias sociales
Comunicación, periodismo y publicidad	Ciencias sociales
Medicina	Ciencias humanas
Bellas artes y diseño	Ciencias humanas
Ciencias agropecuarias	Ciencias naturales y técnicas
Ciencias naturales y exactas	Ciencias naturales y técnicas
Arquitectura y urbanismo	Ciencias sociales
Humanidades	Ciencias humanas
Militar y policial	Ciencias sociales
Recreación y deportes	Ciencias humanas
Ciencias militares y navales	Ciencias naturales y técnicas
Artes-diseño-comunicación	Ciencias sociales

Nota. Elaboración propia.

Como resultado de todos los procesos descritos, se obtuvo un nuevo repositorio de datos denominado T97068A35, limpio y transformado, listo para aplicarle las técnicas de minería de datos. La descripción del repositorio T97068A35 con 97068 registros y 35 atributos se muestra en la tabla 49, y están organizados por los factores sociodemográficos, económicos, institucionales y clases.

Tabla 49
Diccionario de datos del repositorio final T97068A35

N.°	Atributo	Descripción	Valores
		Sociodemográficos	
1	estu_genero	Género	M ,F
2	estu_edad	Edad del estudiante en el momento de presentar la prueba	Ver tabla 15
3	estado_civil	Estado civil del estudiante	Soltero, casado Separado/divorciado, Unión libre, viudo
4	estu_hogar_actual	Tipo de hogar actual donde reside el estudiante	Habitual/permanente Temporal
5	estu_sn_cabeza_fmlia	Si el estudiante es cabeza de familia o no	Sí, no

(continúa)

(viene)

N.°	Atributo	Descripción	Valores
6	estu_pers_cargo	Si el estudiante tiene personas a cargo o no	Sí, no
7	fami_nivel_educa_padres	Máximo nivel educativo completo entre el padre y la madre	Primaria, secundaria Técnico/tecnológico Profesional, postgrado Ninguno
8	fami_ocup_padre	Ocupación del padre	Directivo, empleado Empresario, hogar Independiente, otra Pensionado, profesional
9	fami_ocup_madre	Ocupación de la madre	Los mismos valores de la ocupación del padre
		Económicos	
10	estu_financiacion_matricula	Forma de financiar el pago de la matrícula	Propios, crédito, padres Beca y todas las combinaciones posibles entre estos valores
11	estu_estrato	Estrato socioeconómico del estudiante	Estratos 1 a 6 Zona rural sin estrato
12	fami_nivel_sisben	Nivel de clasificación en el Sisben al que pertenece el estudiante	Niveles 1, 2, 3, Otro nivel, No está en Sisben
13	econ_condicion_vivienda	Condición de la vivienda del estudiante	Buena, mala, regular
14	eco_condicion_hogar	Condición del hogar del estudiante	Buena, mala, regular
15	eco_condicion_transporte	Condición de transporte en el hogar del estudiante	Particular, público
16	eco_condicion_tic	Condición de uso de TIC en el hogar del estudiante	Buena, regular, mala
17	eco_condicion_vive	Condición de vida del estudiante	Sin hacinamiento Hacinamiento medio Hacinamiento crítico
18	fami_ing_fmliar_mensual	Ingresos mensuales familiares en salarios mínimos	Ver tabla 33
19	estu_trabaja	Si estudiante trabaja o no	Ver tabla 34
		Académicos	
20	estu_metodo_prgm	Metodología del programa académico bajo la cual cursa el estudiante	A distancia Presencial

(continúa)

(viene)

N.°	Atributo	Descripción	Valores
21	estu_area_conoc	Nombre del área de conocimiento a la que pertenece el programa académico del estudiante	Ver tabla 22
22	area_grupo_referencia	Área del grupo de referencia de los programas	Ciencias humanas Ciencias sociales Ciencias naturales y técnicas
23	estu_pje_creditos	Porcentaje de créditos cursados por el estudiante al realizar la prueba	Más del 90 Entre 81 y 90 Entre el 75 y el 80 Menos del 75 No sigue el sistema de créditos
24	estu_titulo_bto	Tipo de título de bachillerato obtenido	Académico, normalista Técnico
		Institucionales	
25	inst_tipo	Tipo de institución del estudiante	Oficial, privada Régimen especial
26	inst_caracter_ academico	Carácter académico de la IES	Escuela tecnológica Institución tecnológica Institución universitaria Técnica profesional Universidad
27	inst_acreditada	Si la institución donde pertenece el estudiante es acreditada o no según CNA	Acreditada, No acreditada
28	inst_prog_acreditado	Si el programa al cual pertenece el estudiante es acreditado o no según CNA	Acreditado, No acreditado
29	inst_programa_zona	Zona OCAD donde se ofrece el programa	Bogotá, eje cafetero Caribe, centro oriente Pacífico, centro sur Llano
30	num_estudiantes_zona	Número de estudiantes por zona OCAD	Alto, medio, bajo
31	num_instituciones_zona	Número de IES por zona OCAD	Alto, medio, bajo
		Clases	
32	mod_lectura_critica_ desemp	Desempeño del estudiante en el módulo de lectura crítica	Sobre la media (>= la media) Bajo la media (< la media)

(continúa)

(viene)

N.°	Atributo	Descripción	Valores
33	mod_comunica_ escrita_desemp	Desempeño del estudiante en el módulo de comunicación escrita	Sobre la media (>= la media) Bajo la media (< la media)
34	mod_razona_ cuantitativo_desemp	Desempeño del estudiante en el módulo de razonamiento cuantitativo	Sobre la media (>= la media) Bajo la media (< la media)
35	mod_ingles_desemp	Desempeño del estudiante en el módulo de inglés	Sobre la media (>= la media) Bajo la media (< la media)

Nota. Elaboración propia.

A partir del repositorio T97068A34 y con el fin de descubrir patrones asociados al rendimiento académico en competencias genéricas, se construyó por cada competencia (clase) un repositorio de datos. Para cada repositorio se tomaron los 31 atributos descritos en la tabla 49 y el respectivo atributo clase, dependiendo de la competencia. En cada repositorio se eliminaron los registros con valores nulos en el atributo clase (competencia). En la tabla 50 se presenta la relación de repositorios y competencias.

