ISSN 1870-4069

Reducción de la dimensión de registros de evaluaciones académicas aplicando el algoritmo K-means

Victoria-Elizabeth Padilla-Morales[1], Saturnino Job Morales Escobar[1], Maricela Quintana López[1], José Martín Flores Albino[1], Oscar Herrera Alcántara[1,2]

[1] Centro Universitario UAEM Valle de México, Atizpán de Zarzagoza, México

[2] Universidad Autónoma Metropolitana, Departamento de Sistemas, Azcapotzalco, México

vpadillam002@alumno.uaemex.mx, {sjmoralese, mquintanal, jmfloresa, oherreraa}@uaemex.mx, oha@azc.uam.mx

Resumen. En un ambiente educativo existe una gran cantidad de datos que pueden ser analizados y utilizados en el proceso de la toma de decisiones. En la actualidad, debido al tamaño de su dimensión, los datos tienden a ser más complejos que los datos convencionales y requieren una reducción de su dimensión. La Minería de Datos Educacional permite utilizar técnicas de Minería de Datos para analizar información académica con el fin de identificar patrones que no son evidentes. Este artículo presenta resultados obtenidos en una investigación de un caso de estudio sobre el rendimiento académico de alumnos de Ingeniería del Centro Universitario UAEM Valle de México. En el análisis de los datos se utiliza el algoritmo K-means, el software WEKA y R Studio. Se propone utilizar el agrupamiento para reducir la dimensión de las variables académicas en función de los registros de las calificaciones obtenidas durante los últimos períodos cursados, luego se trabajará con una medida promedio para predecir el rendimiento académico de un alumno. Se utiliza R Studio para contrastar los grupos obtenidos por WEKA.

Palabras clave: minería de datos educacional, reducción de la dimensión, agrupamientos, K-means.

Dimension Reduction of Academic Evaluation Records by Applying the K-Means Algorithm

Abstract. In an educational environment there is a huge data quantity, this data can be analyzed, and it can be used in decision making process. Nowadays data tends to be more complex due to the size than conventional data and need dimension reduction. Educational Data Mining lets using Data Mining techniques for analyzing academic information in order to identify patterns that are not evident. This article presents results obtained in a research of a case of study where regard the academic performance of undergraduate students of Engineering of the Centro Universitario UAEM Valle de México. In the data analysis is used the K-means algorithm, WEKA and R Studio. We propose the use of Clustering to reduce the dimension of academic variables based on their grade registers getting during last periods then we work with some average measure of in order to predict the academic performance of a student. It is used R Studio for contrast the Clusteres obtained by WEKA.

Keywords: educational data mining, reduce dimension, clustering, K-Means.

1. Introducción

Vivimos en la llamada era de la información, en un mundo donde una gran cantidad de datos se colectan y almacenan diariamente [1]. De hecho, es impresionante la cantidad de datos con los que convive el ser humano día a día, y conforme avanza el tiempo, esta cantidad crece cada vez más y de manera muy rápida, llegando a un crecimiento exponencial. Lo anterior hace prácticamente imposible que los seres humanos procesen estos datos manualmente en búsqueda de información valiosa. Esta es una de las razones por la cual la Minería de Datos DM (por sus siglas en inglés, Data Mining) [2], también llamada descubrimiento de conocimiento en bases de datos KDD (por sus siglas en inglés, Knowledge Discovery in Databases), es un área en auge que ofrece técnicas y herramientas para diversos propósitos y cuya aplicación en ámbitos como el de la mercadotecnia, el aprendizaje automático (machine learning), las grandes bases de datos, el reconocimiento de patrones y la visualización por computadora [3], ha tenido buenos resultados.

Los procesos principales de la DM son: la extracción de información mediante relaciones de similitud entre los datos y la búsqueda de patrones o agrupamiento de datos aparentemente inconexos. La DM ha permitido el análisis de datos complejos de manera automática con el objetivo de entender su comportamiento y de esta manera apoyar en la mejora del proceso de la toma de decisiones. Lo anterior es fundamental en cualquier organización con el fin de obtener resultados que funcionen para optimizar y potenciar sus actividades.

Entre los problemas fundamentales que se abordan con la DM, está el agrupamiento o clustering, el cual se puede realizar con diversas técnicas y algoritmos [2], pero que se basan en el mismo principio "los objetos que están en un grupo son más parecidos (similares) entre ellos, que a objetos pertenecientes a otros grupos".

En años recientes, la aplicación de la DM se ha incrementado de manera extraordinaria en diversos ámbitos del quehacer humano, se pueden citar como ejemplos: la toma de decisiones [1], la inteligencia de negocios [4], la recuperación de información con motores de búsqueda web [1], la predicción de tiempos de graduación [5] y el pronóstico de evaluaciones en el campo educativo [6], entre otros.

Una de las aplicaciones de la DM es conocida como Minería de Datos Educacional (EDM por sus siglas en inglés, Educational Data Mining), la cual en los últimos años ha sido un campo de amplio estudio y que ha brindado posibilidades de cambio para las instituciones educativas en los diferentes ámbitos que atienden, logrando mejoras en sus procesos.

En la presente investigación, se utiliza el agrupamiento de datos para reducir la dimensión del espacio de representación, lo que permite encontrar elementos representativos de cada grupo para formar una matriz de menor dimensión.

