Carreño G., Cipamocha L., Machuca L. & Pirazan G. (2020). Evaluación de la calidad del aire del relleno sanitario de la vereda Pirgua en la ciudad de Tunja (Boyacá). Cuaderno Activa, 12, 13-19.

Cuaderno Activa
•REVISTA CIENTÍFICA DE LA FACULTAD DE INGENIERÍA•

Analítica de datos aplicada al contexto universitario. Caso de estudio: pruebas Saber Pro

Data analytics applied to the university context. Case study: Saber Pro tests

José Sebastián Rengifo Collazos[1], Cristian David Sánchez Cobo[2], Cristian Fabián Delgado Manquillo[3], Cristian Olmedo Solarte Sarria[4], Fredy Alonso Vidal Alegría[5], Ricardo Timarán Pereira[6]

Recibido: 6 agosto de 2020 **Aprobado:** 7 septiembre de 2020

Resumen: En este artículo se presentan los resultados de un proyecto de investigación que tuvo como objetivo identificar patrones de rendimiento académico de los estudiantes de carreras profesionales afines con la informática, sistemas y computación, en las Pruebas Saber Pro de los años 2015 a 2017. Se utilizó la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), uno de los métodos más utilizados en proyectos de minería de datos. Esta metodología contempla seis fases: comprensión de negocio, comprensión del problema, comprensión de los datos, modelado, evaluación e implementación. De acuerdo con esta metodología, se construyó un repositorio de datos a partir de las bases de datos del ICFES de las Pruebas Saber Pro. Este repositorio se limpió, se transformó y se le aplicaron técnicas descriptivas en minería de datos para la obtención de patrones de rendimiento académico. Finalmente, se evaluaron y se interpretaron los resultados. El conocimiento generado servirá como soporte a instituciones de educación superior que ofrecen programas relacionados con informática, sistemas y computación, con el fin de generar estrategias que permitan mejorar la calidad de la educación que se imparte en estas carreras.

Palabras clave: CRISP-DM, Identificación de patrones, Minería de datos, Rendimiento académico en Pruebas Saber Pro, Técnicas descriptivas.

1 Autor correspondiente: Ingeniero Informático. Institución Universitaria Colegio Mayor del Cauca. Popayán, Colombia. Correo electrónico: sebastianrengifo@unimayor.edu.co

2 Autor correspondiente: Ingeniero en Automática Industrial. Institución Universitaria Colegio Mayor del Cauca. Popayán, Colombia. Correo electrónico: cdsc-cdsc@hotmail.com

3 Autor correspondiente: Ingeniero en Informática. Institución Universitaria Colegio Mayor del Cauca. Popayán, Colombia. Correo electrónico: cdelgado@unimayor.edu.co

4 Autor correspondiente: Ingeniero Informático. Institución Universitaria Colegio Mayor del Cauca. Popayán, Colombia. Correo electrónico: solarte@unimayor.edu.co

5 Autor correspondiente: Magíster en Educación para la Diversidad. Institución Universitaria Colegio Mayor del Cauca. Popayán, Colombia. Correo electrónico: fvidal@unimayor.edu.co

6 Autor correspondiente: Doctor en Ingeniería Ciencias de la Computación. Institución Universitaria Colegio Mayor del Cauca. Popayán, Colombia. Correo electrónico: ritimar@udenar.edu.co

Abstract: This article presents the results of a research project that had as goal to identify patterns of academic performance of the students of professional careers related to the computer science, systems and computation in the Saber-Pro tests of the years 2015 to 2017. CRISP-DM (Cross Industry Standard Process for Data Mining) methodology was used, one of the most used methods in data mining projects. This methodology contemplates 6 phases: business understanding, problem understanding, data understanding, modeling, evaluation and implementation. In accordance with this methodology, a data repository was built from the Saber Pro test ICFES databases. This repository was cleaned, transformed and descriptive techniques were applied in data mining to obtain patterns of academic performance. Finally, the results were evaluated and interpreted. The knowledge generated will serve as support to higher education institutions that offer programs related to computer science, systems and computing in order to generate strategies to improve the quality of education provided in these careers.

Keywords: Academic performance in Saber Pro, CRISP-DM, Data Mining, Descriptive techniques, Pattern identification.

Introducción

El examen de Estado de las Pruebas Saber Pro es un instrumento estandarizado para la evaluación externa de la calidad de la educación superior, además forma parte con otros procesos y acciones, de un conjunto de instrumentos que el Gobierno Nacional dispone para evaluar la calidad del servicio público educativo y ejercer su inspección y vigilancia [1]. La Prueba Saber Pro está compuesta por los módulos de comunicación escrita, razonamiento cuantitativo, lectura crítica, competencias ciudadanas e inglés. Para su respectiva medición, cada uno de estos módulos se califica de 1 a 300 puntos, luego se suman todos los puntajes obtenidos en cada uno de ellos, y se divide entre los 5 componentes para sacar el promedio global [2].

Según algunos autores [3], [4], [5], [6], los estudios que se han realizado hasta el momento, con respecto al análisis de los resultados de las Pruebas Saber Pro, se basan en información procesada mediante un análisis estadístico, donde fundamentalmente se consideran variables y relaciones primarias, sin tener en cuenta las verdaderas interrelaciones que, por lo general, están ocultas y que únicamente se pueden descubrir utilizando un tratamiento más complejo de los datos, el cual es posible con la minería de datos.