Tabla 50
Repositorios por cada competencia

Repositorio	Descripción
T97055A32LEC	Repositorio para análisis de la competencia de lectura crítica. Contiene 97055 registros y 32 atributos.
T95337A32ESC	Repositorio para análisis de la competencia de comunicación escrita. Contiene 95337 registros y 32 atributos.
T97057A32CUA	Repositorio para análisis de la competencia de razonamiento cuantitativo. Contiene 97057 registros y 32 atributos.
T96946A32ING	Repositorio para análisis de la competencia de inglés. Contiene 96946 registros y 32 atributos.

Nota. Elaboración propia.

Modelado

En esta fase se seleccionó la tarea de clasificación con árboles de decisión, como la técnica de minería de datos más adecuada para solucionar el problema objeto de la investigación. Con esta tarea se pretende obtener un modelo que permita predecir para los nuevos casos de estudiantes de programas profesionales los factores socio-demográficos, económicos, académicos e institucionales asociados a un probable buen o mal desempeño académico en las cuatro competencias genéricas, evaluadas en las pruebas Saber Pro 2011-2.

Descubrimiento de patrones de desempeño académico con árboles de decisión

La técnica de clasificación que se utilizó para el descubrimiento de patrones de desempeño académico en las competencias genéricas de las pruebas Saber Pro 2011-2 fue árboles de decisión. El modelo de clasificación basado en árboles de decisión es probablemente el más utilizado y popular por su simplicidad y facilidad para entender (Han y Kamber, 2001), (Sattler y Dunemann, 2001), (Timarán y Millán, 2006). La importancia de los árboles de decisión se debe a su capacidad de construir modelos interpretables, siendo este un factor decisivo para su aplicación. La clasificación con árboles de decisión considera clases disjuntas, de forma que el árbol conducirá a una y solo una hoja, asignando una única clase a la predicción (Hernández y Lorente, 2009).

El algoritmo de la herramienta Weka (Hall, Frank y Witten, 2011) utilizado para obtener el modelo de clasificación con árboles de decisión fue J48, el cual implementa al algoritmo C.45 (Quinlan, 1993). El algoritmo J48 se basa en la utilización del criterio del coeficiente de ganancia de información (*information gain ratio*). De esta manera, se consigue evitar que las variables con mayor número de posibles valores salgan beneficiadas en la selección. Además, el algoritmo incorpora una poda del árbol de clasificación una vez que este ha sido inducido (Hernández y Lorente, 2009). El parámetro más importante que se debe tener en cuenta para la poda es el factor de confianza C (*confidence level*), que influye en el tamaño y capacidad de predicción del árbol construido. Cuanto más baja se haga esa probabilidad, más se exigirá que la diferencia en los errores de predicción antes y después de podar sea más significativa para no podar. El valor por defecto de este factor es del 25%, y conforme va bajando este valor se permiten más operaciones

de poda; por lo tanto, se puede llegar a árboles cada vez más pequeños (García y Álvarez, 2010). Otra forma de variar el tamaño del árbol es a través del parámetro M que especifica el mínimo número de instancias o registros por nodo del árbol; es menos importante puesto que depende del número absoluto de instancias en el conjunto de datos de partida (Hall, Frank y Witten, 2011).

Antes de construir un modelo se debe definir un procedimiento para probar la calidad del modelo y su validez. Por tanto, para entrenar y probar un modelo de clasificación, el diseño de prueba específica divide los datos en dos conjuntos: entrenamiento y prueba. Existen diferentes medidas de evaluación del clasificador en Weka (Hall, Frank y Witten, 2011):

- Usar el conjunto de datos de entrenamiento (*Use training set*): se emplea todo el conjunto de datos para entrenar el modelo y después se prueba (esta técnica puede ser muy buena para ese conjunto de datos, pero puede ser poco precisa para nuevos datos).
- Proveer un conjunto de datos de prueba (*Supplied test set*): se emplea un conjunto de datos para entrenar y otro conjunto independiente al universo de los datos con los que se está trabajando para prueba (se corre el riesgo de que el conjunto de prueba no refleje o se corresponda con las características de los datos que se emplean para entrenar el modelo).
- Porcentaje de Partición (*Percentage Split*): se emplea un porcentaje aleatorio de datos para entrenar y otro porcentaje para probar; este método difiere del anterior por cuanto ambos conjuntos pertenecen al universo de datos con el que se está trabajando, por lo cual se elimina el riesgo que corre el anterior.
- Validación cruzada (*Cross validation*): este mecanismo permite reducir la dependencia del resultado del experimento en el modo como se realiza la partición (Hernández, Ramírez y Ferri, 2005). Para este caso, se utiliza el método de evaluación validación cruzada con *n* pliegues (n-fold cross validation). Esta es la opción por defecto y la más comúnmente utilizada. Este método consiste en dividir el conjunto de entrenamiento en *n* subconjuntos disjuntos de similar tamaño, llamados pliegues (*folds)* de forma aleatoria. El número de subconjuntos se puede introducir en el campo *Folds*. Posteriormente se realizan *n* iteraciones (igual al número de subconjuntos definido), donde, en cada una se reserva un subconjunto diferente para el conjunto de prueba y los restantes n-1 (uniendo todos los datos) para construir el modelo (entrenamiento). En cada iteración se calcula el error de muestra parcial del modelo. Por último,

se construye el modelo con todos los datos y se obtiene su error promediando los obtenidos anteriormente en cada una de las iteraciones. Otra ventaja de la validación cruzada es que la varianza de los *n* errores de muestra parciales permite estimar la variabilidad del método de aprendizaje con respecto al conjunto de datos. Comúnmente, se suelen utilizar 10 particiones (*10-fold cross validation*) (Hernández, Ramírez y Ferri, 2005).

Por otra parte, si se dispone de la matriz de confusión, es bastante sencillo evaluar o estimar el coste de un clasificador para un determinado conjunto de ejemplos. La matriz de confusión (*Confusion Matrix*) representa de forma detallada el número de instancias que son predichas por clase. La suma de los registros que se representan en cada fila i, i =1…n constituyen el número de instancias que realmente pertenecen a la clase i. Similarmente, la sumatoria de los ejemplos o registros en cada columna j, j = 1…n son las instancias que ha predicho el algoritmo al valor j de la clase. Los valores en la diagonal son los aciertos y el resto son los errores de clasificación (ejemplos que pertenecían a la clase i de la fila i y que fueron clasificados incorrectamente en otra) (Fernández, 2009).

