ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Negocios y Administración Pública

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL DE ESPECIALIZACIÓN

Modelos predictivos del rendimiento académico en Álgebra a través de la minería de datos educativa.

Estudio de Patrones de Interacción en la Plataforma Moodle del Campus Virtual CBC

AUTOR: ANDREA GACHE
MENTOR: DRA. MARÍA JOSÉ BIANCO

DICIEMBRE 2021

Resumen

El objetivo del presente trabajo es generar modelos predictivos del rendimiento académico, empleando métodos de la minería de datos educacional, que sirvan como herramientas de detección de factores vinculados a la interacción de los estudiantes en el aula virtual de Álgebra para Ciencias Económicas , del Ciclo Básico Común de la Universidad de Buenos Aires. Para ello se consolida una base con datos de los estudiantes referidos a la interacción con los recursos en el aula Moodle durante el primer cuatrimestre del 2021. Se emplean técnicas de minería de datos educacional, siguiendo las etapas del Proceso de Extracción de Conocimiento en Bases de Datos. Con R Studio, se aplican métodos de clusterización, para analizar las interacciones con los recursos y los foros. Para la elección del modelo predictivo se desarrolla con RapidMiner un análisis comparativo de diferentes métodos de clasificación. Los resultados muestran que el ensamble Gradient Boosted Tree, clasifica el 98,2% de las instancias correctamente con un margen de error mínimo del 1,8% Se logra identificar atributos predictores, concluyendo que el grado de participación influye en el rendimiento académico positivamente. Todo ello subraya la importancia de la implicación activa del estudiante,frente al mero acceso a la información disponible.

Palabras clave: Análisis de Aprendizaje en Línea, Plataforma Moodle, Rendimiento académico, Minería de Datos Educativa. Pandemia

Índice

Introducción

En marzo del 2020 a días del comienzo del Primer cuatrimestre, el contexto sanitario que tocó atravesar al país aceleró el proceso de creación y puesta en marcha del Campus Virtual para el Ciclo Básico Común de la UBA, dando inicio al cursado virtual de las asignaturas entre ellas Álgebra para Ciencias Económicas.

Promediando ya el segundo año académico desde su creación, es relevante tanto para docentes responsables de Cátedra como para la propia Institución, analizar como el alumno interactúa con la plataforma de educación virtual. Cobra aquí importancia la Minería de Datos educacional (MDE).

El uso de la minería de datos en el contexto educativo tiene el potencial de trasformar los modelos existentes de enseñanza-aprendizaje al proporcionar nuevas herramientas de análisis, interacción e intervención (Aldowah, Al-Samarraie y Fauzy, 2019). Convirtiendo los datos masivos obtenidos desde la plataforma en información que posea impacto en la práctica e investigación educativa. (Ayala Franco,2021).

Esto último, lleva al planteo del siguiente interrogante ¿Qué factores vinculados a la actividad académica virtual son predictores del rendimiento académico?

La pregunta anterior conduce al objetivo general de este trabajo que es detectar a través del uso de técnicas de minería de datos educacional, los factores qué vinculados a la participación y comunicación en el aula virtual, son predictores del rendimiento académico de los alumnos de Álgebra.

Los datos por analizar corresponden a las comisiones de la Sede Paternal del Ciclo Básico Común (CBC) del Primer cuatrimestre del año 2021. Cursos desarrollados virtualmente en plena etapa de emergencia educativa por la pandemia del Covid-19.

En el primer apartado se aplicarán las etapas del proceso KDD. Descubrimiento de conocimiento en bases de datos (Knowledge Discovery in Databases) . Las que comprenden la selección , limpieza/preprocesamiento , transformación de la información y la aplicación de técnicas de la minería de datos, en particular, el clustering para seleccionar atributos asociados al uso que los alumnos hacen de los recursos disponibles en el aula virtual. Que posteriormente serán empleados como predictores del rendimiento académico.

En el segundo apartado se busca también seleccionar atributos en función de su capacidad predictiva.