La minería de datos en la educación no es un tema nuevo, su estudio y aplicación han sido muy relevantes en los últimos años, se pueden utilizar sus técnicas para explicar y/o predecir cualquier fenómeno dentro del campo educativo [7]. Por ejemplo, utilizando las técnicas de minería de datos, se puede predecir, con un porcentaje muy alto de confiabilidad, la probabilidad de deserción de cualquier estudiante [7], [8], [9], [10]. Las instituciones de educación pueden usar la minería de datos para hacer análisis comprensivos de las características de sus estudiantes, métodos evaluativos que develan procesos exitosos o, por el contrario, detectan fraudes o inconsistencias [9].

En cuanto a la aplicación de la minería de datos en las pruebas Saber Pro, se realizó un estudio en Colombia en el que se buscaba dar respuesta a la pregunta ¿Cuáles son los patrones sociodemográficos, económicos, académicos e institucionales asociados al desempeño académico de los estudiantes en las competencias genéricas de los programas profesionales en las pruebas Saber Pro 2011-2? [3]. En el estudio utilizaron la técnica Clasificación, de minería de datos, basada en árboles de decisión. En los patrones descubiertos se destaca que la acreditación institucional y la modalidad de estudio son dos atributos importantes asociados al desempeño académico de los estudiantes en las competencias genéricas: lectura crítica, composición escrita, razonamiento cuantitativo e inglés de las Pruebas Saber Pro 2011-2. En la bibliografía consultada, no se encontró evidencia sobre la aplicación de la minería de datos para detectar patrones de desempeño en las Pruebas con estudiantes de carreras profesionales afines a la informática, sistemas y computación.

ISSN: 2027-8101 - e-ISSN: 2619-5232. Enero-diciembre 2019 • página 105 de 154

En este artículo se presentan los resultados de un proyecto de investigación que tuvo como objetivo identificar patrones de rendimiento académico en las pruebas Saber Pro 2015-2017 a nivel nacional de los estudiantes de carreras profesionales afines con sistemas, informática, y computación; se emplearon técnicas descriptivas de minería de datos con el fin de generar información de calidad que permita a los directivos de las Instituciones de Educación Superior tomar decisiones efectivas con respecto a la formulación de planes y proyectos conducentes a mejorar la calidad de la educación que se imparte en estos programas.

Metodología

Para el desarrollo se utilizó el proceso CRISP-DM (Cross Industry Standard Process for Data Mining) que proporciona una descripción normalizada del ciclo de vida de un proyecto estándar de análisis de datos, de forma análoga a como se hace en ingeniería de software con los modelos de ciclo de vida de desarrollo de software [11], [12], [13], [14]. El modelo CRISP-DM cubre las fases de un proyecto, sus tareas respectivas y las relaciones entre estas tareas. En este nivel de descripción no es posible identificar todas las relaciones; las relaciones podrían existir entre cualquier tarea de acuerdo con los objetivos, el contexto y el interés del usuario sobre los datos.

La metodología CRISP-DM contempla el proceso de análisis de datos como un proyecto profesional, de este modo establece un contexto mucho más rico que influye en la elaboración de los modelos. Este contexto tiene en cuenta la existencia de un cliente que no es parte del equipo de desarrollo, así como el hecho de que el proyecto no acaba una vez se encuentra el modelo idóneo (ya que después se requiere un despliegue y un mantenimiento), sino que está relacionado con otros proyectos, por lo que es preciso documentarlo de forma exhaustiva para que otros equipos de desarrollo utilicen el conocimiento adquirido y trabajen a partir de él. En la Figura 1 se muestran las fases del proceso CRISP-DM.

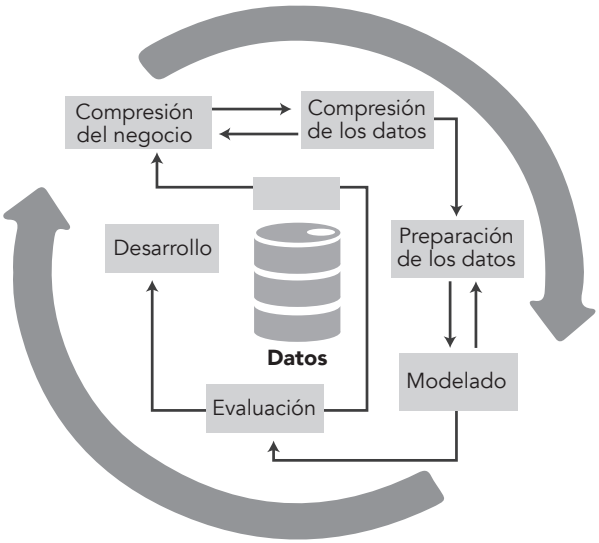

Figura 1. Fases de la Metodología CRISP-DM.
Fuente: Elaboración propia

Fase I: Comprensión del Negocio (*Business Understanding*): En esta fase se profundizó sobre las Pruebas Saber Pro, apoyados en la documentación existente en el repositorio del ICFES y que este nos aportó, se investigó su enfoque legal y su historia para poder entrar en contexto.

Fase II: Comprensión de los Datos (*Data Understanding*): En esta fase se recopiló la base de datos genérica de las Pruebas Saber Pro de los años 2015, 2016 y 2017, se estudiaron los diccionarios de los diferentes años para comprobar a qué hace referencia cada uno de los atributos contenidos en estas bases de datos.

Fase III: Preparación de los Datos (*Data Preparation*): En esta fase se cargaron las bases de datos al gestor, se realizó el proceso de transformación de los datos (ETL), se integraron en un repositorio las bases de datos de los años 2015, 2016 y 2017, este repositorio final quedó limpio y transformado, y sobre él cual se llevó a cabo el proceso de minería de datos.

