Encuentro Internacional de Educación en Ingeniería ACOFI
2º Congreso Latinoamericano de Ingeniería
2019 10 al 13 de septiembre - Cartagena de Indias, Colombia
RETOS EN LA FORMACIÓN DE INGENIEROS EN LA ERA DIGITAL

ANÁLISIS EXPLORATORIO DE DATOS A UNA BASE DE DATOS DE LA BIBLIOTECA DE LA UNIVERSIDAD DE LA SALLE

Paula Katherinne Mila Deaz, Ediwn Iván Gómez Oliveros, Yamile Adriana Jaime Arias

Universidad de La Salle
Bogotá, Colombia

Resumen

Mediante el Análisis Exploratorio de Datos (EDA) y la aplicación de técnicas de Minería de Datos se realizó un estudio en la biblioteca de la Universidad de La Salle (ULSA) para determinar el comportamiento de los miembros de la facultad de ingeniería. El conjunto de datos utilizado en la investigación se obtuvo de la base de datos que registra el uso de elementos como: libros, casilleros, revistas, tesis, entre otros. El proceso de Descubrimiento de Conocimiento de Bases de Datos (KDD) inició con la limpieza a la base de datos eliminando atributos que pudieran llegar a causar ruido, después, se realizó un análisis estadístico y la representación gráfica correspondiente. Una vez hecho dicho análisis, se encontraron tendencias en libros solicitados, áreas de investigación frecuentes entre los estudiantes de Ingeniería, fechas de solicitud del material y relación entre el uso de la biblioteca y el año de ingreso de los estudiantes.

El Estudio se fortaleció con la aplicación de técnicas de minería de datos, lo que permitió entender la dinámica entre los estudiantes de la facultad de ingeniería con sus asignaturas. Como resultado se logra obtenga información para desarrollar nuevas estrategias e incentivos hacia los estudiantes para el uso de los elementos y servicios prestados por la biblioteca.

Palabras clave: minería de datos; visualización; arboles de decisión; biblioteca

Abstract

Through Exploratory Data Analysis and in conjunction with Data Mining techniques, a study was made in the library of the La Salle University to determine the behavior of the members of the faculty

of engineering. The data set used in the research refers to the use of the elements belonging to the Library, such as Books, lockers, magazines, theses, among others. The process of Discovery of Knowledge of Databases began with the cleaning of the database eliminating attributes that could cause noise, then a statistical analysis and the corresponding graphic representation was made. Once this analysis was done, trends were found in the books requested, the relationship between the use of the library and the year of entry of students, the most frequent research areas among engineering students and the customary application dates for material in library.

The study was strengthened with the application of data mining techniques, which allowed to understand the dynamics among the students of the faculty of engineering with their subjects. As a result, information is obtained to develop new strategies and incentives for students to use the elements and services provided by the library.

***Keywords**: data mining; visualization; decision trees; library*

1. Introducción

Actualmente, las organizaciones y/o personas naturales vinculadas a modelos de negocios que generan grandes cantidades de datos, presentan tanto la oportunidad como la necesidad de realizar su análisis para convertirlos en información relevante que luego apoye una posterior toma de decisiones que favorezca el rendimiento del entorno o negocio en el cual se encuentren. Por ello, ha surgido la minería de datos, la cual fusiona técnicas tradicionales de análisis de datos con sofisticados algoritmos de agrupamiento y predicción posibles de aplicar a grandes cantidades de datos (*Big Data*) con el fin de descubrir información útil que no se presenta de una manera explícita (Steinbach, Tan, & Kumar, 2005)

En esta investigación, mediante uso de técnicas de minería de datos, visualización y resúmenes estadísticos, se caracterizaron usuarios de la Biblioteca pertenecientes a la Facultad de Ingeniería en el año 2017. En particular, a los programas de Ingeniería Industrial, Civil y Ambiental, con el fin de encontrar comportamientos y patrones. Para lo cual se recopilaron datos relacionados con temáticas de libros solicitados, fechas de préstamos, año y periodos de ingreso a la ULSA de los usuarios, junto con que otros tipos de elementos que solicitan. Con la unificación de todas las variables, se obtuvo una la base de datos para el estudio a la cual se le efectuó una limpieza en cuanto a estructura y contenido, posteriormente se estableció el grado de incidencia de cada atributo y se realizó el AED, mediante el proceso que se presenta en la figura 1.