Pero, analizando exclusivamente las acciones vinculadas a la comunicación asincrónica a través de los foros. De igual modo que en apartado anterior las tareas a realizar para encontrar esos atributos incluyen desde la preparación de los datos hasta la aplicación de técnicas de agrupamiento propias de la minería de datos.

Finalmente, en el tercer y último apartado y en función de los atributos seleccionados en los dos anteriores, se intentará determinar los modelos más efectivos para la predicción del rendimiento académico de los alumnos de Álgebra. Se ajustarán los parámetros de los modelos seleccionados y se evaluará su capacidad predictiva para elegir el que más se adecué al objetivo planteado.

Con la búsqueda automática de conocimiento valioso en los datos, y el descubrimiento de patrones de comportamiento, analizando cuestiones relacionadas con el uso de la plataforma, y el aprovechamiento de los recursos, se espera poder dar cuenta si lo actuado hasta aquí en el proceso de aprendizaje en línea de los alumnos requiere algún cambio de rumbo. En particular, se quiere conocer qué ocurre dentro del entorno para poder extraer conclusiones, de cara a la evaluación y mejora de las prácticas de enseñanza-aprendizaje, a fin de mantener la calidad educativa en este particular contexto.

1.Minería de datos educacional aplicada al análisis del aprendizaje en línea

En el contexto de educación virtual, la incorporación de analíticas de aprendizaje cobra vital importancia para comprender y optimizar el aprendizaje y los entornos en que tiene lugar, con el fin de mejorarlos. Éstas facilitan y aceleran el análisis de grandes volúmenes de información generados por las plataformas virtuales e implican la medida, recopilación, análisis e informe de datos sobre los estudiantes.

La analítica académica (Campbell y Oblinger, 2007) combina los datos institucionales, el análisis estadístico y los modelos predictivos. Permiten la exploración de datos para identificar informaciones nuevas y útiles para atender las expectativas y necesidades estratégicas de las organizaciones de educación superior (Rodríguez Almeida y da Silva Camargo, 2015).

La minería de datos aplicada a la educación proporciona un conocimiento intrínseco del proceso de enseñanza y aprendizaje, que, mediante la generación de modelos, pueden responder a preguntas relacionadas con el rendimiento escolar y sus problemáticas asociadas y facilitar una planificación educativa efectiva (Anoopkumar y Rahman, 2016).

En tal sentido, este apartado aborda cuestiones que tienen que ver con el análisis de las acciones que sobre los objetos y los recursos hicieron los alumnos de Álgebra para Ciencias Económicas en el Campus Virtual del Ciclo Básico Común (CBC) de la Sede Paternal durante el Primer cuatrimestre del 2021.

A partir de la aplicación de las fases o etapas KDD, ya enumeradas, se busca determinar características de interacción con el entorno destinadas a trazar perfiles de alumnos en función de su comportamiento en el aula virtual. Se propone para ello, en primer término, la obtención de la información desde la Plataforma y su preprocesamiento para la aplicación posterior de técnicas de minería de datos. En segundo lugar, la exploración descriptiva de los mismos y el uso de técnicas de clustering a fin de obtener grupos de estudiantes con características similares de uso de los recursos.

En tercer lugar, establecer perfiles que sirvan de base para analizar mediante modelos predictivos como los mismos inciden en el rendimiento académico.

1.1 Obtención y preprocesado de los datos

Para dar inicio al proceso de generación de modelos predictivos, el primer paso fue recabar los datos que pudieran aportar información relacionada con factores asociados al rendimiento académico a partir del uso que los estudiantes hacen de los recursos disponibles en el aula virtual. Cuanto más completa se realice la fase de obtención y preprocesado, más factible será el descubrimiento de nuevos patrones, con resultados más confiables.

Desde el bloque Administración de la Plataforma Moodle se consultó el informe sobre Participación en el curso, el que proporciona información por cada recurso si el estudiante accedió o no al mismo y el número de vistas realizadas.

A partir de la obtención y consolidación de los datos, se eliminaron los atributos Nombre y Apellido e IP de las computadoras de acceso, para evitar la identificación de los alumnos, de esta forma se garantiza el anonimato en la manipulación de la información durante el proceso de análisis.