Entre los algoritmos de agrupamiento más conocidos y utilizados, se encuentra el algoritmo K-means. Este algoritmo puede trabajar con una gran cantidad de datos, donde K es el número de grupos que se obtienen. El valor asignado al parámetro k es proporcionado por el usuario y entonces el algoritmo ubica a los objetos en el grupo que le corresponda a través del criterio de su distancia más corta a una medida central [7].

Los datos analizados en este trabajo corresponden a los registros de evaluación que obtuvieron los alumnos en cada una de sus asignaturas durante los 10 semestres cursados a lo largo de su carrera (en un plan de 5 años). Estas evaluaciones son complementadas con datos curriculares de la carrera, mismos que se mencionan más adelante.

A partir de los historiales académicos individuales, los alumnos tendrán registros formados por a lo más 53 evaluaciones. La muestra con la que se hicieron las pruebas se obtuvo de alumnos egresados entre el 2015 y 2017.

En la investigación se utilizó el clustering como una herramienta para la reducción de variables aplicando el algoritmo K-means [7], el software WEKA (Waikato Environment Knowledge Analysis) [2] y el Lenguaje R [8].

La reducción de la dimensión o espacio de representación es importante en el análisis de grupos, ya que, no solo permite hacer manejables los conjuntos de datos de grandes dimensiones y reducir el costo computacional, sino que provee al usuario de una imagen más clara y visual para el análisis de los datos en cuestión [9].

Sin embargo, el reto más grande del clustering es la validación de los agrupamientos logrados. Esto se debe, a que, a diferencia de la clasificación supervisada, no se tiene a *priori* forma de validar los resultados. Por ejemplo, ¿cómo saber si la cantidad de grupos es la correcta? y si ¿los objetos están adecuadamente asignados al grupo? Estas, son consideraciones que hacen complicado evaluar la calidad de cualquier agrupamiento. No obstante, en este sentido, en la literatura se reportan diversos criterios para la validación, encontrado incluso que las medidas utilizadas varían de acuerdo con los índices y lo que éstos consideran para hacer la evaluación de los grupos como de buena calidad [10]. Así, por ejemplo, se tienen índices para evaluar: la separación entre los grupos, la cohesión entre los elementos, como es el caso del índice de homogeneidad biológica (BHI) y la estabilidad biológica [11].

2. Planteamiento del problema

En el caso de estudio del Centro Universitario UAEM Valle de México, la información de los alumnos de nivel superior es almacenada en una base de datos. La hipótesis es la siguiente: Al utilizar algoritmos de DM, los datos pueden ser usados para entender el comportamiento de los alumnos, apoyar a los profesores, evaluar y mejorar la enseñanza presencial, entre otros. Por otra parte, la estructura curricular de la carrera significa información adicional, así como los datos de profesores y de otros procesos de enseñanza-aprendizaje, los cuales también pueden analizarse utilizando la DM para obtener relaciones significativas y no conocidas de manera explícita, de tal manera que se generen conclusiones que apoyen a la toma de decisiones en los sistemas escolares.

Otro ejemplo de la aplicación de la DM, en este ámbito, sería identificar las causas o variables que influyen en el desempeño académico de los alumnos, y de esta manera hacer una predicción de éste. Con base en los resultados, se podrían tomar acciones útiles y constructivas que ayuden a las autoridades a mejorar sus procesos en la toma de decisiones, incrementar el desempeño académico de los alumnos, disminuir la deserción escolar, reducir la reprobación, entender mejor el comportamiento de los alumnos, brindar apoyo a profesores y mejorar los procesos de enseñanza-aprendizaje.

Dada la importancia de este tema, existen diversos estudios alrededor de mundo, que están en busca de conclusiones y desarrollando investigaciones que aporten en el análisis de datos del ámbito educativo [12]. Entre estos proyectos, algunos indagan sobre las causas que dan pauta al resultado del alumno, mismo que refleja su desempeño mediante la calificación. Una conclusión a la que han llegado los autores es que dichas causas son multifactoriales, por ejemplo, en [13] se consideran factores académicos, personales y económicos de los alumnos. Por otro lado, en [14] el estrato social, situación económica, el entorno de los alumnos, así como el nivel académico de los padres se contemplan como variables que influyen en ese resultado.

En México, la educación es un tema que también ha sido objeto de mucho estudio con el fin de encontrar o desarrollar estrategias educativas que involucren métodos y herramientas innovadoras que mejoren el proceso de enseñanza-aprendizaje y que tal mejora, se vea reflejada en el desempeño académico de los alumnos en los diferentes niveles educativos, desde nivel básico (preescolar, primaria y secundaria) hasta nivel medio superior en donde se han implementado

reformas Integrales de educación para lograr un desempeño deseable en las evaluaciones tipo PISA[1]

En ese mismo sentido, en las instituciones de educación superior también se hacen grandes inversiones en aspectos como infraestructura, capacitación y actualización docente e investigación educativa, con el objetivo de mejorar los procesos involucrados en la educación y de esta manera, mejorar los resultados de sus egresados.

Los datos con los que se trabajaron se obtuvieron del plan de estudios de la carrera de Ingeniería en Sistemas y Comunicaciones del Centro Universitario UAEM Valle de México, así como de las trayectorias académicas de alumnos egresados entre el 2015 y 2017.

Predecir el desempeño de los alumnos permite identificar a aquellos que necesitan atención especial para reducir la tasa de fracaso y tomar acciones apropiadas para las siguientes pruebas semestrales.

De igual manera, el presente trabajo de investigación puede ayudar a la detección de alumnos en riesgo, para que las autoridades correspondientes asignen a un profesor-tutor para evitar el rezago y posteriormente la deserción del alumno.