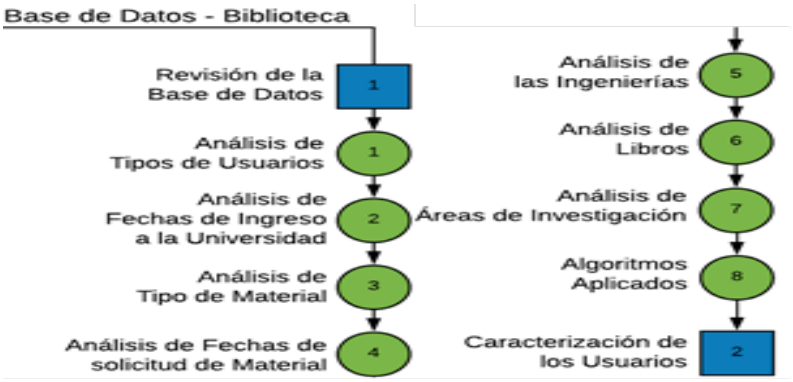

Fig. 1 Diagrama de operaciones

La orientación de la investigación se encaminó a identificar si los estudiantes relacionan las temáticas de estudio con los espacios académicos que toman y a verificar si requieren el material de consulta para el interior de la Universidad o para realizar la investigación en casa, para lo que se usaron técnicas de predicción como arboles de decisión, algoritmos basados en reglas y técnicas bayesianas.

2. Marco referencial

En 1970 el estadístico John Tukey creo el área conocida como Análisis Exploratorio de Datos o EDA, la cual se enfoca en temáticas como: resumen estadístico o analítica descriptiva, basándose en el planteamiento de una hipótesis y la visualización para la detección de patrones (Steinbach, Tan, & Kumar, 2005). El EDA está fundamentado en generar conocimiento a partir de unas bases de datos, detectar información implícita, extraer variables relevantes para la información predicha, y probar hipótesis originadas a partir de supuestos (NIST, 2018).

A su vez, el resumen estadístico o analítica descriptiva presenta las características que definen un conjunto de datos o variables específicas. Estos resúmenes, en cuanto a variables o atributos singulares, pueden incluir medidas de tendencia central, dispersión, posición (Fernandez & Dias, 2001). En cuanto a conjuntos de datos se utiliza correlación y covarianza entre los atributos. (Steinbach, Tan, & Kumar, 2005).

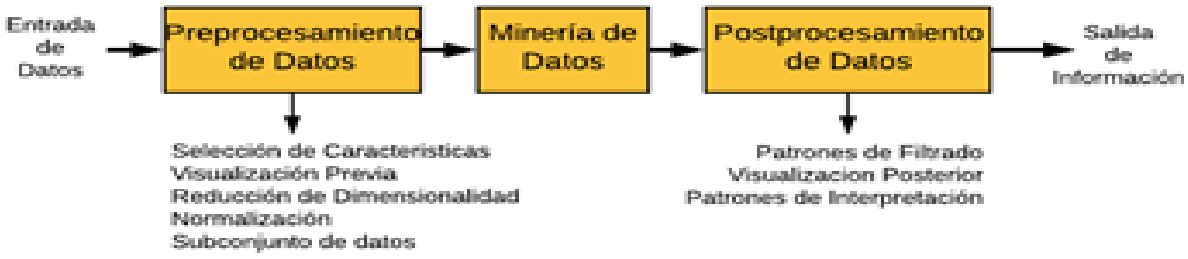

Fig. 2 Proceso de descubrimiento de conocimiento (KDD)