A lo largo del cuatrimestre en estudio que abarca desde los primeros días de abril hasta finales de julio del 2021, se registraron un total de 149.326 accesos realizados por los 499 alumnos inscriptos. Al momento de consolidar la información, mediante operaciones de ordenamiento, filtrado y copiado de datos se obtuvo la base de datos inicial almacenada en un archivo en formato XLSX.

La Tabla 1 muestra los nombres y la descripción de los 149 atributos, a partir de los cuales se realizará el estudio.

Tabla 1

Descripción del nombre de los atributos

	Nombre de la Variable	Descripción
ID	ID	Número asignado a cada uno de los alumnos en forma aleatoria
ARCHIVOS	ARCH_1 a ARCH_4	Número de accesos Archivos teóricos-prácticos
	ARCH_1.1 a ARCH_4.1	Número de accesos Trabajos Prácticos resueltos
	PROGRAMA	Número de accesos Programa de la materia
	POL_FORO	Número de accesos Información sobre us de los foros de comunicación
	ORIENTACIONES	Número de accesos Información sobre contenido de la materia
	CRONO	Número de accesos Cronograma de actividades
	ORGANIZADOR	Número de accesos Organizador de activades sugeridas para el cursado de la materia
	INFO_EF	Número de accesos Información sobre modalidad de evaluación, fechas, condiciones de aprobación.
	INFO_E_FINAL	Número de accesos Información sobre modalidad de examen final, fechas, condiciones de aprobación.
	BIBLIO	Número de accesos url de acceso a la Bibliografía Obligatoria
	P_1 a P_5	Número de accesos Trabajos Prácticos por Unidad
	P_1R a P5R	Número de accesos Trabajos Prácticos Resueltos por Unidad
URL	V_1 a V_34	Número de accesos Videos de contenido teórico
	V_A_1 al V_D_1	Número de accesos Videos de contenido práctico a la Unidad 1
	V_A_2 al V_N_2	Número de accesos Videos de contenido práctico a la Unidad 2
	V_A_3 al V_F_3	Número de accesos Videos de contenido práctico a la Unidad 3
	V_A_4 al V_N_2	Número de accesos Videos de contenido práctico a la Unidad 4
	V_A_4 al V_N_2	Número de accesos Videos de contenido práctico a la Unidad 5
FOROS	F1 a F22	Número de accesos a los Foros de comunicación
	VF_N	Número de accesos al foro de Novedades y Orientaciones administrativas (Atributo derivado)
	VF_ACM	Número de acceso a los Foros del Acompañamiento académico (Atributo derivado)
	VF_U1 a VF_U5	Número de accesos Foros Unidad 1 a 5 (Atributo derivado)
	VF_N	Número de mensajes al foro de Novedades y Orientaciones administrativas (Atributo derivado)
	MF_ACM	Número de mensajes Foros del Acompañamiento académico (Atributo derivado)
	MF_U1 a MF_U2	Numero de mensajes en los foros Unidad 1 a 5 (Atributo derivado)
	VT_F	Número total de accesos a los foros (Atributo derivado)
	MT_F	Número total de mensajes en los foros (Atributo derivado)

EVALUACIÓN	Num_RC_EF1	Número de respuestas correctas en la Primer Evaluación Formativa
	Num_RC_EF2	Número de respuestas correctas en la Segunda Evaluación Formativa
	Cond_Reg	Condición de regularización (Atributo derivado)
	E_FINAL	Número de respuestas correctas en la Primer Evaluación Formativa
	RENDIMIENTO_ACAD	Condición de Rendimiento (Atributo derivado)

Fuente: Elaboración propia

A partir de las variables originales se derivaron dos variables cualitativas. La Tabla 2 muestra la asignación de las categorías de la variable Condición de Regularización (Cond_Reg), para ello se adopta el criterio de regularización de la materia que fija el CBC.