Como se ha visto, la aplicación de la EDM ha permitido encontrar aspectos que impactan en el desempeño escolar de los alumnos.

Los supuestos que motivan la realización de este trabajo de investigación son:

a. Las evaluaciones obtenidas por un alumno en las materias cursadas dan pauta para predecir su desempeño en las materias del semestre inmediato siguiente
b. Si se detectan a tiempo alumnos en riesgo, basándose únicamente en las evaluaciones, se podrían aplicar estrategias para evitar el rezago educativo y la deserción académica.

Es evidente que, a mayor cantidad de dimensiones en una matriz, mayormente complejo es el análisis de los datos y esto puede llevar a lo que se conoce como la maldición de la dimensionalidad [9] durante el proceso de clustering. Así, uno de los grandes beneficios de la reducción de dimensión es que brinda mayor facilidad en el manejo de los datos.

El problema abordado en este trabajo de investigación es cómo reducir la dimensión de los registros de las evaluaciones de los alumnos (objetos) ya que las filas que representan a cada alumno tienen dimensiones 6, 11, 19, 26, 31, 36, 41, 46, 52 y 53, dependiendo del grado de avance en la carrera. Estos números, corresponden al total de evaluaciones obtenidas al terminar 1º, 2º, hasta el décimo semestre. Como se puede apreciar, entre mayor grado de avance, mayor cantidad de variables, así, reducir la cantidad sin pérdida de información, claramente tiene muchos beneficios.

Para almacenar las evaluaciones, se definió una matriz de 98 filas (registros de alumnos) por 53 columnas(materias).

3. Trabajo relacionado

En [12] se hace la revisión del estado del arte en la EDM y considera más de 240 trabajos de investigación para describir 222 enfoques diferentes y 18 herramientas utilizadas en EDM.

Entre los trabajos analizados se encuentran propuestas de soluciones para mejorar el desempeño académico de los alumnos. Por ejemplo: la identificación de alumnos con desempeño destacado para otorgarles becas o identificar alumnos con bajo desempeño para proporcionarles recursos de tutoría de manera oportuna [15].

Así, la EDM busca, principalmente modelar el comportamiento de los alumnos y encontrar posibles soluciones a los problemas detectados, por ejemplo, en [16] analizan los datos de alumnos universitarios de la UAEM CU VM para determinar su perfil académico mediante el uso de un ensamble de clasificadores. Las clases que consideraron son excelente, bueno y regular, y

[1] https://www.inee.edu.mx/index.php/evaluaciones-internacionales/reactivos-liberados-pisa-2016

cada alumno es clasificado con base en su promedio general, actividades de estudio, formas de aprendizaje y hábitos de estudio. Los resultados obtenidos lograron un 80.4 % de acierto en la clasificación.

En el caso de [17] en la Universidad Autónoma Metropolitana (UAM) trabajaron con datos de alumnos universitarios de primer ingreso, asociados a un curso de nivelación de matemáticas y con base a éstos se hace la predicción de dos cursos posteriores. Lo anterior con el fin de identificar el impacto de la calificación obtenida en el curso de nivelación de matemáticas, sobre la calificación de los cursos de Complementos de Matemáticas e introducción al Cálculo y de esta manera, determinar si el curso ayuda a los alumnos a tener un mejor desempeño; en el trabajo se utiliza la técnica de reglas de asociación. La conclusión a la que llegaron es que el curso de nivelación no está ayudando a los alumnos a acreditar los dos cursos posteriores.

En [18], encuentran relaciones utilizando reglas de asociación entre los factores de preadmisión de los alumnos y el primer año dentro de la universidad, buscaban identificar sí la manera en la que son elegidos los alumnos permite detectar a aquellos que tendrán un buen desempeño académico, tomando como variables: la escuela de procedencia, el promedio general de la preparatoria y la edad de ingreso del alumno; como resultados encuentran que la escuela de procedencia no es relevante, sin embargo, la edad de ingreso y el promedio general de la preparatoria sí lo son.

En [19] buscan relaciones entre las actividades diarias de alumnos universitarios y su desempeño académico universitario mediante el uso de sensores en smartwatches que permiten el monitoreo de sus actividades. Los resultados del reconocimiento de actividades con Random Forest fueron del 86.9 % de acierto.

4. Propuesta metodológica

Para el desarrollo del presente trabajo, se aplicó una metodología que tiene como base la metodología KDD, y que consta de 5 etapas para la reducción de la dimensión de los registros de evaluaciones de los alumnos, como se puede apreciar en la figura 1.

Fig. 1. Metodología aplicada en el presente trabajo.

En la etapa 1, se realizó un análisis de la currícula de la carrera a modelar y de los historiales académicos para determinar las variables que se van a utilizar: es decir, total de materias, materias por semestre, dependencias entre materias, el tipo de materia (obligatoria u optativa), los núcleos de formación, oportunidades para acreditar una materia.

Como resultado del análisis, en el caso de estudio de ISC, se determinó tomar para la descripción de los objetos, las evaluaciones obtenidas por el alumno en examen ordinario la primera vez que cursa la materia, puesto que tiene derecho a 6 oportunidades para acreditarla, las primeras 3 (Ordinario, Extraordinario y Título de Suficiencia) son en periodo regular o primera vez que cursa la materia y las mismas tres oportunidades tendrá, en caso de no acreditarla y volver a cursarla.