Por lo demás, la visualización es una técnica, dentro de la ciencia de datos, que permite presentar gran cantidad de información entendible para todo tipo de usuario. Además, de ser un medio de

detección de patrones de interés dentro del objetivo del procesamiento, permiten identificar característica y comportamientos de variables en la base de datos. Todas estas áreas forman parte de la Minería de Datos, cuyo el fin último es llegar al *Knowledge Discovery in Databases*, este enfoque consiste en la transformación de datos en información útil para la toma de decisiones, cuyo proceso presenta en la figura 2. (Steinbach, Tan, & Kumar, 2005).

3. Análisis y procesamiento

La base de datos inicial contaba con 12 atributos y 47.566 registros. El análisis de cada atributo se presenta en la tabla 1.

TABLA 1: DESCRIPCIÓN DE LOS ATRIBUTOS

#	Atributo	T. dato	Descripción
1	Ing.	Nominal	Programa de Ingeniería
2	No. De cuenta	Nominal	ID que identifica al usuario.
3	Año ingreso	Ratio	Año en el que ingresa el usuario a la universidad
4	Periodo de ingreso	Ordinal	Semestre en el que el usuario ingresa a la universidad
5	Código de barras	Nominal	Código de barras del artículo perteneciente a la biblioteca
6	Número de clasificación	Nominal	Cadena de caracteres para la clasificación de libros.
7	Fecha del préstamo	Ratio	Fecha en el que se realizó el préstamo por el usuario.
8	Hora del préstamo	Intervalo	Hora en el que el usuario realiza el préstamo.
9	Fecha de devolución	Ratio	Fecha el que el usuario regresa el artículo a la biblioteca
10	Perfil	Nominal	Define el perfil del usuario
11	Categoría usuario	Nominal	Tipo de usuario que realiza el préstamo en la biblioteca.
12	Tipo de material	Nominal	Material de consulta y casilleros

En la revisión de las características de los datos, se identificó que el atributo 10, "*Perfil*", no es relevante ya que el atributo11, "*Categoría de usuario*" brinda, de forma más precisa, el tipo de usuario con valores como: profesor, estudiante de pregrado o estudiante de posgrado. Además, los estudiantes de posgrado y los profesores no cuentan con el atributo 3 y 4, "*Año de ingreso*" y "*Periodo de ingreso*" respectivamente. El atributo 5 "*Código de barras*" representa redundancia, ya que el atributo 6 "*Numero de clasificación*" identifica la categoría del material disponible.

Tipos de usuarios: La figura 3 presenta los perfiles de usuario, mostrando que la mayoría son estudiantes de pregrado. La figura 4 presenta la categoría del usuario, en donde igualmente se e observa una mayoría de estudiantes de pregrado. Además, para la categoría profesores, se encontró que, de los 24 profesores, 10 de ellos realizan un posgrado en la ULSA. Dado que estudiantes de posgrado y profesores son muy pocos respecto al nivel de pregrado, para análisis posteriores, se omiten estas dos categorías.