Fase IV: Modelado (*Modelling*): Se utilizaron técnicas predictivas de minería de datos para construir un modelo óptimo que brinde información útil sobre el rendimiento académico

de los estudiantes de carreras profesionales afines con la informática, los sistemas y la computación.
Fase V: Evaluación (*Evaluation*): Los modelos obtenidos se interpretaron desde el enfoque de las carreras afines a tecnología, se analizaron los atributos útiles para la obtención del conocimiento requerido para los objetivos planteados.

Fase VI: Despliegue (*Deployment*): Se elaboró un reporte final con el cual se presentó el resultado obtenido, junto con las recomendaciones, con base en los patrones de desempeño identificados. Estas recomendaciones sirven de soporte para que las instituciones de educación superior tomen las decisiones estratégicas que ayuden a mejorar la calidad de la educación en los niveles local, regional y nacional.

Resultados

En esta sección se presentan los datos obtenidos después de todo el proceso antes mencionado en la Metodología.

Comprensión del Problema

Las Pruebas Saber Pro son un instrumento estandarizado para la evaluación externa de la calidad de la educación superior. Los lineamientos para el diseño del examen Saber Pro se definieron de acuerdo con la política de formación por competencias del Ministerio de Educación Nacional (MEN), tanto en el nivel universitario como en el nivel tecnológico y técnico profesional, y en su desarrollo han participado las comunidades académicas, asociaciones y redes de facultades y programas. Todos los examinandos son evaluados en las competencias genéricas y en una combinatoria de módulos de competencias específicas.

Los periodos analizados fueron: 2015, 2016 y 2017:
Periodo 2012-2015 (Saber Pro): Durante este periodo, todos los estudiantes fueron evaluados en las competencias genéricas: razonamiento cuantitativo, lectura crítica, comunicación escrita, inglés y competencias ciudadanas. Los resultados de los módulos de competencias genéricas tienen una escala con media 10 y desviación estándar 1.

Periodo 2016 en adelante (Saber Pro): Los estudiantes que actualmente toman el examen de Saber Pro son evaluados en cinco competencias genéricas: razonamiento cuantitativo, lectura crítica, comunicación escrita, inglés y competencias ciudadanas. A pesar de que se evalúan las mismas competencias que en el periodo anterior, a partir del 2016 se produjeron cambios en el diseño de los módulos de estas pruebas y en su calificación que imposibilitan la comparación entre años del periodo anterior con el actual. Los resultados de los módulos de competencias genéricas resultaron en una nueva escala histórica con año base en 2016, con media 150 y desviación estándar 30.

Además de los puntajes en las competencias genéricas, en este período inició la publicación de un puntaje global del examen Saber Pro, construido a partir de un promedio simple entre los puntajes de las cinco competencias genéricas. Por esta razón, este puntaje global también cuenta con media 150 y desviación 30. En la Tabla 1 se describen las competencias genéricas evaluadas en las pruebas Saber Pro, y en la Tabla 2 se muestran las escalas de estas competencias entre los periodos 2015, 2016 y 2017..

Tabla 1. Descripción de las competencias genéricas.

Competencia genérica	Descripción
Comunicación escrita	Evalúa la competencia para comunicar ideas por escrito referidas a un tema dado. Los temas sobre los que yace la escritura son de dominio público, no requieren conocimientos especializados.
Razonamiento cuantitativo	Evalúa competencias relacionadas con las habilidades matemáticas para desempeñarse adecuadamente en contextos cotidianos que involucran información de carácter cuantitativo.

ISSN: 2027-8101 - e-ISSN: 2619-5232. Enero-diciembre 2019 • página 107 de 154

Competencia genérica	Descripción
Lectura crítica	Evalúa las capacidades de entender, interpretar y evaluar textos que pueden encontrarse, tanto en la vida cotidiana como en ámbitos académicos no especializados.
Competencias ciudadanas	Evalúa los conocimientos y habilidades que posibilitan la construcción de marcos de comprensión del entorno, los cuales promueven el ejercicio de la ciudadanía y la coexistencia inclusiva según la Constitución Política
Inglés	Evalúa la competencia para comunicarse efectivamente en inglés. Esta competencia, alineada con el Marco Común Europeo, permite clasificar a los examinados según su nivel de desempeño.

Fuente: Elaboración propia

Tabla 2. Competencias genéricas - Escalas del 2015-2016-2017.

Competencias Genéricas - Escalas del 2015-2016-2017				
Prueba	**Media**	**Desviación Estándar**	**Mínimo**	**Máximo**
Razonamiento cuantitativo	150	30	0	300
Lectura crítica	150	30	0	300
Comunicación escrita	150	30	0	300
Inglés	150	30	0	300
Competencias ciudadanas	150	30	0	300
Puntaje global	150	30	0	300

Fuente: Elaboración propia

Comprensión de los Datos
Análisis del repositorio de datos inicial: En la base de datos de las pruebas Saber Pro están almacenados los resultados obtenidos de estas pruebas, donde podremos encontrar sus resultados, información socioeconómica, entre otros.

Los datos almacenados están organizados en archivos de texto plano con sus respectivos nombres:

- SaberPro_Genéricas_20151.txt,
- SaberPro_Genéricas_20152.txt,
- SaberPro_Genéricas_2016.txt,
- SaberPro_Genéricas_2017.

En la Figura 2 se muestra la cantidad de atributos en cada tabla de la base de datos y la relación entre ellas.