En la etapa 2, Se mapearon las materias y las evaluaciones de cada alumno en una matriz en Excel para poder procesarlas. Las matrices obtenidas contienen registros de 98 alumnos con diferentes dimensiones desde 6 hasta 53 atributos, los cuales corresponden a las evaluaciones finales de las materias. En la Fig. 2, se muestra la imagen de una parte de la matriz generada a partir de los historiales académicos individuales de los alumnos considerando las 53 materias.

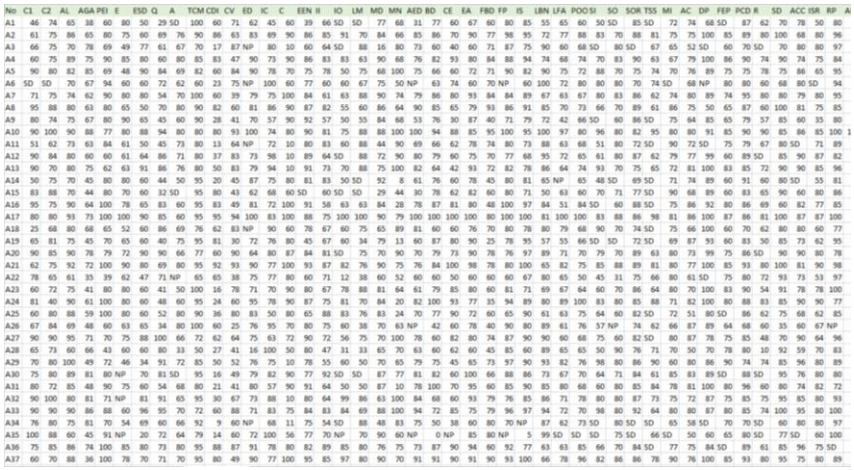

Fig. 2. Parte de la Matriz 53 x 98 con registros de evaluaciones.

En la etapa 3, se quitaron de la matriz aquellos datos que no aportan a la solución del problema, se completan datos faltantes y se realizan transformaciones para que cumplan con los requisitos para aplicar los algoritmos. En este caso, se realizó la revisión de los datos concentrados en la matriz de Excel y debido a que se utilizó el algoritmo K-means era necesario que todos los datos fueran de tipo numérico, lo cual no se cumplía porque existen evaluaciones alfabéticas SD (Sin Derecho) o NP (No Presentó) que para efectos de procesamiento en el departamento de Control Escolar se toman como 0, ya que la materia no fue acreditada por el alumno. Por este motivo, SD y NP fueron sustituidos por 0.

La etapa 4 se refiere a la formación de grupos candidatos de materias, en este punto se generaron matrices de Excel por semestre convertidas en archivos .csv para ser procesadas en WEKA. Aplicando el algoritmo K-means con diferentes parámetros, se obtuvieron varios agrupamientos por semestre los cuales se utilizan en la etapa 5 para elegir el mejor agrupamiento.

En la etapa 5, se reduce de la dimensión de los registros de evaluaciones.

Primero se identifica el mejor agrupamiento, utilizando criterios basados en medidas internas mediante los índices de conectividad, índice Dunn y ancho de la silueta (contenidos en el paquete clValid de R Studio) [8]. Estos índices, son de los más utilizados para este tipo de problemas.

La conectividad indica el grado de conectividad de los Clústeres o grupos según lo determinado por los k vecinos más cercanos (el argumento neighbSize del índice). La conectividad tiene un valor entre el 0 y el infinito y debería ser minimizada. Ambos, tanto el ancho de silueta como el índice Dunn combinan la compactación y la separación de los Clústeres. El ancho de silueta es el promedio de cada valor de observación de silueta. El valor de silueta mide el grado de confianza en una asignación de agrupación en particular y se encuentra en el intervalo [-1,1], con objetos agrupados que tienen valores cercanos a 1 y observaciones agrupadas pobremente que tienen valores cercanos a -1. El índice de Dunn es la relación entre la distancia más pequeña entre los objetos que no están en el mismo grupo y la distancia mayor dentro del clúster. Tiene un valor entre 0 e infinito y debe maximizarse [11].

Una vez determinado el mejor agrupamiento, el centroide de cada grupo será elegido como el representante de todas las materias que contiene, logrando así la reducción de la dimensión de los registros de las evaluaciones.

La Tabla 1 muestra el concentrado de las materias que conforman el plan de estudios de la carrera de Ingeniero en Sistemas y Comunicaciones del Centro Universitario UAEM Valle de México y las etiquetas asignadas a éstas. De las 59 materias, el alumno debe cursar solamente 53.

Tabla 1. Las 59 materias del plan de estudios de la carrera de Ing. en Sistemas y Comunicaciones.