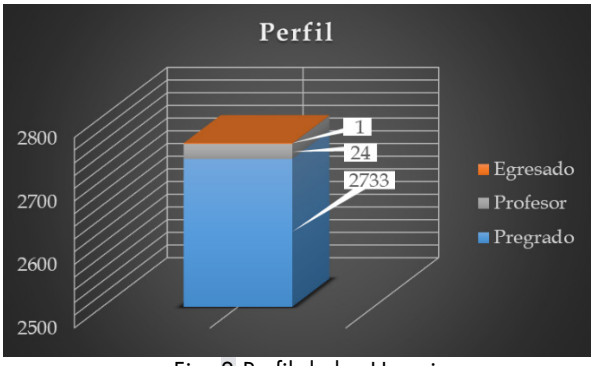

Fig. 3 Perfil de los Usuarios

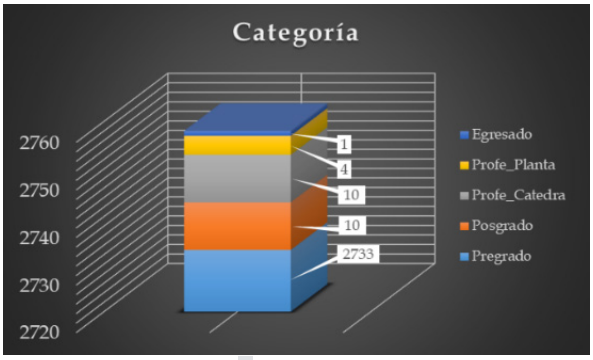

Fig. 4 Categoría de los usuarios

Fechas de ingreso a la Universidad: La figura 5, presenta el año de ingreso de los usuarios; el mayor valor corresponde a estudiantes de pregrado ingresados en el año 2015, seguido del 2016 y 2017, esto muestra una relación directa con respecto al programa del gobierno "Ser Pilo Paga", SPP, ya que fue precisamente en el año 2015 cuando inició este programa, siendo la ULSA, uno de los centros de educación superior con mayor acogida a estudiantes de este programa, ubicando en el segundo periodo del programa SPP, a la ULSA en el segundo lugar entre las Instituciones de Educación Superior con mayor número de estudiantes en la carrera de Ingeniería Industrial con 913 becarios (Ministerio de educacion nacional, 2016). Los datos mostraron además estudiantes con fecha de ingreso del 2001 al 2011 lo cual se entendió como error debido a la duración de las carreras por lo que se procedió a eliminan esos registros.

Como se observa en la figura 6, el periodo de ingreso de los estudiantes corresponde con un 67%al primer semestre de cada año. Esto se debe a dos razones: la primera es que existe mayor comunidad estudiantil escolar en el calendario A, y la segunda, corroborando lo anterior, es que el programa SPP cobija a los estudiantes de dicho calendario.

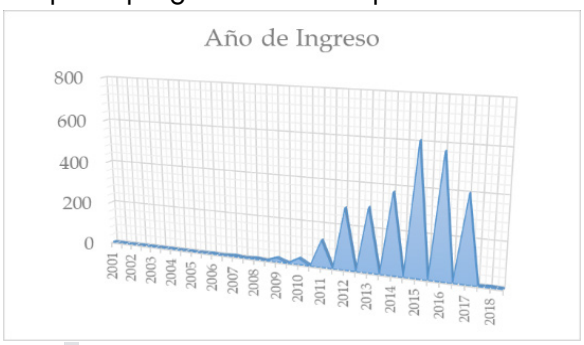

Fig. 5 Año de ingreso de estudiantes de pregrado a la universidad

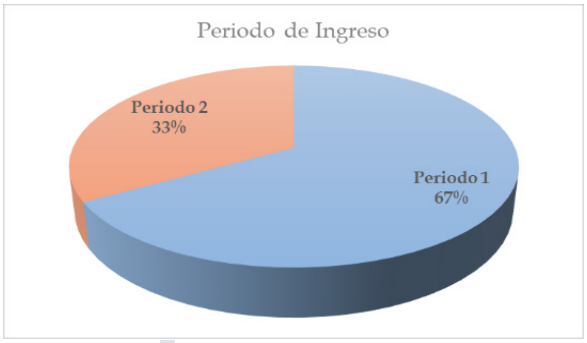

Fig. 6 Periodo de ingreso de estudiantes a la universidad

Tipo de material: la figura 7, muestra la distribución de préstamos según el material solicitado, siendo los libros generales y casilleros los materiales con mayor uso. Esto hace que se genere una duda para futuras investigaciones *¿Por qué las tesis, mapas, archivos, DVD, Cd-Rom, revistas y libros de reserva no se utilizan?* Con base en esta pregunta se retiran esas categorías pues su porcentaje de participación es muy bajo. Adicionalmente, como el uso del casillero es obligatorio para el ingreso a la biblioteca también se elimina esa categoría.