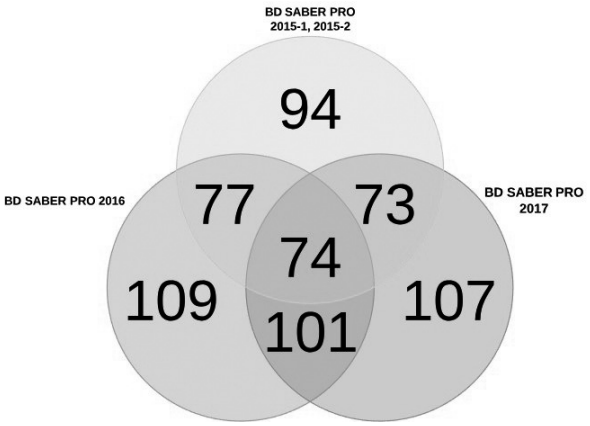

Figura 2. Atributos Pruebas Saber Pro 2015, 2016 y 2017.
Fuente: Elaboración Propia

Por otra parte, en la Tabla 3 se evidencia parte del diccionario de datos que corresponde a los atributos de las Pruebas Saber Pro en los periodos entre 2015 y 2017.

Tabla 3. Diccionario de Atributos Pruebas Saber Pro 2015-2017.

Diccionario de atributos Saber Pro periodo 2015-2019							
	Campo	**2015-1**	**2015-2**	**2016**	**2017**	**Descripción del campo**	**Operaciones de respuesta**
1	Estu_tipo documentio					Tipo de documento	CC - Cédula de ciudadanía CE - Cédula extranjera CR - Certificado registraría PC - Pasaporte colombiano RC - Registro civil de nacimiento TI - Tarjeta de identidad V - Verificar
2	Estu_Nacionalidad					Nacionalidad	Texto
3	Estu_Genero					Genero	F - Femenino M - Masculino
4	Estu_Fecha de nacimiento					Fecha de nacimiento	(DD/MM/AAAA)
5	Estu_Exterior					Indica al estudiante	No Si

ISSN: 2027-8101 - e-ISSN: 2619-5232. Enero-diciembre 2019 • página 109 de 154

Diccionario de atributos Saber Pro periodo 2015-2019							
	Campo	**2015-1**	**2015-2**	**2016**	**2017**	**Descripción del campo**	**Operaciones de respuesta**
6	Periodo					Abierta (ejemplo 2010- 20102)	20173 20172 20163 20162 20154 20153 20152
7	Estu_Consecutivo					Id público del estudiante	Estudiante
8	Estu_Estado civil					Estado civil	Texto
9	Estu_Estudiante					Indica si realiz´p la inscripción pormedio de la institución de Educación Superior (estudiante) o indovidual.	No Si
10	Estu_Pais_Reside					Codigo del pais donde reside actualmente el estudiante	
11	Estu_Tiene Etnia					¿Pertenece usted a un grupo etnico	

Fuente: Elaboración propia

Construcción de un repositorio de datos

Analizado los atributos y los diferentes elementos que contiene estos archivos de las Pruebas Saber Pro genéricas, se procede a hacer un repositorio único que contenga agrupados todos estos datos, y sobre este repositorio se realizan todas las operaciones de limpieza y de transformación.
Esta información se encuentra en formato de texto plano, separados por un símbolo (¬), ya que en este formato no se le puede aplicar ninguna operación de limpieza y de trasformación, se hace una migración hacia un administrador de bases de datos, en este caso se utilizó PostgreSQL, para poder acceder a los datos de una forma más eficiente y poderles aplicar un lenguaje de consulta SQL mucho más potente. En la Tabla 4 se muestra la cantidad de registros existentes de las bases de datos y la cantidad de registros obtenidos después de la integración.

Tabla 4. Cantidad de atributos y de registros existentes por cada base de datos.

Base de datos	Cantidad de atributos	Cantidad de registros
2015	94	371.562
2016	109	242.629
2017	107	245.593
Total	**73**	**859.785**

Fuente: Elaboración propia

Análisis de la calidad de los datos

Se analizó el repositorio saberpro2015_2017, donde se encuentran consolidados todos los datos de las diferentes bases, así podemos observar cómo están agrupados los datos de los 73 atributos y tomar decisiones de limpieza y de transformación para estos datos.

En este análisis identificamos cantidad de datos nulos, no nulos, y una muestra de los diferentes datos que se encuentran por atributo.

Preparación de datos

Dentro del proceso de limpieza y transformación podemos encontrar tareas de procesamiento, como las siguientes:

- Reemplazo de valores ausentes.
- Corrección de valores atípicos (inconsistentes, fuera de rango).
- Generación de atributos derivados.
- Eliminación de atributos que no aportan información.
- Limpieza de filas que no aportan información.

Todas estas operaciones se realizan dentro de la base de datos llamada saberpro_limpio, donde se encuentran todos los datos de la base saberpro2015_2017.

Remplazo de valores ausentes: "Los valores ausentes pueden estar representados en el origen de datos de varias maneras: como valores NULL, como celdas vacías o como un valor artificial".

Como parte de este proceso, se llevaron a cabo una serie de tareas:

- Se rellenaron los datos de manera manual.
- Se utilizó una constante global en función del análisis de cada atributo.
- Se utilizó la moda para los datos categóricos, la cual se utiliza para los valores con más frecuencia.
- Se utilizó la media para los valores numéricos.

Corrección de valores atípicos: Los valores atípicos pueden estar representados por un valor fuera del intervalo, o puede ser que los datos se hayan especificado de forma incorrecta y generado una distorsión en los resultados del análisis. Como parte de este proceso se realizaron las siguientes tareas:

- Se aplicó una generalización de atributos mediante jerarquías de conceptos, entre los que se encuentra la zona geográfica y el tipo de institución.
- Se aplicó discretización, que consiste en dividir el rango de valores contenidos en un conjunto de intervalos, esto se evidencia en los puntajes de las competencias genéricas.

Generación de atributos derivados: Se trata de añadir nuevos atributos, basados en el análisis de los valores existentes, con el fin de obtener un nuevo atributo de evolución. Los atributos derivados los vemos plasmados en los nuevos atributos de las competencias genéricas, y en la creación del atributo de la zona geográfica y el de carreras afines.