MATERIA	ID	MATERIA	ID	MATERIA	ID
Inglés C1	C1	Métodos Numéricos	MN	Redes	R
Inglés C2	C2	Algoritmos y Estructura de Datos	AED	Sistemas Digitales	SD
Álgebra Lineal	AL	Base de Datos	BD	Administración de Centros de Cómputo	ACC
Álgebra y Geometría Analítica	AGA	Circuitos Eléctricos	CE	Interconexión y Seguridad de Redes	ISR
Probabilidad y Estadística para Ingenieros	PEI	Electrónica Analógica	EA	Residencia Profesional	RP
Electromagentismo	E	Fundamentos de Base de Datos	FBD	Administración de Base de Datos	ABD
Estática y Dinámica	ESD	Fundamentos de Programación	FP	Calidad de Software	CS
Química	Q	Ingeniería de Software	IS	Programación Avanzada	PA
Administración	A	Lenguajes de Bajo Nivel	LBN	Inteligencia Artificial	IA
Técnicas de Comunicación	TC	Lenguajes Formales y Automátas	LFA	Seminario de Titulación	ST
Cálculo Diferencial e Integral	CDI	Programación Orientada a Objetos	POO	Sistemas de Tiempo Real	STR
Cálculo Vectorial	CV	Sistemas de Información	SI	Comunicación Vía Microondas y Satelital	CVM
Ecuaciones Diferenciales	ED	Sistemas Operativos	SO	Teoría del Control	TC
Introducción a la Computación	IC	Sistemas Operativos para Red	SOR	Planeación Estratégica	PE
Contabilidad	C	Temas Selectos de Sistemas	TSS	Auditoría y Seguridad Informática	ASI
Ecología Ética y Normatividad	EEN	Metodología de la Investigación	MI	Sistemas de Instrumentación y Control	SIC
Intrducción a la Ingeniería	II	Arquitectura de Computadoras	AC	Sistemas Electrónicos de Comunicación	SEC
Ivestigación de Operaciones	IO	Desarrollo de Proyectos	DP	Transmisión y Comunicación de Datos	TCD
Lógica Matemática	LM	Formulación y Evaluación de Proyectos	FEP	Sistemas Expertos	SE
Matemáticas Discretas	MD	Protocolos de Comunicación de Datos	PCD		

5. Experimentos

A partir de la muestra obtenida en la etapa 2, se realizaron experimentos con varias corridas variando parámetros del algoritmo K-means (K y la semilla), los experimentos, se describen a continuación:

Experimento 1. Para las corridas en WEKA, se consideraron diferentes valores para k y se variaron las semillas en cada uno de los semestres; obteniéndose los resultados que se muestran en la Tabla 2.

Notar que el primer semestre consta de 6 materias, el segundo de 11 y el sexto de 36 materias:

Tabla 2. Resultado de corrida de K-means en WEKA para primer, segundo y sexto semestre.

Sem	K	Seed	Distribución	Conformación de grupos
1	**2**	**10**	**3, 3**	**(A, AGA, ESD) (TCM, IC, II)**
1	**2**	**5**	**5, 1**	**(AGA, ESD, IC, II, TCM) (A)**
1	2	3	3,3	(AGA, TCM, IC) (A, ESD, II)
2	3	10	4, 2, 5	(A, IC, II, CDI) (AGA, ESD) (TCM, AL, FP, AC, Q)
2	3	5	5, 1, 5	(AGA, ESD, II, CDI, Q) (A) (TCM, IC, AL, FP, AC)
2	3	3	6, 3, 2	(TCM, IC, II, AL, FP, AC) (AGA, ESD, CDI) (A, Q)
6	5	10	9, 14, 3, 6, 4	(IC,AL,FP,AC,EEN,AED,LBN,C2,POO) (TCM,II,ED,C,MD,PEI,E,BD,CE,SOR,IA) (A,CDI,LM,IO) (Q,CI,MN,FBD,SO,LBN,SI) (AGA,ESD,CV,R,ACC)

En la Tabla 2: la primera columna corresponde al semestre en que se aplica el clustering, la columna K corresponde al número de grupos a formar, la columna Seed corresponde al valor asignado a la semilla, la columna Distribución corresponde al número de materias que forman cada grupo según la K y la columna Conformación de grupos corresponde a las materias (identificadas con el id mostrado en la Tabla 1) que conforman los grupos de acuerdo con la distribución.

Como puede observarse en la Tabla 2, el cambio en el valor de la semilla modifica la distribución de las materias en los grupos y aparentemente se ven más equilibrados. En el caso de la primera fila, trabajando con primer semestre, el valor para K = 2, a la semilla (seed) se asigna valor 10, después de aplicar el algoritmo K-means, se obtiene una distribución de 3 elementos para el grupo 1 y 3 elementos para el grupo 2, se puede observar en la columna Conformación de grupos, que el primero está formado por las materias (A, AGA, ESD) y que el segundo por las materias (TCM, IC, II)

Así en la segunda fila, el valor de K no cambia, se varía la semilla a 5 y la distribución cambia a 5 elementos para el primer grupo y un elemento para el segundo, la columna conformación nos dice qué elementos quedaron en los grupos.

En la literatura se encontró que el algoritmo K-means en WEKA [20], también trabaja con valores nominales (SD y NP) y para este caso computa la moda para tratar los valores nominales como valores numéricos. Para el caso particular de este trabajo no funcionó tomar la moda; por lo tanto, fue importante realizar la limpieza de los datos (etapa 2 de la propuesta metodológica).

Experimento 2. En el caso de la validación de grupos se encontró que es muy difícil poder llegar a validar el valor óptimo de K [11,21,22]. Para evaluar el resultado de los algoritmos de agrupamiento se utilizan índices de validación que miden distintos aspectos de las características de los grupos logrados [23]. En el caso particular de este trabajo de investigación, se trabajó con un paquete de validación de clustering ya construido del lenguaje R llamado clValid [11]. Este contiene funciones para validar los resultados del análisis de clustering. El paquete ofrece 3 tipos de medidas de validación de clustering "internas", "de estabilidad" y "biológicas". En el presente trabajo se utilizaron las medidas internas y el algoritmo K-means.

La ejecución en el lenguaje R para la validación de los agrupamientos logrados arrojó como resultados los datos de la Tabla 3.

Como podemos observar, en la Tabla 3 se muestran los valores óptimos para K en cada uno de los semestres, desde el primero hasta el sexto.