Fecha de solicitud del material: al realizar el análisis a la distribución de fechas de uso de la biblioteca, se evidenció una preferencia por los periodos de febrero a mayo y agosto a noviembre, lo que concuerdan con el desarrollan los semestres académicos.

Ingenierías: Como se puede observar en la figura 8, la ingeniería civil y ambiental presentan mayor uso de los materiales de la biblioteca que ingeniería industrial, esto debido a que estas carreras tienen más tiempo en la ULSA ya que ingeniería industrial se ofrece desde el año 2011.

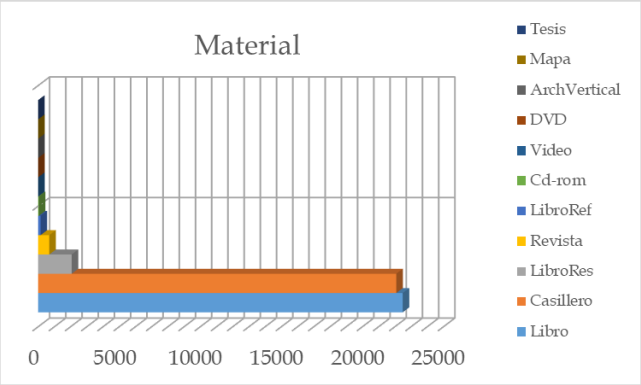

Fig. 7 Tipo de material

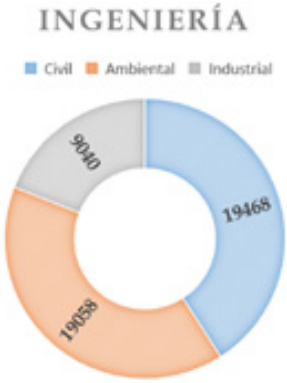

Fig. 8 Programas de Ingeniería

Libros generales: la tabla 2 presenta los 10 libros de mayor consultaron durante el 2017. Dichos libros pertenecen a las áreas de física, cálculo y mecánica de fluidos; Dentro de esta lista también encontramos dos títulos pertenecientes al programa *"Canon de los 100 libros"*, el cual promueva la literatura entre la comunidad estudiantil.

TABLA 2: LIBROS MAS SOLICITADOS EN EL 2017

#	Libro	Frecuencia	Canon de los 100 libros
1	Física universitaria	874	
2	Cálculo de una variable: trascendentes tempranas	872	
3	Mecánica de fluidos	707	
4	La vida secreta de los números: cómo piensan y trabajan los matemáticos	397	*
5	Termodinámica	342	
6	Probabilidad y estadística para ingeniería y ciencias	280	
7	Pre cálculo: matemáticas para el calculo	259	
8	Voces del Planeta	232	*
9	Reglamento colombiano de construcción sismo resistente	226	
10	Hidráulica de canales abiertos	207	