Tabla 2

Categorías de la variable: Condición de Regularización

Nombre de la Categoría	Criterio
Cond_Reg.: AUSENTE	Num_RC_EFI o Num_RC_EFII no realizada
Cond_Reg.: INSUFICIENTE	Num_RC_EFI o Num_RC_EFII menor que 6
Cond_Reg.: REGULAR	Num_RC_EFI y Num_RC_EFII mayor o igual a 6

Fuente: Elaboración propia

La variable dependiente Rendimiento académico se derivó a partir de las variables Condición de Regularización y número de respuestas correctas en el examen final, como se observa en la Tabla 3.

Tabla 3

Categorías de la variable: Rendimiento académico

Nombre de la Categoría	Condición
ABANDONÓ	Cond_Reg.: AUSENTE
NO SATISFACTORIO	Cond_Reg.: INSUFICIENTE o
	Cond_Reg.: REGULAR y Num_RC comprendido entre 0 y 11
POCO SATISFACTORIO	Cond_Reg.: REGULAR y Num_RC : AUSENTE o
	Cond_Reg.: REGULAR y Num_RC comprendido entre 12 y 15
SATISFACTORIO	Cond_Reg.: REGULAR y Num_RC comprendido entre 16 y 17
MUY SATISFACTORIO	Cond_Reg.: REGULAR y Num_RC comprendido entre 18 y 19
EXCELENTE	Cond_Reg.: REGULAR y Num_RC 20

Fuente: Elaboración propia

Sobre la base construida se aplican técnicas de preprocesado, normalización y discretización, según el tipo de variable. La normalización se efectúo sobre todos los atributos numéricos. Con las variables Rendimiento académico y Condición de regularización se realizó directamente la discretización manual, como ya se ha referido en párrafos anteriores.

Cerrada la primera fase del Knowledge discovery in Databases (KDD), que comprende la extracción y transformación de la información con el objetivo de preparar los datos para la aplicación de técnicas de minería, en el próximo apartado se abordará el análisis estadístico de las variables definidas y la aplicación de métodos de clustering.

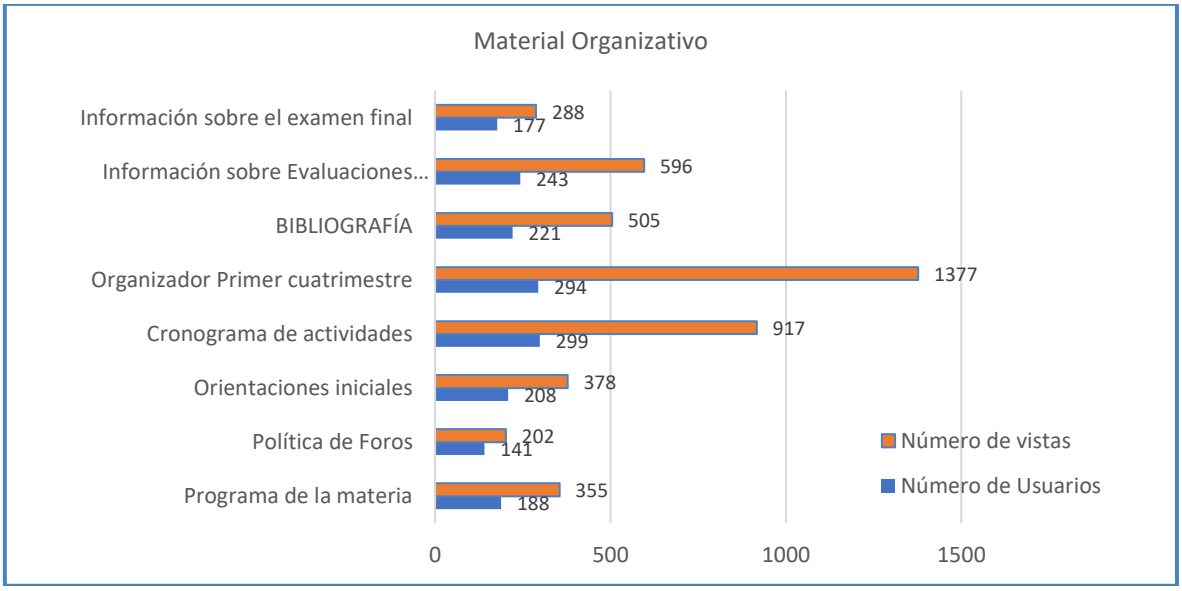
1.2 Aplicación de minería de datos educacional: Clustering

Con el propósito de analizar el grado de participación de los estudiantes se indagó inicialmente el número de accesos a cada recurso, el número de usuarios que accedieron a ese recurso y el porcentaje de acceso respecto del total de participantes.