La columna semestre corresponde al semestre en el que se va a hacer el agrupamiento (va desde el 1 hasta el 10), la columna número de materias muestra la dimensión de la matriz y por lo tanto es la cantidad de materias a agrupar, la columna Connectivity muestra el valor óptimo para K según el índice de conectividad, la columna Dunn muestra el valor óptimo para K según el índice Dunn, la columna Silhouette muestra el valor óptimo para k según el índice de ancho de la silueta, la columna valor óptimo K muestra el valor óptimo para K después de elegir; en [24] proponen un método de ensamble para la validación de clustering, basado en las medidas internas trabajadas en la presente investigación, tomando en consideración la votación de mayoría simple.

Entonces teniendo la matriz de dimensión 36 (Ver Figura 3) quedará reducida a dimensión 5 (Ver figura 4)

Tabla 3. Número de materias por semestre y valor óptimo para k**.**

Semestre	No de materias	Connectivity	Dunn	Silhouette	Valor óptimo K
Primero	6	2	5	2	2
Segundo	11	2	10	2	2
Tercero	19	2	18	6	6
Cuarto	26	2	25	2	2
Quinto	31	2	30	4	4
Sexto	**36**	**2**	**35**	**5**	**5**

ETQ	AGA	ESD	A	TCM	IC	II	AL	CDI	FP	AC	Q	C1	ED	C	EEN	MD	MN	AED	LBN	C2	PEI	E	CV	LM	FBD	SO	BD	CE	IO	POO	R	LFA	SI	SOR	ACC	IA
A1	38	50	0	100	45	66	65	60	80	74	29	46	62	60	39	77	68	31	55	74	60	80	71	0	60	0	77	60	0	60	62	65	50	85	78	75
A2	65	60	76	90	69	85	86	86	98	75	69	61	83	90	86	84	66	85	72	75	80	75	63	70	77	70	86	70	91	88	80	77	83	88	68	100
A3	78	77	67	70	80	64	70	17	87	52	61	66	0	10	60	16	80	73	90	75	69	49	87	88	71	80	60	40	0	68	0	60	0	0	80	0
A4	75	80	80	85	73	83	89	83	88	79	60	60	90	90	86	90	68	76	74	75	90	85	47	63	84	83	82	93	83	74	74	68	70	90	74	85
A5	85	90	69	82	78	78	82	60	90	76	84	90	90	70	75	68	100	75	90	80	69	48	84	75	71	70	66	60	50	72	78	75	88	75	86	0
A6	67	60	62	60	100	60	70	23	0	68	72	0	0	60	77	75	50	0	100	0	94	60	75	67	70	80	63	74	60	80	60	72	80	70	80	70
A7	62	80	70	100	75	61	74	60	84	80	54	71	79	100	84	90	74	79	67	75	90	80	39	88	84	83	86	80	63	67	80	63	80	86	79	85
A8	63	50	80	90	86	82	80	82	86	75	70	95	81	90	87	86	64	90	85	88	80	65	60	60	93	70	85	65	55	73	60	70	66	89	81	90
A9	67	65	60	90	57	57	75	28	71	64	45	80	70	90	92	84	68	53	72	74	80	90	41	55	40	60	76	30	50	66	57	42	0	86	60	75
A10	88	88	80	80	74	81	90	80	100	80	94	90	100	80	90	88	100	100	100	100	77	80	93	88	95	80	94	88	75	80	90	97	96	82	86	75
A11	63	50	73	80	72	83	73	13	80	72	45	51	0	10	80	44	90	69	88	62	84	61	64	88	74	80	66	62	60	68	67	63	51	72	0	75
A12	60	64	71	80	98	64	80	37	77	77	86	90	73	10	89	72	90	80	95	84	60	61	83	88	70	80	79	60	0	65	0	72	61	87	90	80
A13	75	91	76	80	94	73	80	50	82	81	86	90	79	10	91	75	100	82	86	70	62	63	83	88	72	70	64	42	70	74	72	64	93	75	90	62
A14	45	60	50	95	75	83	70	20	80	74	44	50	87	80	81	92	8	61	65	75	80	80	45	0	45	0	76	60	50	65	60	0	48	69	0	85
A15	44	60	0	95	68	60	70	80	80	68	32	83	62	60	0	29	44	30	50	88	80	70	43	0	60	71	78	62	0	60	65	63	70	77	60	85
A16	64	65	60	95	72	58	90	83	100	86	83	95	81	100	91	84	28	78	84	75	100	78	49	63	48	60	87	81	63	84	69	51	0	88	82	85
A17	73	90	60	95	83	75	93	95	100	86	85	80	100	100	88	90	79	100	81	80	100	100	94	100	80	88	100	100	100	100	81	100	83	86	87	80
A18	68	60	69	76	90	67	80	62	80	66	86	25	0	60	78	65	89	81	80	68	65	52	83	75	70	70	60	60	60	68	62	79	90	74	80	60
A19	45	60	75	95	76	67	75	81	78	87	40	65	72	80	45	79	13	60	57	81	70	65	30	34	25	0	87	80	60	66	50	55	0	72	73	85
A20	78	90	66	77	80	81	90	60	76	73	90	90	64	87	84	70	90	70	89	85	79	72	90	75	78	70	79	73	0	70	0	71	79	89	90	75
A21	72	80	80	95	77	87	92	92	80	77	69	62	90	100	93	90	75	76	65	75	100	90	93	76	78	88	84	100	82	75	80	82	85	89	81	100
A22	35	47	0	65	77	71	61	65	60	61	71	78	75	80	60	60	52	60	80	65	39	62	38	38	60	31	60	50	12	50	72	65	45	75	73	70
A23	41	60	50	100	70	67	75	16	81	70	41	60	71	90	80	81	64	61	69	72	80	80	78	88	60	70	79	85	78	64	54	67	60	86	78	85
A24	61	60	60	95	78	75	90	24	94	82	48	81	95	90	87	84	20	82	80	40	100	80	60	70	35	80	100	93	81	100	83	89	83	85	90	90
A25	59	60	80	90	50	88	88	36	65	51	52	60	83	80	65	83	24	70	61	80	100	80	80	76	60	60	77	90	83	75	62	63	64	82	68	85
A26	48	65	80	100	95	75	69	60	90	87	34	67	76	70	80	70	63	0	89	84	60	63	25	38	40	0	42	60	60	76	60	61	57	74	60	87
A27	71	88	66	72	63	72	95	62	87	87	100	90	75	72	90	70	100	78	90	90	70	75	64	75	74	60	60	82	56	68	48	60	75	82	90	64
A28	66	60	33	50	100	47	60	27	85	50	80	65	16	50	80	65	70	63	89	73	43	60	41	33	45	90	60	62	31	65	10	65	50	76	59	60
A29	49	34	72	85	75	55	100	50	97	80	91	70	76	10	78	70	65	79	93	80	72	46	52	50	73	80	75	45	60	76	74	82	98	86	96	68
A30	81	70	0	95	82	92	89	16	88	83	81	75	79	90	77	87	77	81	73	80	80	0	49	0	66	71	82	60	0	70	0	67	64	84	76	85
A31	48	60	68	80	57	64	85	21	85	81	54	80	80	90	91	87	10	78	85	72	90	75	41	50	60	80	100	70	50	68	60	80	60	85	74	70
A32	81	81	65	95	88	64	80	30	76	72	91	90	73	10	80	63	100	84	86	100	71	0	67	86	79	80	68	60	99	78	75	71	80	87	85	85
A33	86	96	70	72	83	83	90	60	96	80	95	90	71	75	84	88	100	94	94	90	88	60	88	69	79	80	72	85	84	70	74	72	98	92	95	65
A34	81	69	66	92	68	54	75	9	70	58	60	76	0	11	75	48	83	75	87	80	70	54	60	88	80	80	50	38	0	73	0	62	0	0	80	0
A35	45	20	64	79	100	70	60	14	0	50	72	100	72	56	77	90	60	0	99	88	91	0	60	70	80	75	0	0	0	0	0	0	0	0	0	0
A36	74	80	80	95	78	89	86	88	92	75	73	75	91	80	82	76	75	73	63	85	100	85	87	80	60	70	87	90	85	85	61	63	66	84	96	85
A37	36	70	70	95	77	85	88	80	93	76	71	60	90	100	95	90	70	91	66	70	100	78	49	80	90	86	91	90	97	96	80	78	82	86	75	100