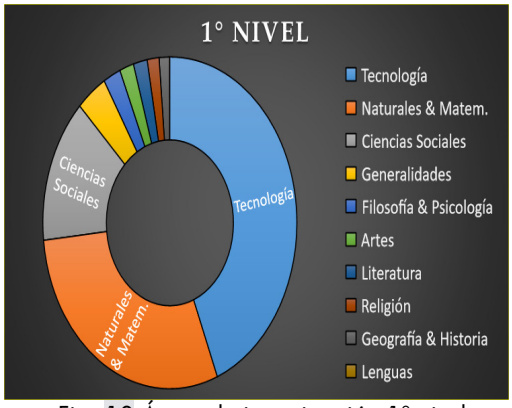

Fig. 10 Áreas de investigación 1° nivel

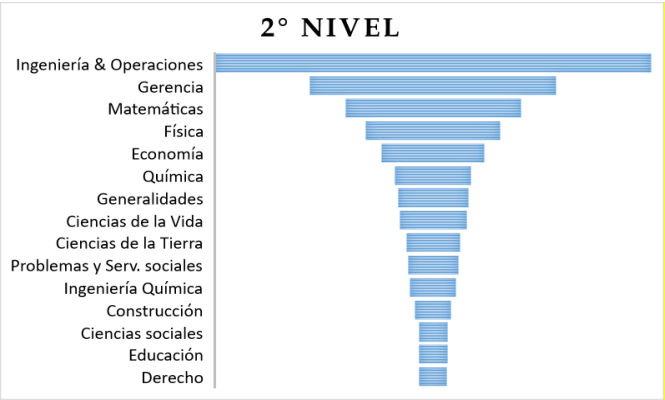

Fig. 11 Áreas de investigación 2° nivel

Áreas de Investigación: la ULSA, maneja un *Sistema de Clasificación Decimal Dewey* , el cual se encuentra estructurado en tres niveles, de sumario general a sumario especifico, con clasificación jerárquica (Derwey, 2000). En la figura 10, se encuentran las frecuencias de uso de libros clasificados en primer nivel y se observa que las áreas de tecnología, naturales, matemáticas y ciencias sociales son de los de mayor uso de manera conjunta por todos los estudiantes. Para el siguiente nivel (2), se observa que las áreas de mayor uso son Ingeniería y Operaciones, Gerencia y Matemáticas. En el nivel inferior se encuentra la clasificación por libros en su forma o estado más específico. Se pudo determinar que dentro de las tres ramas de mayor frecuencia cada programa tiene presencia, siendo Administración, Ingeniería Sanitaría e Ingeniería Civil, áreas relacionadas respectivamente con Ingeniería Industrial, Ambiental y Civil respectivamente.

4. Uso minería de datos

Con el objetivo de desarrollar un estudio con más profundo se construye una nueva base de datos, resultado de los análisis de variables realizados anteriormente. Los atributos incluidos se describen en la tabla 3, obteniendo una base de datos de nueve atributos y 25.470 registros, lo que redujo en un 40% el volumen de datos y es en este conjunto en el cual se implementaron técnicas de clasificación propias de la minería de datos.

TABLA 3: BASE DE DATOS PARA PREDICCIÓN

#	Atributo	Descripción	
1	Ingeniería	Programa de Ingeniería.	
2	AñoIngreso	Año en el que ingresa el usuario a la universidad.	
3	PeriodoIngreso	Semestre en el que el usuario ingresa a la universidad.	
4	DiaPrestamo	Día de la semana en que solicita el material de consulta.	
5	MesPrestamo	Mes en que solicitó el material de consulta.	
6	IntervaloHora	Hora en que solicitó el material de consulta.	
7	TipoMaterial	Clasificación del material de consulta	
8	Tema	Área de investigación	
9	Clase	Se refiere a si el estudiante utiliza el material de consulta en la ULSA o se lo lleva a su casa	
		V	Universidad
		F	Casa

Algoritmos Seleccionados: la clasificación incluye algoritmos basados en árboles de decisión, generación de reglas y métodos bayesianos. Para este estudio los algoritmos utilizados se presentan en la tabla 4.

TABLA 4: ALGORITMOS APLICADOS AL CONJUNTO DE LA BIBLIOTECA

Arboles de decisión	**J48**	Clasificadores basados en reglas	**DecisionTable**	Bayesianos	
	DecisionStump		OneR-B		Naive-Bayes
	HoeffdingTree		Part-M		MultinomialTest

Métricas de Rendimiento: Una vez aplicados los algoritmos seleccionados, haciendo uso de la herramienta Weka, se analizan las matrices de confusión para cada algoritmo y las métricas de rendimiento respectivas, que incluyen: precisión, *recall*, curva ROC, medida de exactitud (Accuracy) y tasa de error. El rendimiento general de cada algoritmo se calcula sobre la clase, teniendo en cuenta que los verdaderos positivos corresponden a clasificar adecuadamente la clase definida, con lo cual a menos falsos positivos la medida de precisión aumenta. A su vez, el valor de *recall* aumenta cuando hay pocos falsos negativos.