La Figura 1 muestra el número de vistas y el número de usuarios que accedieron durante el cuatrimestre a los recursos vinculados a cuestiones metodológicas y organizativas, consideradas relevantes para la estructuración del estudio de la materia.

Figura 1

Número de vistas por recursos organizativos y por usuarios

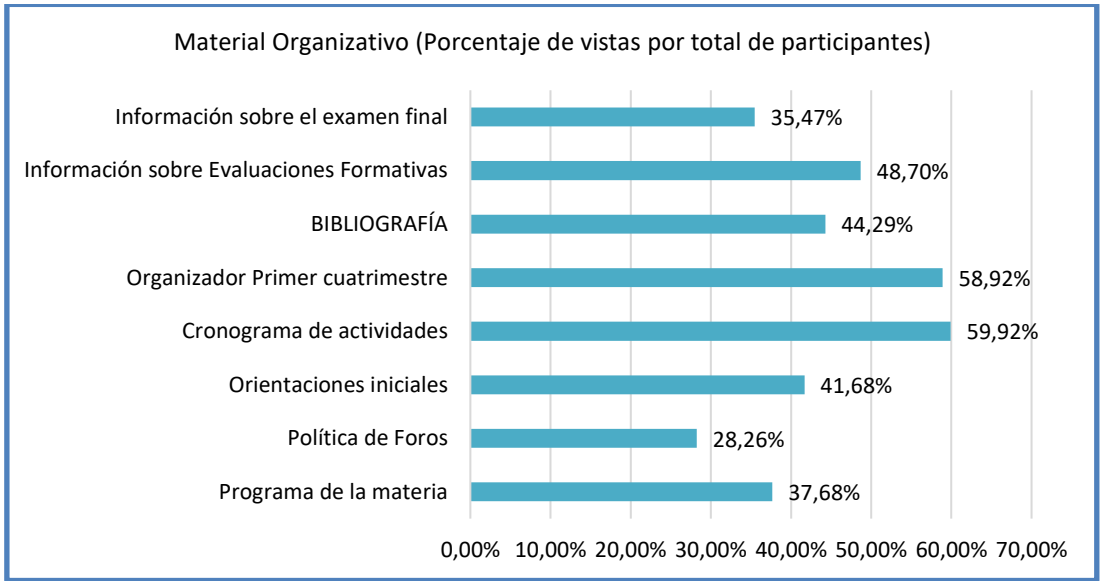


Fuente: Elaboración propia

La Figura 2 presenta el porcentaje de accesos a los recursos organizativos para la cursada sobre el total de participantes. Puede inferirse una interacción por parte de los estudiantes baja sobre estos recursos.

Figura 2

Porcentaje de participación sobre los recursos organizativos



Fuente: Elaboración propia

El análisis de ambas figuras permite un primer acercamiento al nivel de acceso y participación en el aula por parte de los estudiantes sobre recursos que deberían tener por su relevancia mayores niveles de intervención.

A partir del conjunto inicial de variables enumeradas en la Tabla 1, para el proceso de clustering sólo se consideraron las relacionadas con los materiales de estudio, las que refieren a material metodológico y las correspondientes a las respuestas de las Evaluaciones Formativas y del Examen Final. Se conforma un nuevo Data set con 109 atributos. No se tienen en cuenta las variables vinculadas a la comunicación a través de foros, dado que ese análisis será desarrollado en el siguiente apartado.

Como se mencionó en la introducción de esta sección, uno de los objetivos del trabajo es lograr una segmentación del conjunto de datos en clústers según la similitud en el uso de los recursos disponibles en el aula.