Teniendo en cuenta los parámetros de evaluación anteriores y los repositorios de datos descritos en la tabla 43, se procedió a construir los diferentes árboles de decisión con el algoritmo J48. Se escogió como clase el desempeño en la competencia respectiva, y para evaluar la calidad del modelo y su validez, el método de validación cruzada, específicamente la validación cruzada con 10 pliegues por los mejores resultados que se obtienen.

Con el fin de obtener diferentes modelos de árboles por competencia y reglas de clasificación generalizadas, hasta reglas más detalladas, se establecieron cuatro porcentajes de prepoda del árbol para el factor *M* igual a 10%, 5%, 1% y 0,5% del total de registros del repositorio de datos, manteniendo constante el factor confianza C en el 25%. Por los mejores resultados obtenidos y por la facilidad de análisis de los patrones, se escogió el árbol construido con los parámetros M=1% y C=25%. Una vez construidos los árboles, se aplicó un proceso de pospoda para dejar las ramas y, por ende, las reglas más representativas, que son aquellas que sobrepasan un mínimo soporte del 0,5% y una confianza del 60%.

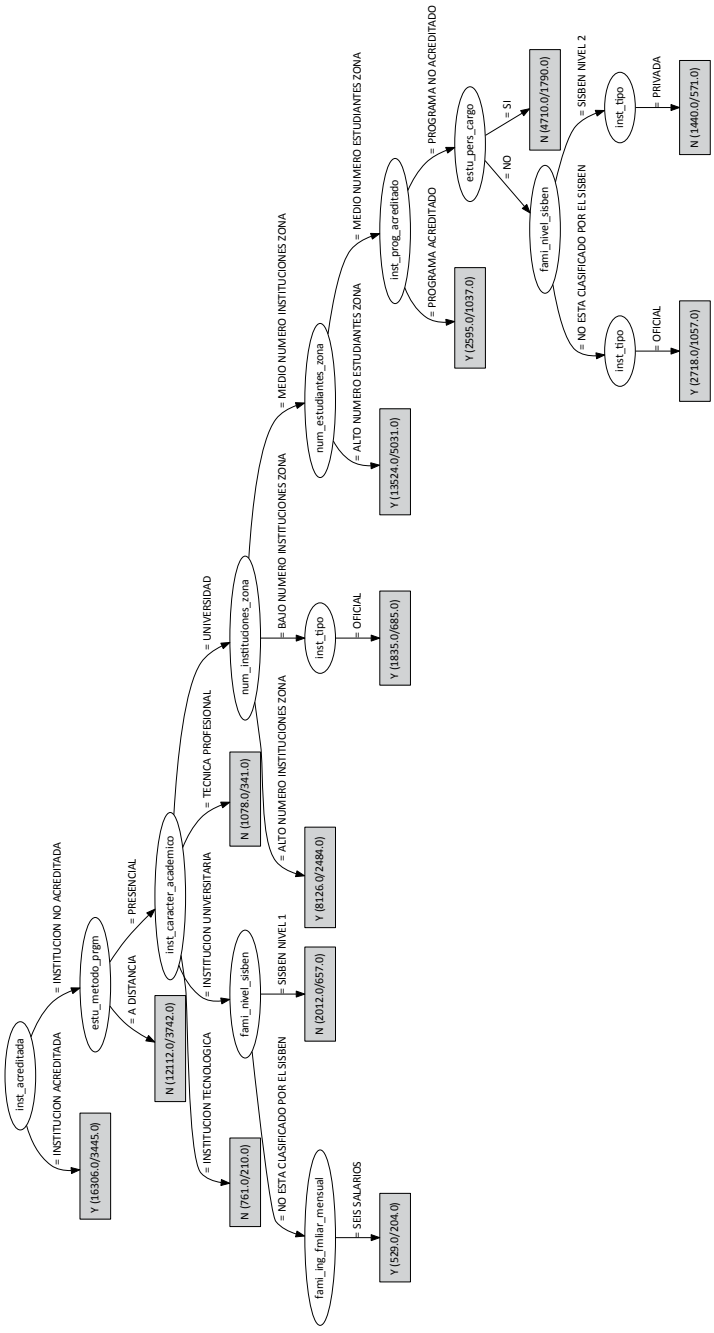

Figura 10. Árbol para lectura crítica

Nota. Elaboración propia.

```
=== 10 Fold Cross Validation ===
=== Summary ===
Correctly Classified Instances          62781               64.686 %
Incorrectly Classified Instances        34274               35.314 %
Kappa statistic                          0.2831
Mean absolute error                      0.442
Root mean squared error                  0.4704
Relative absolute error                 89.1798 %
Root relative squared error             94.4938 %
Coverage of cases (0.95 level)          99.999 %
Mean rel. region size (0.95 level)      99.9995 %
Total Number of Instances               97055

=== Confusion Matrix ===
     a     b <-- classified as
 37498 15548 | a = Y
 18726 25283 | b = N
```

Figura 11. Precisión y matriz de confusión del árbol de lectura crítica
Nota. Elaboración propia.

Descubrimiento de patrones de desempeño en lectura crítica

Para la construcción del árbol de decisión para el descubrimiento de patrones de desempeño en la competencia de lectura crítica en las pruebas Saber Pro 2011-2, se utilizó el conjunto de datos T97055A32LEC. El árbol construido con los parámetros M=971 (1%) y C=0,25 para la prepoda y confianza mayor o igual al 60% y soporte mayor o igual al 0,5%, se muestra en la figura 10. En la figura 11 se puede ver la precisión del árbol y su matriz de confusión, y en la figura 12 están las reglas más representativas con un soporte>=0,5% y confianza>=60%.

Descubrimiento de patrones de desempeño en comunicación escrita

Para la construcción del árbol de decisión para el descubrimiento de patrones de desempeño en la competencia de comunicación escrita, en las pruebas Saber Pro 2011-2, se utilizó el conjunto de datos T95337A32ESC. El árbol construido con los parámetros M=954 (1%) y C=0,25 para la prepoda y confianza mayor o igual al 60% y soporte mayor o igual al 0,5% se muestra en la figura 13. En la figura 14 puede verse la precisión del árbol y su matriz de confusión, y en la figura 15 se presentan las reglas más representativas con un soporte>=0,5% y confianza>=60%.