5. Resultados

Algoritmos y Métricas: Una vez realizados experimentos, con diferentes algoritmos de árboles de decisión, clasificación basados en reglas y métodos bayesianos se seleccionó el que produjo mejores resultados. La tabla 5 presenta el resumen de las siglas utilizadas para posteriormente explicar las métricas obtenidas por cada algoritmo, las cuales se presentan en las tablas 6, 7 y 8 de acuerdo los resultados arrojadas por la herramienta Weka.

TABLA 5: RESUMEN DE SIGLAS

ALG	Algoritmo	**RC**	Recall	**C**	CLASE	**ER**	Error Rate
PC	Precisión	**RA**	ROC Área	**ACC**	Accuracy		

TABLA 6: MÉTRICAS PARA ÁRBOLES DE DECISIÓN

ALG	PC	RC	RA	C	ACC	ER
J48	0,711	0,41	0,756	V	75,8498	24,1502
	0,769	0,922	0,756	F		
Decisión Stump	0,978	0,083	0,538	V	70,6870	29,3130
	0,699	0,999	0,538	F		
Hoeffding Tree	0,642	0,438	0,761	V	74,2739	25,7261
	0,771	0,886	0,761	F		

TABLA 7: MÉTRICAS PARA CLASIFICACIÓN CON REGLA

ALG	PC	RC	RA	C	ACC	ER
Decision Table	0,621	0,468	0,763	V	73,8948	26,1052
	0,776	0,866	0,763	F		
One R-B	0,97	0,102	0,55	V	71,230	29,000
	0,703	0,999	0,55	F		
Part-M	0,651	0,529	0,783	V	75,918	24,082
	0,797	0,867	0,783	F		

TABLA 8: MÉTRICAS PARA LOS MÉTODOS BAYESIANOS

ALG	PC	RC	RA	C	ACC	ER
Naive-Bayes	0,663	0,364	0,769	V	73,811	26,189
	0,754	0,913	0,769	F		
Multinomial Test	-	0	0,5	V	68,0897	31,9103
	0,681	1	0,5	F		
Updateable	0,663	0,364	0,769	V	73,811	26,189
	0,754	0,913	0,769	F		
Bayes Net	0,662	0,365	0,769	V	73,7871	26,2129
	0,754	0,913	0,769	F		

Selección de mejores resultados: Basados en las tablas 6, 7 y 8, los algoritmos seleccionados son el J48, part-M Y Naive-Bayes, que corresponden a árboles de decisión, clasificación basada en reglas y métodos bayesianos respectivamente. A su vez, se analizan los perfiles de los estudiantes que se llevan el material a sus casas dependiendo del intervalo de hora, el tema estudiado, entre algunas otras características.

El árbol de decisión generado a partir del algoritmo J48 tiene en cuenta principalmente el tema de búsqueda relacionado con el intervalo de horas en que los libros salen de la biblioteca. Entre los hallazgos se encuentra:

- Física: Cuando el usuario solicita el material entre las 6 y 8 a.m., en más de un 78% los estudiantes regresan el material de consulta el mismo día. Por otra parte, si el usuario solicita el material después de las 10 am. en más de un 79% de las veces, lo lleva para la casa.
- Administración General, Ciencias Sociales y Matemáticas si se solicitan en las horas de la tarde, son llevados a casa.
- Ingeniería Sanitaría, Literatura y Retórica: con más de un 79% de precisión, independientemente de la hora del día, estos libros se llevan para la casa.