Eliminación de atributos que no aportan información: Basándonos en las etapas de este proyecto, se decidió eliminar atributos que no son relevantes ni para la finalidad del estudio ni para el proceso de minería de datos.

Limpieza de filas que no aportan información: En esta fase se evidencia que los datos que van a ser analizados son referentes solo a los estudiantes de Colombia, por lo que se eliminaron registros en los que los estudios o el naciente hayan sido en el exterior. Es importante alcanzar la acreditación de los programas, ya que en las Instituciones Universitarias acreditadas con domicilio en la Región Pacífico, el 94,57 % de los programas alcanzan un puntaje superior a la media nacional en las Pruebas Saber Pro.

Construcción de una vista minable: Después del proceso de limpieza realizado en la base de datos saberpro_limpia, se continuó con el proceso de transformación en la base de datos saberpro_transformado, con el fin de normalizar

ISSN: 2027-8101 - e-ISSN: 2619-5232. Enero-diciembre 2019 • página 111 de 154

datos y generar nuevos atributos necesarios para la creación de una vista minable, de esta base saberpro_transformado se extrajo un repositorio llamado saberpro_final, el cual contiene 859.785 valores y 34 atributos, y que se va a utilizar para extraer las reglas.

Este es el caso particular de los atributos de desempeño en las competencias genéricas, cuyos valores son "Bajo la media" y "Sobre la media". De esta forma, se conoce el nivel de desempeño general sin realizar análisis específicos sobre la puntuación obtenida, lo cual es muy útil en el proceso de identificación de patrones por medio de árboles de clasificación [15].

Modelado

Se construyó un Modelo descriptivo con el cual se logró identificar patrones de rendimiento general de las Pruebas Saber Pro 2015-2017, empleando el Módulo de Asociación de la Herramienta WEKA, donde se seleccionó el Algoritmo Apriori, el cual es uno de los más precisos y usados para los estudios de minería de datos, que permitió generar reglas a partir de atributo desem_global [16].

En la Figura 3 se muestran los resultados generados por la Herramienta WEKA empleando la tarea de Asociación:

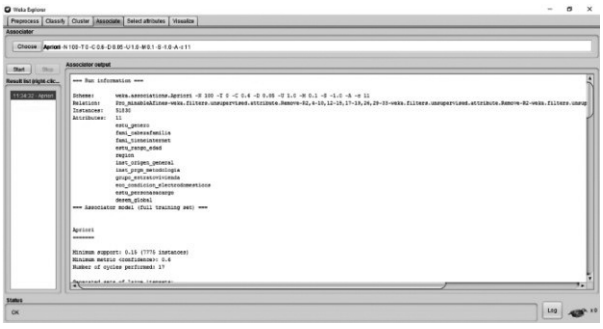

Figura 3. Resultados de Reglas de Asociación con el algoritmo A priori con Weka. Fuente: Elaboración propia

Experimentación de la muestra afines a Sistemas: En esta etapa se debe analizar el desempeño global que obtuvieron los estudiantes afines a carreras de sistemas, se generó un nuevo repositorio solo con los registros donde el atributo afines_sistema sea igual a "SI", con lo cual se obtuvo un total de 51.830, que equivalen al 6 % de la información del repositorio minable inicial para la construcción del modelo, correspondiente al desempeño global de las estudiantes de carreras afines con sistemas.

La parametrización que se usó para el Algoritmo Apriori fue la siguiente: confianza mínima de 0.7, en base a esta condición se encontró que el soporte mínimo que fue de 0.05 para las reglas de asociación generadas, que para este caso se fijó un valor de 100 reglas generadas. La Herramienta WEKA encontró 29 conjuntos de ítems de un solo ítem, 110 conjuntos de ítems con 2 ítems, 163 conjuntos de ítems con 3 ítems, 91 conjuntos de ítems con 4 ítems, 28 conjuntos de ítems con 5 ítems, y 7 conjuntos de ítems con 6 ítems, que cumplían con las condiciones de confianza mínima, lo cual arroja un total de 21 ítems, de los cuales se seleccionaron los más representativos, se ponderaron, y se tomaron solo 10 de ellos que luego fueron interpretados. En la Tabla 5 se muestran las reglas descubiertas.

Tabla 5. Reglas Algoritmo A priori Muestra Afines Sistemas.

#	Reglas	Confianza
1	estu_genero=M fami_cabezafamilia=No fami_tieneinternet=Si region=ANDINA inst_prgm_metodologia=PRESENCIAL eco_condicion_electrodomesticos=BUENA estu_personasacargo=No 11569 ==> desem_global=SOBRE LA MEDIA 7994	conf:(0.69)
2	fami_tieneinternet=Si inst_prgm_metodologia=PRESENCIAL grupo_estratovivienda=ENTRE 3 Y 4 estu_personasacargo=No 11530 ==> desem_global=-SOBRE LA MEDIA 7866	conf:(0.68)