	Antecedente	Consecuente	Confianza (%)	Soporte (%)
1	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y inst_caracter_academico = INSTITUCION TECNOLOGICA	N	72.40	0.78
2	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = A DISTANCIA	N	69.11	12.48
3	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y inst_caracter_academico = TECNICA PROFESIONAL	N	68.37	1.11
4	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y inst_caracter_academico = INSTITUCION UNIVERSITARIA y fami_nivel_sisben = SISBEN NIVEL 1	N	67.35	2.07
5	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y inst_caracter_academico = UNIVERSIDAD y num_instituciones_zona = MEDIO NUMERO INSTITUCIONES ZONA y num_estudiantes_zona = MEDIO NUMERO ESTUDIANTES ZONA y inst_prog_acreditado = PROGRAMA NO ACREDITADO y estu_pers_cargo = SI	N	62.00	4.85
6	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y inst_caracter_academico = UNIVERSIDAD y num_instituciones_zona = MEDIO NUMERO INSTITUCIONES ZONA y num_estudiantes_zona = MEDIO NUMERO ESTUDIANTES ZONA y inst_prog_acreditado = PROGRAMA NO ACREDITADO y estu_pers_cargo = NO y fami_nivel_sisben = SISBEN NIVEL 2 y inst_tipo = PRIVADA	N	60.35	1.48
7	inst_acreditada = INSTITUCION ACREDITADA	Y	78.87	16.80
8	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y inst_caracter_academico = UNIVERSIDAD y num_instituciones_zona = ALTO NUMERO INSTITUCIONES ZONA	Y	69.43	8.37
9	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y inst_caracter_academico = UNIVERSIDAD y num_instituciones_zona = MEDIO NUMERO INSTITUCIONES ZONA y num_estudiantes_zona = ALTO NUMERO ESTUDIANTES ZONA	Y	62.80	13.93
10	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y inst_caracter_academico = UNIVERSIDAD y num_instituciones_zona = BAJO NUMERO INSTITUCIONES ZONA y inst_tipo = OFICIAL	Y	62.67	1.89
11	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y inst_caracter_academico = INSTITUCION UNIVERSITARIA y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y fami_ing_fmliar_mensual = SEIS SALARIOS	Y	61.44	0.55
12	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y inst_caracter_academico = UNIVERSIDAD y num_instituciones_zona = MEDIO NUMERO INSTITUCIONES ZONA y num_estudiantes_zona = MEDIO NUMERO ESTUDIANTES ZONA y inst_prog_acreditado = PROGRAMA NO ACREDITADO y estu_pers_cargo = NO y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y inst_tipo = OFICIAL	Y	61.11	2.80
13	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y inst_caracter_academico = UNIVERSIDAD y num_instituciones_zona = MEDIO NUMERO INSTITUCIONES ZONA y num_estudiantes_zona = MEDIO NUMERO ESTUDIANTES ZONA y inst_prog_acreditado = PROGRAMA ACREDITADO	Y	60.04	2.67

Figura 12. Reglas más representativas de lectura crítica

Nota. Elaboración propia.

Descubrimiento de patrones de desempeño en razonamiento cuantitativo

Para la construcción del árbol de decisión para el descubrimiento de patrones de desempeño en la competencia de razonamiento cuantitativo en las pruebas Saber Pro 2011-2, se utilizó el conjunto de datos T97057A32CUA. El árbol construido con los parámetros M=971 (1%) y C=0,25 para la prepoda y confianza mayor o igual al 60% y soporte mayor o igual al 0,5% se muestra en la figura 16. En la figura 17 está la precisión del árbol y su matriz de confusión, y en la figura 18 se ven las reglas más representativas con un soporte>=0,5% y confianza>=60%.

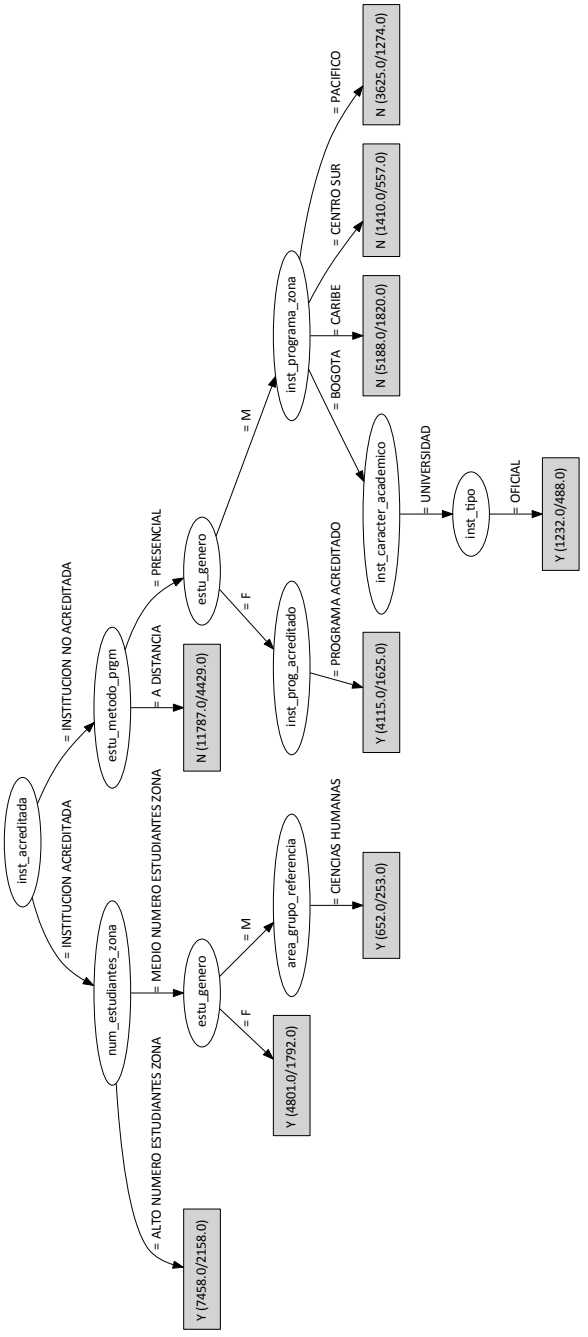

Figura 13. Árbol para comunicación escrita

Nota. Elaboración propia.

```
=== 10 Fold Cross Validation ===
=== Summary ===
Correctly Classified Instances          56669               59.4407 %
Incorrectly Classified Instances        38668               40.5593 %
Kappa statistic                          0.1876
Mean absolute error                      0.4762
Root mean squared error                  0.4882
Relative absolute error                 95.2592 %
Root relative squared error             97.6493 %
Coverage of cases (0.95 level)         100 %
Mean rel. region size (0.95 level)      99.9969 %
Total Number of Instances               95337

=== Confusion Matrix ===
   a     b <-- classified as
31234  17117 | a = N
21551  25435 | b = Y
```