Fig. 3. Matriz de sexto semestre dimensión 36.

6. Resultados

Al llevar a cabo la clusterización y la validación de los clústeres, se esperaba que los agrupamientos logrados tuvieran materias pertenecientes a la misma área de conocimiento, de las 4 que se tienen en la estructura curricular de la carrera elegida, las cuales son Ciencias Básicas y Matemáticas (CBM), Ciencias Sociales y Humanidades (CSH), Ciencias de la Ingeniería (CI) e Ingeniería Aplicada (IA); o por núcleo de formación, sin embargo, lo anterior no ocurrió, en este caso, los agrupamientos que se formaron dependían de otros factores tales como el profesor asignado a la materia y el turno en el que se oferta.

En la tabla 2, se puede ver que la última columna la reducción de dimensión para la matriz de sexto semestre es de dimensión 36 para convertirse a una matriz de dimensión 5, quedando de esta manera (Fig. 4).

ETQ	CDI	Q	AGA	SOR	A
A1	45.00	41.75	57.17	85.00	61.39
A2	92.33	74.00	64.50	88.00	82.39
A3	5.67	55.00	71.33	0.00	58.89
A4	83.67	71.25	71.50	90.00	82.72
A5	36.67	80.13	82.50	75.00	75.44
A6	51.00	68.88	59.33	70.00	57.17
A7	69.33	75.75	68.50	86.00	80.06
A8	75.67	69.13	69.00	89.00	82.72
A9	51.00	45.88	57.17	86.00	73.39
A10	76.67	92.50	88.83	82.00	87.44
A11	49.33	69.75	48.33	72.00	64.72
A12	39.00	68.38	74.50	87.00	73.06
A13	60.67	80.63	78.50	75.00	73.44
A14	51.67	25.63	43.33	69.00	76.06
A15	55.00	50.63	58.67	77.00	58.22
A16	77.00	50.25	72.67	88.00	83.50
A17	91.67	87.00	87.33	86.00	90.61

Fig. 4. Matriz dimensión 5 para sexto semestre.

En la Fig. 4, se observa un fragmento de la matriz de dimensión 5 para sexto semestre, que en un principio era de dimensión 36. Como se puede ver se conservan los centroides de cada grupo y es con los que se procederá a trabajar posteriormente.

7. Conclusiones y trabajo futuro

Lo que se puede concluir de los resultados mostrados en la sección anterior, es que por medio de la aplicación del algoritmo K-means se puede reducir la dimensión de los registros de las evaluaciones de los alumnos. Esto se logra tomando los centroides del mejor agrupamiento, lo que permite trabajar ahora con un representante de cada grupo de materias.