#	Reglas	Confianza
3	inst_prgm_metodologia=PRESENCIAL grupo_estratovivienda=ENTRE 3 Y 4 estu_personasacargo=No 11915 ==> desem_global=SOBRE LA MEDIA 8073	conf:(0.68)
4	fami_tieneinternet=Si region=ANDINA inst_prgm_metodologia=PRESENCIAL eco_condicion_electrodomesticos=BUENA estu_personasacargo=No 15288 ==> desem_global=SOBRE LA MEDIA 10271	conf:(0.67)
5	fami_cabezafamilia=No region=ANDINA inst_prgm_metodologia=PRESENCIAL eco_condicion_electrodomesticos=BUENA estu_personasacargo=No 15312 ==> desem_global=SOBRE LA MEDIA 10231	conf:(0.67)
6	fami_tieneinternet=Si grupo_estratovivienda=ENTRE 3 Y 4 eco_condicion_electrodomesticos=BUENA estu_personasacargo=No 12253 ==> desem_global=SOBRE LA MEDIA 8182	conf:(0.67)
7	estu_genero=M fami_cabezafamilia=No fami_tieneinternet=Si region=ANDINA inst_prgm_metodologia=PRESENCIAL eco_condicion_electrodomesticos=BUENA 13750 ==> desem_global=SOBRE LA MEDIA 9177	conf:(0.67)
8	fami_cabezafamilia=No fami_tieneinternet=Si inst_prgm_metodologia=PRESENCIAL grupo_estratovivienda=ENTRE 3 Y 4 eco_condicion_electrodomesticos=BUENA 12255 ==> desem_global=SOBRE LA MEDIA 8155	conf:(0.67)
9	fami_cabezafamilia=No inst_prgm_metodologia=PRESENCIAL grupo_estratovivienda=ENTRE 3 Y 4 eco_condicion_electrodomesticos=BUENA 12481 ==> desem_global=SOBRE LA MEDIA 8266 conf:(0.66)	conf:(0.66)
10	fami_tieneinternet=Si region=ANDINA inst_prgm_metodologia=PRESENCIAL grupo_estratovivienda=ENTRE 3 Y 4 eco_condicion_electrodomesticos=BUENA 11879 ==> desem_global=SOBRE LA MEDIA 7854	conf:(0.66)

Fuente: Elaboración propia

Modelo de Clustering: Se construyó un Modelo de Clustering para establecer el agrupamiento de los datos de manera independiente y determinar los más predominantes en las diferentes muestras [17].

WEKA cuenta con diferentes algoritmos generados específicamente para manejar los clústeres; para las pruebas realizadas se seleccionó el algoritmo SimpleKMeans, uno de los más precisos y que es usado para los estudios de minería de datos realizados en esta herramienta. Para la generación del clúster y para las pruebas realizadas, no fue necesario seleccionar una determinada variable, debido a que WEKA permite de manera general trabajar con todos los atributos. En la Figura 4 se observan los clústeres generados por la Herramienta WEKA.

ISSN: 2027-8101 - e-ISSN: 2619-5232. Enero-diciembre 2019 • página 113 de 154

Figura 4. Clústeres generados con la Herramienta WEKA. Fuente: Elaboración propia

Gracias a los clústeres fue posible determinar los datos más predominantes en la base de datos, de forma general y por muestras, determinados aleatoriamente por el algoritmo SimpleKMeans, con el fin de encontrar los segmentos de las Pruebas Saber Pro 2015 a 2017. Se seleccionaron todos los datos del repositorio referentes a los estudiantes con carreras afines a sistemas, y se utilizaron los atributos más relevantes en los estudiantes. Los parámetros fueron los siguientes:

- max-candidates =100,
- periodic-pruning= 10000,
- min-density = 2.0,
- canopy T1 = -1.25 canopy T2= -1.0,
- Numclusters = 3,
- distance function = "weka.core. EuclideanDistance-R first-last" Maxiteracions= 500 -NumExecutionSlots = 1,
- Seed = 10.

El modo seleccionado para generar los clústeres fue utilizar todo el conjunto de datos por la opción "use training set" de WEKA, lo cual permitió trabajar con todos los datos, los resultados se evidencian en la Figura 5.
En la Tabla 6 se muestran los clústeres 0,1 y 2 relacionados con las carreras afines con sistemas, informática y computación con su respectivo atributo en las Pruebas Saber Pro.

```
=== Run information ===
Scheme:     weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-
density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation:   Pro_minableAfines-weka.filters.unsupervised.attribute.Remove-R2,4-10,12-15,17-19,26,29-
weka.filters.unsupervised.attribute.Remove-R2-weka.filters.unsupervised.attribute.Remove-R8
Instances:  51830
Attributes:  11
        estu_genero
        fami_cabezafamilia
        fami_tieneinternet
        estu_rango_edad
        region
        inst_origen_general
        inst_prgm_metodologia
        grupo_estratovivienda
        eco_condicion_electrodomesticos
        estu_personasacargo
        desem_global
Test mode:  evaluate on training data
=== Clustering model (full training set) ===
kMeans
======
Number of iterations: 3
Within cluster sum of squared errors: 147171.0
Initial starting points (random):

Cluster 0: M,No,Si,[<23),ANDINA,PRIVADO,PRESENCIAL,'ENTRE 0 Y 2',BUENA,No,'BAJO LA MEDIA'
Cluster 1: M,No,Si,[<23),ANDINA,OFICIAL,PRESENCIAL,'ENTRE 3 Y 4',BUENA,No,'SOBRE LA MEDIA'
Cluster 2: M,No,Si,[23-32),ANDINA,PRIVADO,PRESENCIAL,'ENTRE 3 Y 4',BUENA,Si,'BAJO LA MEDIA'
Missing values globally replaced with mean/mode
```

Figura 5. Resultados de la tarea de clúster con el algoritmo K-means utilizando WEKA. Fuente: Elaboración propia

Tabla 6. Clúster General Carreras afines con Sistemas, Informática y Computación.