Figura 14. Precisión y matriz de confusión del árbol de comunicación escrita

Nota. Elaboración propia.

	Antecedente	Consecuente	Confianza (%)	Soporte (%)
1	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y estu_genero = M y inst_programa_zona = CARIBE	N	64.92	5.44
2	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y estu_genero = M y inst_programa_zona = PACIFICO	N	64.86	3.80
3	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = A DISTANCIA	N	62.42	12.36
4	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y estu_genero = M y inst_programa_zona = CENTRO SUR	N	60.50	1.48
5	inst_acreditada = INSTITUCION ACREDITADA y num_estudiantes_zona = ALTO NUMERO ESTUDIANTES ZONA	Y	71.06	7.82
6	inst_acreditada = INSTITUCION ACREDITADA y num_estudiantes_zona = MEDIO NUMERO ESTUDIANTES ZONA y estu_genero = F	Y	62.67	5.04
7	inst_acreditada = INSTITUCION ACREDITADA y num_estudiantes_zona = MEDIO NUMERO ESTUDIANTES ZONA y estu_genero = M y area_grupo_referencia = CIENCIAS HUMANAS	Y	61.20	0.68
8	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y estu_genero = F y inst_prog_acreditado = PROGRAMA ACREDITADO	Y	60.51	4.32
9	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y estu_genero = M y inst_programa_zona = BOGOTA y inst_caracter_academico = UNIVERSIDAD y inst_tipo = OFICIAL	Y	60.39	1.29

Figura 15. Reglas más representativas de comunicación escrita

Nota. Elaboración propia.

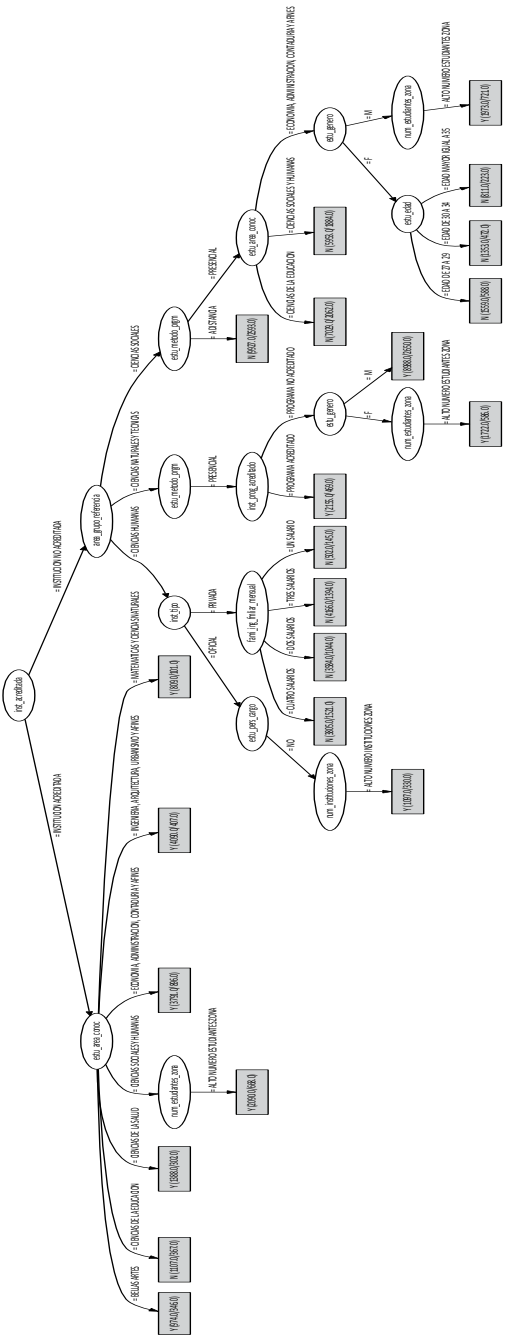

Figura 16. Árbol para razonamiento cuantitativo

Nota. Elaboración propia.

```
=== 10 Fold Cross Validation ===
=== Summary ===
Correctly Classified Instances          64549               66.5063 %
Incorrectly Classified Instances        32508               33.4937 %
Kappa statistic                          0.3295
Mean absolute error                      0.4273
Root mean squared error                  0.4626
Relative absolute error                 85.4605 %
Root relative squared error             92.5244 %
Coverage of cases (0.95 level)         100 %
Mean rel. region size (0.95 level)      99.9923 %
Total Number of Instances               97057

=== Confusion Matrix ===
   a     b <-- classified as
29093 19051 | a = Y
13457 35456 | b = N
```

Figura 17. Precisión y matriz de confusión del árbol de razonamiento cuantitativo
Nota. Elaboración propia.

Descubrimiento de patrones de desempeño en inglés

Para la construcción del árbol de decisión para el descubrimiento de patrones de desempeño en la competencia de inglés en las pruebas Saber Pro 2011-2, se utilizó el conjunto de datos T96946A32ING. El árbol construido con los parámetros M=970 (1%) y C=0,25 para la prepoda y confianza mayor o igual al 60% y soporte mayor o igual al 0,5%, se muestra en la figura 19. En la figura 20 se presenta la precisión del árbol y su matriz de confusión, y en la figura 21 pueden observarse las reglas más representativas con un soporte>=0,5% y confianza>=60%.