Por su parte, el clasificador basado en reglas usando el algoritmo part-M aporta información sobre el tipo de perfiles que extraen libros dependiendo de su carrera, tema consultado y hora de solicitud. Entre las reglas generadas se encuentran las siguientes:

- Los estudiantes de Ingeniería Civil entre las 10:00 am y 12:00m acuden a consultar matemáticas. Además, si ingresaron a la ULSA en el año 2016 tienen una alta tendencia a sacar los libros de la biblioteca. Esta regla arrojo un 80% de precisión.
- En el programa de Ingeniería Ambiental, se consultan de manera significativa el área de ciencias de la tierra entre las 12:00m hasta las 2:00pm; estos libros son solicitados para la casa en un 89%.
- Los temas de administración general son los más consultados por la Ingeniería Industrial; además, si son solicitados entre 4:00 y 6:00 PM son llevados para consultar en casa.

Finalmente, basados en el método bayesiano, se obtienen resultados más concretos según la carrera, el año y el periodo de ingreso, el mes, hora, tipo de material y tema consultado, los cuales se muestran en la tabla 9.

TABLA 9: RESULTADOS MÉTODO BAYESIANO

Atributo		Consultados Casa	Consulta en Universidad	Atributo		Consultados Casa	Consulta en Universidad
Ingeniería	ndustrial	1302	3285	**Préstamo**	Febrero	1198	2456
	Civil	3669	6585		marzo	1337	2245
	mbiental	3030	7199		abril	958	1721
Año de ingreso	2011	183	459		agosto	1142	2542
	2012	523	1153		eptiembre	922	2259
	2013	792	1517		octubre	981	2251
	2014	1365	2302		oviembre	421	1219
	2015	2171	5334	**Tema**	Física	1487	2987
	2016	1721	3794		Matemáticas	1562	2881
	2017	1140	2267				

En resumen, los estudiantes de Ingeniería Civil son quienes más consultando libros y quienes además solicitan más libros tanto para ser consultados dentro de la ULSA y en casa. Se observa que en el transcurso de los años hay un crecimiento del uso de los libros que puede ser debido al ingreso de estudiantes del programa SPP, lo que ha generado una cultura de mayor uso de la biblioteca. Por otra parte, los temas más solicitados deben ser tenidos en cuenta para garantizar que todo usuario tenga disponible el material que requiere. El estudio logró determinar periodos de préstamos, incluso definir horas de entregas de libros a diario teniendo en cuenta el flujo correspondiente al tema y el periodo del año. En un trabajo futuro, sería posible obtener resultados interesantes con datos de solicitudes realizadas en años diferentes y así generar perfiles más precisos de acuerdo con la carrera y épocas del año.

6. Referencias

Libros

- Derwey, M. (2000). Sistema de clasificacion decimal de Dewey Ed 21. Bogotá.
- Ho yu, C. (2010). Esplotatory data analysis in the context of data minin and resampling. En *International Journal of Psychological Research* (págs. 9-22). Arizona: Journal of Psychological Research.
- Steinbach, M., Tan, P.-N., & Kumar, V. (2005). *INTRODUCTION TO DATA MINING.* Boston: Pearson.

Fuentes electrónicas

- Fernandez, P., & Dias, P. (2001). *Fisterra.* Obtenido de Estadistica descriptiva de los datos: https://www.fisterra.com/mbe/investiga/10descriptiva/10descriptiva.asp#estadistica.
- NIST. (2018). *ENGINEERING STATISTICSHANDBOOK*. Obtenido de NITS: https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm

Noticias

- *Ministerio de educacion nacional*. (2016). Obtenido de Ser pilo paga 2 en cifras.

Sobre los Autores

- **Paula Katherinne Mila Deaz**: Estudiante de Ingeniería Industrial. Pmila54@unisalle.edu.co
- **Ediwn Iván Gómez Oliveros**: Estudiante de Ingeniería Industrial. Egomez61@unisalle.edu.co
- **Yamile Adriana Jaime Arias**: M.Sc. e Ingeniera de Sistemas y Computación. Profesora titular. Yajaime@unisalle.edu.co

---