Cluster#				
Attribute	**Full Data**	**0**	**1**	**2**
	(51830.0)	**(25050.0)**	**(14618.0)**	**(12162.0)**
estu_genero	m	m	m	m
fami_ cabezafamilia	no	no	no	no
fami_ tieneinternet	si	si	si	si
estu_rango_ edad	[23-32)	[<23)	[23-32)	[23-32)
region	andina	andina	andina	andina
inst_origen_ general	oficial	privado	oficial	privado
inst_prgm_ metodologia	presencial	presencial	presencial	presencial
grupo_ estratovivienda	entre 0 y 2	entre 0 y 2	entre 3 y 4	entre 3 y 4
eco_condicion_ electrodomesticos	buena	buena	buena	buena
estu_ personasacargo	no	no	no	si
desem_global	sobre la media	bajo la media	sobre la media	bajo la media

Fuente: Elaboración propia

En la Figura 6 se muestran la cantidad de estudiantes y el porcentaje con respecto al total de estudiantes por cada clúster generado con la Herramienta WEKA.

ISSN: 2027-8101 - e-ISSN: 2619-5232. Enero-diciembre 2019 • página 115 de 154

Time taken to build model (full training data): 0.57 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 25050 (48%)

1 14618 (28%)

2 12162 (23%)

Figura 6. Instancias por cada clúster.
Fuente: Elaboración propia

En la Figura 7 se muestra visualmente cómo se agrupan los datos en los clústeres 0, 1 y 3, así como la forma en que fueron organizados por el algoritmo, según la distancia HAMMING, la cual determina su similitud por variable categórica.

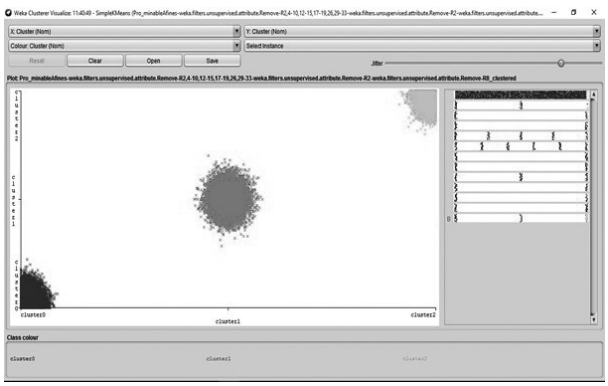

Figura 7. Visualización de los clústeres con Weka.
Fuente: Elaboración propia

De acuerdo con las figuras 7 y 8, el total de registros o instancias del repositorio (51.830) se agrupa en tres clústeres:

Clúster 0: con 25.050 instancias
Clúster 1: con 14.618 instancias
Clúster 2: con 12.162 instancias

En el clúster 0 se agrupan el 48,33 % de todos los estudiantes de programas afines a la computación que presentaron las Pruebas Saber Pro y cuyo desempeño general en estas pruebas está bajo la media. Específicamente, son estudiantes hombres que no son cabeza de familia, que tienen internet, menores de 23 años, de la región Andina, de IES privadas, de un programa presencial, de estrato bajo (1 y 2), con una condición de electrodomésticos buena y sin personas a cargo.

En el clúster 1 se agrupan el 28,2 % de todos los estudiantes de programas afines a la computación que presentaron las Pruebas Saber Pro y cuyo desempeño general en estas pruebas está sobre la media. Específicamente, son estudiantes hombres, que no son cabeza de familia, que tienen internet, cuya edad está entre 23 y 32 años, de la región Andina, de IES oficiales o públicas, de un programa presencial, de estrato medio (3 y 4), con una condición de electrodomésticos buena y sin personas a cargo.

En el clúster 2 se agrupan el 23,46 % de todos los estudiantes de programas afines a la computación que presentaron las Pruebas Saber Pro y cuyo desempeño general en estas pruebas está bajo la media. Específicamente, son estudiantes hombres, que no son cabeza de familia, que tienen internet, cuya edad está entre 23 y 32 años, de la región Andina, de IES privadas, de un programa no presencial, de estrato medio (3 y 4), con una condición de electrodomésticos buena y con personas a cargo.

Interpretación de las reglas

Las reglas más representativas que superan un soporte mínimo de 20 % (de 51.830 registros) y una confianza mínima de 67 % son:

Regla 1. El 69 % de los estudiantes hombres que no son cabeza de familia tienen internet, son de la región Andina, estudian un programa presencial en computación y afines, su condición de electrodomésticos es buena y no tienen personas a cargo, entonces el desempeño general en las Pruebas Saber Pro está sobre la media. El 22,3 % de todos los estudiantes cumplen esta regla.

Regla 2. El 68 % de los estudiantes que tienen internet, estudian un programa presencial en computación y afines, son de estrato medio bajo (1, 2 y 3) y no tienen personas a cargo, entonces el desempeño general en las Pruebas Saber Pro está por debajo de la media. El 22,2 % de todos los estudiantes cumplen esta regla.

Regla 3. El 68 % de los estudiantes que estudian un programa presencial en computación y afines, son de estrato alto (5 y 6) y no tienen personas a cargo, entonces el desempeño general en las Pruebas Saber Pro está sobre la media. El 22,9 % de todos los estudiantes cumplen esta regla.

Regla 4. El 67 % de los estudiantes que tienen internet, son de la región Andina, no estudian un programa presencial en computación y afines, su condición de electrodomésticos es buena y no tienen personas a cargo, entonces el desempeño general en las Pruebas Saber Pro está bajo la media. El 29,5 % de todos los estudiantes cumplen esta regla.

Conclusiones

Es importante alcanzar la acreditación de los programas, ya que el 94,57 % de los programas de las instituciones universitarias acreditadas con domicilio en la Región Pacífico alcanzan un puntaje superior a la media nacional en las Pruebas Saber Pro.