	Antecedente	Consecuente	Confianza (%)	Soporte (%)
1	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS SOCIALES y estu_metodo_prgm = A DISTANCIA	N	73.88	10.23
2	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS SOCIALES y estu_metodo_prgm = PRESENCIAL y estu_area_conoc = ECONOMIA, ADMINISTRACION, CONTADURIA Y AFINES y estu_genero = F y estu_edad = EDAD MAYOR IGUAL A 35	N	72.50	0.84
3	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS HUMANAS y inst_tipo = PRIVADA y fami_ing_fmliar_mensual = UN SALARIO	N	71.12	0.52
4	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS HUMANAS y inst_tipo = PRIVADA y fami_ing_fmliar_mensual = DOS SALARIOS	N	70.87	3.69
5	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS SOCIALES y estu_metodo_prgm = PRESENCIAL y estu_area_conoc = CIENCIAS DE LA EDUCACION	N	70.66	7.24
6	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS SOCIALES y estu_metodo_prgm = PRESENCIAL y estu_area_conoc = CIENCIAS SOCIALES Y HUMANAS	N	68.35	6.13
7	inst_acreditada = INSTITUCION ACREDITADA y estu_area_conoc = CIENCIAS DE LA EDUCACION	N	66.85	1.14
8	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS HUMANAS y inst_tipo = PRIVADA y fami_ing_fmliar_mensual = TRES SALARIOS	N	66.54	4.29
9	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS SOCIALES y estu_metodo_prgm = PRESENCIAL y estu_area_conoc = ECONOMIA, ADMINISTRACION, CONTADURIA Y AFINES y estu_genero = F y estu_edad = EDAD DE 30 A 34	N	65.11	1.39
10	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS SOCIALES y estu_metodo_prgm = PRESENCIAL y estu_area_conoc = ECONOMIA, ADMINISTRACION, CONTADURIA Y AFINES y estu_genero = F y estu_edad = EDAD DE 27 A 29	N	62.28	1.61
11	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS HUMANAS y inst_tipo = PRIVADA y fami_ing_fmliar_mensual = CUATRO SALARIOS	N	60.03	3.92
12	inst_acreditada = INSTITUCION ACREDITADA y estu_area_conoc = INGENIERIA, ARQUITECTURA, URBANISMO Y AFINES	Y	90.05	4.21
13	inst_acreditada = INSTITUCION ACREDITADA y estu_area_conoc = MATEMATICAS Y CIENCIAS NATURALES	Y	87.52	0.83
14	inst_acreditada = INSTITUCION ACREDITADA y estu_area_conoc = CIENCIAS DE LA SALUD	Y	78.24	1.43
15	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS NATURALES Y TECNICAS y estu_metodo_prgm = PRESENCIAL y inst_prog_acreditado = PROGRAMA ACREDITADO	Y	78.24	2.22
16	inst_acreditada = INSTITUCION ACREDITADA y estu_area_conoc = ECONOMIA, ADMINISTRACION, CONTADURIA Y AFINES	Y	76.37	3.91
17	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS NATURALES Y TECNICAS y estu_metodo_prgm = PRESENCIAL y inst_prog_acreditado = PROGRAMA NO ACREDITADO y estu_genero = M	Y	70.52	9.26
18	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS HUMANAS y inst_tipo = OFICIAL y estu_pers_cargo = NO y num_instituciones_zona = ALTO NUMERO INSTITUCIONES ZONA	Y	69.92	1.13
19	inst_acreditada = INSTITUCION ACREDITADA y estu_area_conoc = CIENCIAS SOCIALES Y HUMANAS y num_estudiantes_zona = ALTO NUMERO ESTUDIANTES ZONA	Y	68.04	2.15
20	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS NATURALES Y TECNICAS y estu_metodo_prgm = PRESENCIAL y inst_prog_acreditado = PROGRAMA NO ACREDITADO y estu_genero = F y num_estudiantes_zona = ALTO NUMERO ESTUDIANTES ZONA	Y	65.97	1.77
21	inst_acreditada = INSTITUCION ACREDITADA y estu_area_conoc = BELLAS ARTES	Y	64.48	1.00
22	inst_acreditada = INSTITUCION NO ACREDITADA y area_grupo_referencia = CIENCIAS SOCIALES y estu_metodo_prgm = PRESENCIAL y estu_area_conoc = ECONOMIA, ADMINISTRACION, CONTADURIA Y AFINES y estu_genero = M y num_estudiantes_zona = ALTO NUMERO ESTUDIANTES ZONA	Y	63.46	2.03

Figura 18. Reglas más representativas de razonamiento cuantitativo

Nota. Elaboración propia.