Los resultados obtenidos del análisis mostraron que los agrupamientos dependen de las relaciones entre las evaluaciones, sin embargo, hay factores como los horarios en los que se toma la materia y el profesor asignado que influyen en el resultado.

El modelo presentado en este trabajo, puede ser replicado en otras carreras.

Por otro lado, a pesar de que WEKA es una herramienta altamente utilizada para este propósito, lamentablemente no cuenta con un módulo de validación de clustering; motivo por el cual se recurrió a utilizar R Studio.

Como trabajo futuro se hará la predicción del desempeño académico de los alumnos de la carrera de Ingeniería en Sistemas y Comunicaciones del Centro Universitario UAEM Valle de México, mediante técnicas de clasificación. Asimismo, se pueden incluir otras variables en el análisis, como el tipo de examen con el que se aprobó la materia y número de oportunidad en la

que fue acreditada. También se puede considerar como trabajo futuro el utilizar otros algoritmos de agrupamiento y otras herramientas diferentes.

Referencias

1. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. (2011)
2. Frank, E., Hall, M.A., Witten, I.H.: The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques, Fourth Edition, Morgan Kaufmann (2016)
3. Huapaya, C.R., Lizarralde, F.A., Arona, G.R., Massa, S.M.: Minería de Datos Educacional en Ambientes Virtuales de aprendizaje. In: Workshop de Investigadores en Ciencia de la Computacion (WICC), pp. 996–1000 (2012)
4. Bhawana, D., Hardik, J., Vedant, S., Gandla, R.: CRM Application For Analizing the Sales Using Data Mining. International Journal of Engineering Science and Computing, pp. 16138–16139 (2018)
5. Vassilis, T., Emmanuel, I.E.L.P., Nikos, K., Panagiotis, P.: Prediction of students' graduation time using a two-level classification algorithm. In: Conference: IEEE 1st International Conference on Technology and Innovation in Learning, Teaching and Education (ITHET) (2018)
6. Bhardwaj, B.K., Pal, S.: Data Mining: A prediction for performance improvement using classification. CoRR, vol. abs/1201.3418 (2012)
7. Hartigan, J.A., Clustering algorithms (Wiley series in probability and mathematical statistics) (1975)
8. R. D. C. Team: R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing (2008)
9. Chris, D., Tao, L.: Adaptive Dimension Reduction Using Discriminant Analysis andK-means. In: International Conference on Machine Learning, Corvallis (2007)
10. Alsabti, K., Ranka, S., Singh, V.: An efficient k-means clustering algorithm. Electrical Engineering and Computer Science (1997)
11. Guy, B., Vasyl, P., Susmita, D., Somnath, D.: clValid: An R Package for Cluster Validation. Journal of Statistical Software, vol. 25, pp. 1–22 (2008)
12. Peña-Ayala, A.: Educational data mining: A survey and a data mining-based analysis of recent works. Expert Systems with Applications, pp. 1432–1462 (2014)
13. Ruby, J., David, K.: Analysis of Influencing Factors in Predicting. International Journal of Innovative Research in Computer and Communication Engineering, pp. 1085–1092 (2015)
14. Organista, J., McAnally, L., Henríquez, P.: Clasificación de estudiantes de nuevo ingreso a una universidad pública, con base a variables de desempeño académico, uso de tecnología digital y escolaridad de los padres. Revista Electrónica de Investigación Educativa, pp. 34–55 (2012)
15. Dutt, A., Ismail, M.A.: A Systematic Review on Educational Data Mining. IEEE Access, pp. 15991–16005 (2017)
16. Quintana Lopez, M., Flores Albino, J.M., Lazcano Salas, S., Landassuri Moreno, V.M.: Ensamble de Clasificadores para Determinar el Perfil Académico del Estudiante usando Árboles de Decisión y Redes Neuronales. Research in Computing Science (2018)
17. Gonzalez-Brambila, S.B., Sanchez-Guerrero, L., Ardon-Pulido, I., Figueroa-Gonzalez, J., Gonzalez-Beltran, B.A.: Predicting academic performance of engineering students after

approving a mathematics leveling course using decision trees. Research in Computing Science (2018)

18. González-Brambila, S.B., Figueroa-González, J.: Looking Relationships between Pre-admission. Research in Computing Science, pp. 113–124 (2017)
19. Oscar, H.A., Ari Yair, B.-A., Miguel, G.-M., Castro-Espinoza, F.: Monitoring Student Activities with Smartwatches: On the Academic Performance Enhancement. Sensors, vol. 19, no. 7 (2019)
20. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. SIGKDD Explorations, pp. 10–18 (2009)
21. Charrad, M., Nadia, G., Véronique, B., Azam, N.: NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. Journal of Statistical Software, vol. 61, p. 1–36 (2014)
22. Kassambara, A.: Determining The Optimal Number Of Clusters: 3 Must Know Methods. Sthda.com, http://www.sthda.com/english/articles/29-cluster-validation-essentials/96-determining-the-optimal-number-of-clusters-3-must-know-methods/. Último acceso: 2019/03/25
23. Kassambara, A.: Practical Guide To Cluster Analysis in R Unsupervised Machine Learning. STHDA (2017)
24. Sunghae, J.: An Ensemble Method for Validation of Cluster Analysis. International Journal of Computer Science Issues, vol. 8, no. 6 (2011)