Entre los atributos que forman parte de los patrones descubiertos con las técnicas descriptivas se destacan: el género, el ser o no cabeza de familia, el tener internet, la edad, la región geográfica, el tipo de IES, la presencialidad del programa, el estrato socioeconómico y el tener o no personas a cargo, como factores importantes asociados al buen o bajo desempeño académico de los estudiantes de programas afines a la informática, sistemas y computación en las Pruebas Saber Pro.
Una de las dificultades que se presentó en el desarrollo de la investigación fue la mala calidad de los datos de las bases de datos del ICFES, cuyo proceso de limpieza y transformación consumió la mayor parte de tiempo de la investigación.

Referencias

[1] Instituto Colombiano para la Evaluación de la Educación (Icfes). *Guía de Orientación Saber Pro: Módulo de Competencias Genéricas.* Bogotá D.C., Colombia: Icfes, 2020. Disponible en: https:// www.icfes.gov.co/ documents/ 20143/ 1891934/ Guia+ de+ orientacion+ de+ Modulos+ genericos+ Saber+ Pro-2020.pdf.

[2] Instituto Colombiano para la Evaluación de la Educación (Icfes), *Examen Saber Pro, Módulos de competencias genéricas y específicas. Evaluación de la calidad de la educación superior.* Bogotá D.C., Colombia: Icfes, 2012. Disponible en: http://www.icfes.gov.co/examenes/.../151-saber-pro-modulos-de-competencias.

[3] R. Timarán, I. Hernández, S. Caicedo, A. Hidalgo y J. Alvarado, *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional.* Bogotá: Ediciones Universidad Cooperativa de Colombia, 2016. Disponible en: http://dx.doi.org/10.16925/9789587600490

[4] Instituto Colombiano para la Evaluación de la Educación (Icfes), Saber Pro: Principales resultados en Competencias Genéricas. Santa Marta, Colombia, 2012. Disponible en: www.icfes.gov.co/examenes/.../151-saber-pro-modulos-de-competencias.

[5] L. Zapata, "Factores académicos asociados al bajo rendimiento en inglés en las pruebas ECAES presentadas por los estudiantes de la Facultad de Educación en el año 2009", Trabajo de grado, Fundación Universitaria Luis Amigó, Medellín, Colombia, 2011.

[6] Universidad Nacional de Colombia (Unal). *Análisis de los resultados obtenidos por la Universidad Nacional de Colombia sede Bogotá en las pruebas Saber Pro 2011-2.* Bogotá: Universidad Nacional de Colombia, 2012. Disponible en: www.unal.edu.co/diracad/evaluacion/SaberPro_2012/analisis_de_resultados.pdf.

ISSN: 2027-8101 - e-ISSN: 2619-5232. Enero-diciembre 2019 • página 117 de 154

[7] R. Timarán, A. Calderón, y J. Jiménez, "Aplicación de la minería de datos en la extracción de perfiles de deserción estudiantil", *Ventana Informática,* núm. 28, 2013. Disponible en: http://revistasum.umanizales.edu.co/ojs/index.php/ventanainformatica/article/view/181

[8] S. Valero, *Aplicación de técnicas de minería de datos para predecir deserción.* Puebla, México: Universidad Tecnológica de Izúcar de Matamoros, 2009. Disponible en: http://www.utim.edu.mx/~svalero/docs/MineriaDesercion.pdf

[9] S. Valero, A. Salvador y M. García, *Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos.* Puebla, México: Universidad Tecnológica de Izúcar de Matamoros, 2010. Disponible en: www.utim.edu.mx/~svalero/docs/e1.pdf

[10] K. Ordóñez y P. Valdiviezo, "Aplicación de técnicas de minería de datos para predecir la deserción de los estudiantes de primer ciclo de la Modalidad Abierta y a Distancia de la UTPL", Trabajo de grado, Universidad Técnica Particular de Loja, Ecuador, 2013. Disponible en: http://dspace.utpl.edu.ec/bitstream/123456789/7897/1/Ordonez%20Brice%C3%B1o%20Karla-%20Informatica.pdf

[11] A. Azevedo & M. Santos, "KDD, SEMMA and CRISP-DM: a parallel overview", in *Proceedings of IADIS European Conference on Data Mining* Amsterdam, Netherlands, 2008, pp. 182-185.

[12] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, & R. Wirth, *CRISP-DM 1.0: Step-by-step data mining guide.* CRISP-DM consortium: NCR Systems Engineering Copenhagen (USA and Denmark), DaimlerChrysler AG (Germany), SPSS Inc. (USA), and OHRA Verzekeringen en Bank Groep B.V. (The Netherlands), 2000.

[13] J. Hernández, M. Ramírez, y C. Ferri, *Introducción a la Minería de Datos.* Madrid: Editorial Pearson Educación S. A., 2005. Disponible en: http://dspace.ucbscz.edu.bo/dspace/handle/123456789/526

[14] J. Villena. *CRISP-DM: La metodología para poner orden en los proyectos de Data Science,* 2016. Disponible en: https://data.sngular.team/es/art/25/crisp-dm-la-metodologia-para-poner-orden-en-los-proyectos-de-data-science

[15] Vidhya Analytics. Weka GUI way to learn Machine Learning. 2018. Disponible en: https://www.analyticsvidhya.com/learning-paths-data-science-business-analytics-business-intelligence-big-data/weka-gui-learn-machine-learning/

[16] M. Cano y R. Robles, "Factores asociados al rendimiento académico en estudiantes universitarios", *Revista Mexicana de Orientación Educativa,* vol. 15, núm. 35, pp. 1-25. Disponible en: https://doi.org/10.31206/rmdo072018.

[17] J. Molina y J. García, *Técnicas de análisis de datos.* 2006. Disponible en: http://www.giaa.inf.uc3m.es/docencia/II/ADatos/apuntesAD.pdf