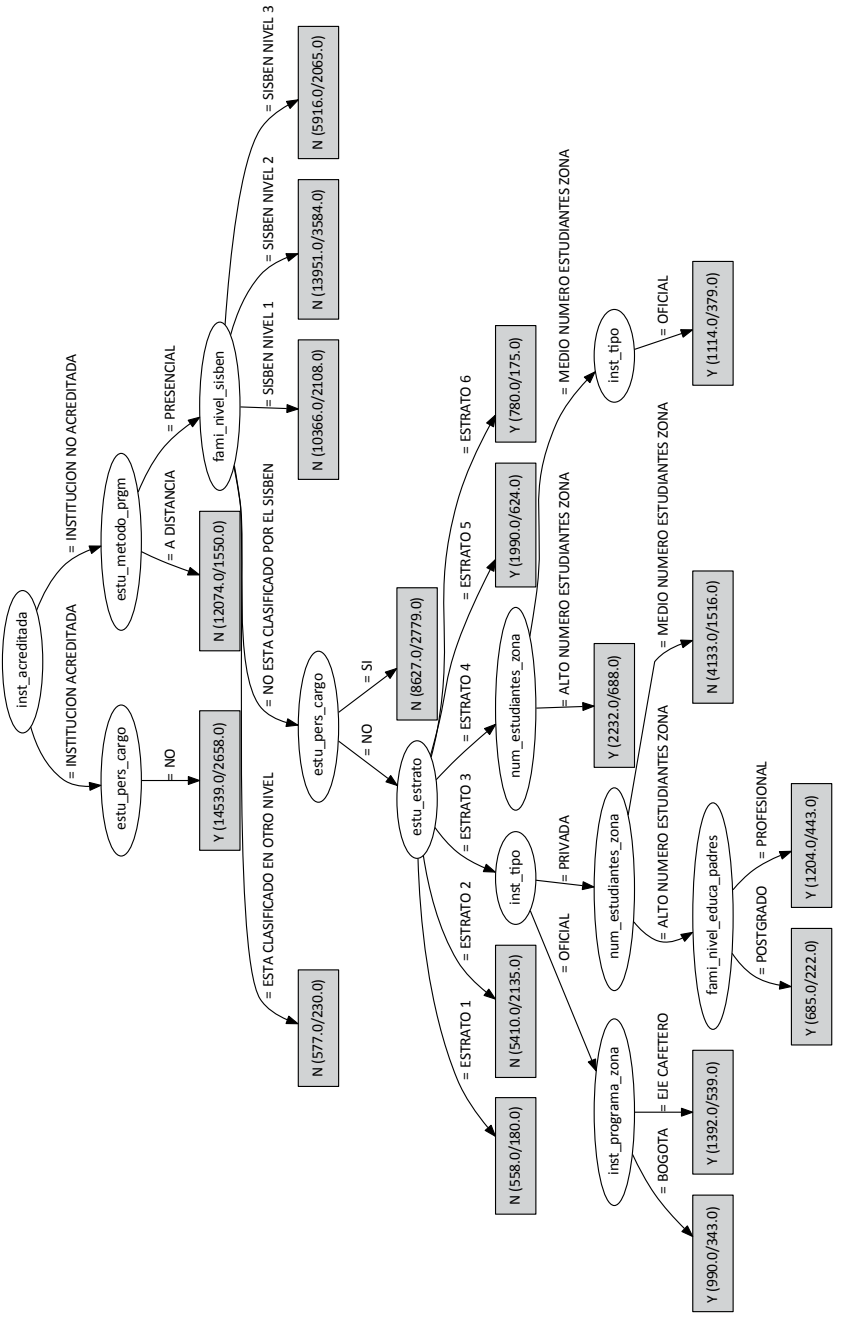

Figura 19. Árbol para inglés

Nota. Elaboración propia.

```
=== 10 Fold Cross Validation ===
=== Summary ===
Correctly Classified Instances          69600               71.7925 %
Incorrectly Classified Instances        27346               28.2075 %
Kappa statistic                          0.3931
Mean absolute error                      0.3815
Root mean squared error                  0.437
Relative absolute error                 78.6852 %
Root relative squared error             88.7538 %
Coverage of cases (0.95 level)          99.9979 %
Mean rel. region size (0.95 level)      99.999 %
Total Number of Instances               96946

=== Confusion Matrix ===
a        b <-- classified as
48836    8073 | a = N
19273    20764 | b = Y
```

Figura 20. Precisión y matriz de confusión del árbol de inglés
Nota. Elaboración propia.

Evaluación

En esta fase se evaluaron los patrones descubiertos con el fin de determinar su validez, remover los patrones redundantes o irrelevantes y traducir los patrones útiles en términos que sean entendibles por el usuario. La evaluación e interpretación de los patrones descubiertos se describe en el capítulo VI de resultados.

Implementación

En esta fase, el conocimiento descubierto se incorpora al existente y se podrá integrar a los procesos de toma de decisiones del ICFES y de las instituciones gubernamentales y académicas que velan por la calidad de la educación superior. Una vez estas instituciones intervengan los factores asociados al desempeño académico de los estudiantes de programas profesionales en las pruebas Saber Pro, será posible analizar los resultados y determinar sus efectos.

	Antecedente	Consecuente	Confianza (%)	Soporte (%)
1	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = A DISTANCIA	N	87.16	12.45
2	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = SISBEN NIVEL 1	N	79.66	10.69
3	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = SISBEN NIVEL 2	N	74.31	14.39
4	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y estu_pers_cargo = SI	N	67.79	8.90
5	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y estu_pers_cargo = NO y estu_estrato = ESTRATO 1	N	67.74	0.58
6	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = SISBEN NIVEL 3	N	65.09	6.10
7	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y estu_pers_cargo = NO y estu_estrato = ESTRATO 3 y inst_tipo = PRIVADA y num_estudiantes_zona = MEDIO NUMERO ESTUDIANTES ZONA	N	63.32	4.26
8	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y estu_pers_cargo = NO y estu_estrato = ESTRATO 2	N	60.54	5.58
9	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = ESTA CLASIFICADO EN OTRO NIVEL	N	60.14	0.60
10	inst_acreditada = INSTITUCION ACREDITADA y estu_pers_cargo = NO	Y	81.72	15.00
11	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y estu_pers_cargo = NO y estu_estrato = ESTRATO 6	Y	77.56	0.80
12	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y estu_pers_cargo = NO y estu_estrato = ESTRATO 4 y num_estudiantes_zona = ALTO NUMERO ESTUDIANTES ZONA	Y	69.18	2.30
13	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y estu_pers_cargo = NO y estu_estrato = ESTRATO 5	Y	68.64	2.05
14	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y estu_pers_cargo = NO y estu_estrato = ESTRATO 3 y inst_tipo = PRIVADA y num_estudiantes_zona = ALTO NUMERO ESTUDIANTES ZONA y fami_nivel_educa_padres = POSTGRADO	Y	67.59	0.71
15	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y estu_pers_cargo = NO y estu_estrato = ESTRATO 4 y num_estudiantes_zona = MEDIO NUMERO ESTUDIANTES ZONA y inst_tipo = OFICIAL	Y	65.98	1.15
16	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y estu_pers_cargo = NO y estu_estrato = ESTRATO 3 y inst_tipo = OFICIAL y inst_programa_zona = BOGOTA	Y	65.35	1.02
17	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y estu_pers_cargo = NO y estu_estrato = ESTRATO 3 y inst_tipo = PRIVADA y num_estudiantes_zona = ALTO NUMERO ESTUDIANTES ZONA y fami_nivel_educa_padres = PROFESIONAL	Y	63.21	1.24
18	inst_acreditada = INSTITUCION NO ACREDITADA y estu_metodo_prgm = PRESENCIAL y fami_nivel_sisben = NO ESTA CLASIFICADO POR EL SISBEN y estu_pers_cargo = NO y estu_estrato = ESTRATO 3 y inst_tipo = OFICIAL y inst_programa_zona = EJE CAFETERO	Y	61.28	1.44

Figura 21. Reglas más representativas de inglés

Nota. Elaboración propia.

Referencias

Artunduaga, M. (2008). *Variables que influyen en el rendimiento académico en la Universidad*. Recuperado de http://es.slideshare.net/1234509876/variables-del-rendimiento-academico-universidad

Colombia, Consejo Nacional de Acreditación [CNA] (2013). *Lineamientos para la acreditación de programas de pregrado*. Bogotá.

Colombia, Ministerio de Educación Nacional [MEN] (s. f.). Propuesta de lineamientos para la formación por competencias en educación superior. Recuperado de http://www.mineducacion.gov.co/1621/articles-261332_archivo_pdf_lineamientos.pdf

Davidson, M. y McKinney, G. (2001). Quantitative Reasoning: An Overview. *Western Washington University*. Recuperado de http://www.wwu.edu/vpue/documents/issue8.pdf

Fernández, G. (2009). *Extracción de Información de la Web usando Técnicas de Minería de Datos*. Recuperado de http://www.tdg-seville.info/Download.ashx?id=48

Garbanzo, G. (2007). Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde calidad de la educación superior pública. *Revista Educación*, *31*(1), 43-63.

García, M. y Álvarez, A. (2010). Análisis de Datos en WEKA –Pruebas de Selectividad. Recuperado el 5 de mayo de 2013, de http://www.it.uc3m.es/jvillena/irc/practicas/06-07/28.pdf

Gómez, G. y Soares, A. (2013). Diferencias de género con relación al desempeño académico en estudiantes de nivel básico. *Alternativas en Psicología*, *XVII*(28), 106-118.

Hall, M., Frank, E. y Witten, I. (2011). Practical Data Mining: Tutorials. *University of Waikato.* Recuperado de www.micai.org/2012/tutorials/Weka%20tutorials%20Spanish.pdf

Han, J. y Kamber, M. (2001). *Data Mining Concepts and Techniques*. San Francisco: Morgan Kaufmann Publishers.

Hernández, J., Ramírez, M. y Ferri, C. (2005). *Introducción a la Minería de Datos.* Madrid: Pearson Educación SA.

Hernández, E. y Lorente, R. (2009). *Minería de datos aplicada a la detección de Cáncer de Mama*. Madrid: Universidad Carlos III. Recuperado de http://tps5to-utn-frre.googlecode.com/svn/trunk/BI/Cancer%20de%20Mama/14.pdf

Instituto Colombiano para la Evaluación de la Educación [ICFES] (2011a). *Examen saber pro noviembre de 2011-II. Módulos de competencias genéricas y específicas disponibles. Evaluación de la calidad de la educación superior*. Recuperado de http://acofartes.org.co/docsweb/documento/ICFES%202011,%20M%C3%93DULOS%20COMPETENCIAS%20GEN%C3%89RICAS%20Y%20ESPEC%C3%8DFICAS.pdf

Moncada, L. y Rubio M. (2011). Determinantes inmediatos del rendimiento académico en los nuevos estudiantes matriculados en el sistema de educación superior a distancia del Ecuador: caso Universidad Técnica Particular de Loja. *Red Internacional de Educación Docente, RIED, 14*(2), 77-95.

Montero, E. y Villalobos, J. y Valverde, A. (2007). Factores institucionales, pedagógicos, psicosociales y sociodemográficos asociados al rendimiento académico y a la repetición estudiantil en la Universidad de Costa Rica: un estudio multinivel. *Revista Relieve, 13*(2), 215-234.

Parra, C., Mejía, L., Valencia, A., Castañeda, E. (2012). *Rendimiento académico de los estudiantes de primer semestre de pregrado de la Facultad de Ingeniería de la Universidad de Antioquia: cohorte 2012-2*. Medellín: Ingeniería y Sociedad. Recuperado de file:///C:/Users/Aspire/Downloads/16537-56603-1-PB%20(1).pdf

Pineda, C. y Pedraza, A. (2011). *Persistencia y graduación. Hacia un modelo de retención estudiantil para Instituciones de Educación Superior.* Bogotá: Arfo Editores e Impresores Ltda.

Quinlan, J. (1993). *C4. 5: programs for machine learning. Vol. 1*. Baltimore: Morgan Kaufmann Publishers. Recuperado de http://books.google.com.co/books?hl=es&lr=&id=HExncpjbYroC&oi=fnd&pg=PR7&dq=Programs+for+Machine+Leraning&ots=nLkbbRq2Yj&sig=Y5h5CQUdtbZjs1Fjd8ilbJfyRLE

Sarramona, J. (2002). *Evaluación de programas de educación a distancia*. Barcelona: Universidad Autónoma de Barcelona.

Sattler, K. y Dunemann, O. (2001). SQL database primitives for decision tree classifiers. En *Proceedings of the tenth international conference on Information and knowledge management* (pp. 379-386). Atlanta: CIKM. Recuperado de http://dl.acm.org/citation.cfm?id=502650

Seibold, J. (2000). La calidad integral en educación. Reflexiones sobre un nuevo concepto de calidad educativa que integre valores y equidad educativa. *Revista Iberoamericana de Educación, 23*. Recuperado de http://www.rieoei.org/rie23a07.htm

Spicker, P., Alvarez, S. y Gordon, D. (s. f.). *hH: Hacinamiento Hambruna*. Recuperado de http://biblioteca.clacso.edu.ar/gsdl/collect/clacso/index/assoc/D9393.dir/h.pdf

Timarán, R. y Millán, M. (2006). New algebraic operators and SQL primitives for mining classification rules. En *Computational Intelligence* [pp. 61-65]. Recuperado de http://www.actapress.com/PaperInfo.aspx?PaperID=29048&reason=500

Toro, J. y Villaveces, J. (2008). *El pensamiento matemático: una competencia genérica emergente.* Recuperado de http://www.mineducacion.gov.co/1621/articles-189357_archivo_pdf_matematica_1B.pdf)

Vásquez, C. y Rodríguez, M. (2007). La deserción estudiantil en educación superior a distancia: perspectiva teórica y factores de incidencia. *Revista Latinoamericana de Estudios Educativos*, *XXXVII*(3 y 4), 107-122.