El término *clustering* hace referencia a técnicas de aprendizaje no supervisado cuya finalidad es encontrar patrones o grupos dentro de un conjunto de observaciones. Agrupando el conjunto de datos basándose en la similitud de los valores de sus atributos.

Su finalidad es revelar concentraciones en los datos o casos para su agrupamiento eficiente en clústers o conglomerados según su homogeneidad y se pueden utilizar tanto variables cualitativas como variables cuantitativas, dado que los grupos se basan en la proximidad o lejanía de unos con otros.

La técnica será utilizada con el objetivo de obtener información oculta en grandes cantidades de datos, agrupando los mismos dentro de un número de clases (Jaramillo, 2015).

Las particiones se establecen de forma que, las observaciones que están dentro de un mismo grupo sean similares entre ellas y distintas a las observaciones de otros grupos. Asignan las observaciones en *k clústers* de forma que la suma de las varianzas internas de todos ellos sea la menor posible. Este tipo de algoritmo es útil para explorar, describir y resumir datos. Su uso puede permitir confirmar (o rechazar) algún tipo de clasificación previa, así como descubrir patrones y relaciones.

Existen un gran número de métodos de clustering. En primer término, se aplicó al conjunto de datos el método K-means clustering (Macqueen, 1967), en el que, cada grupo está representado por el centro o los medios de los puntos de datos pertenecientes al grupo.

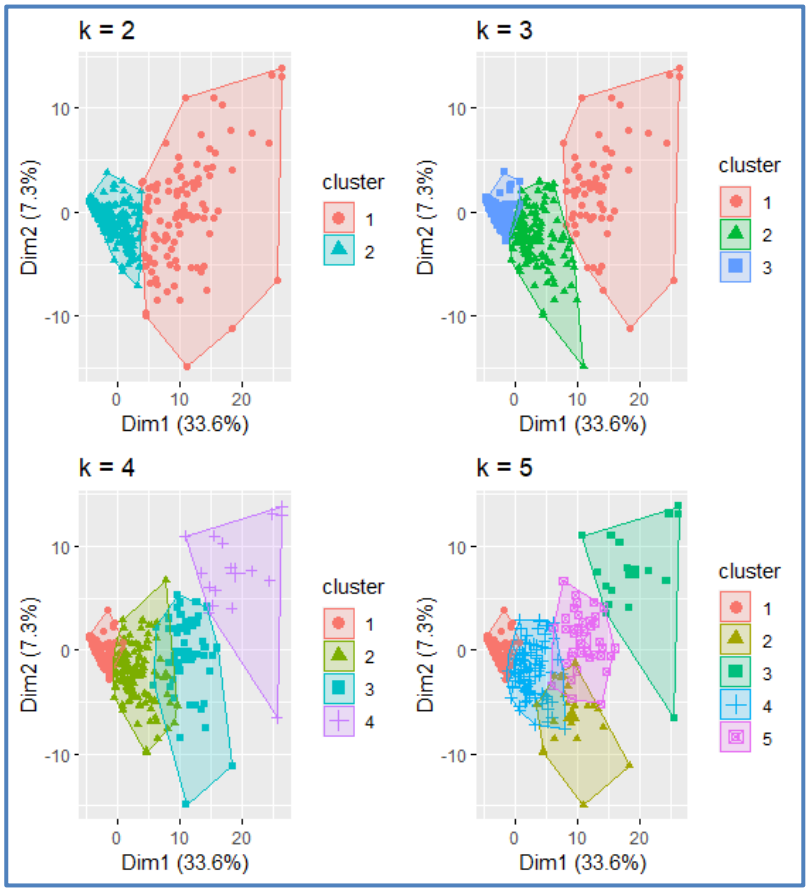
En segundo término, se aplicó K-medoids o PAM (Partitioning Around Medoids), en el que, cada grupo está representado por uno de los objetos en el grupo. Finalmente, en tercer término, se trabajó con una variante de PAM llamada CLARA (Clustering Large Aplications) que se utiliza para analizar grandes conjuntos de datos (Kaufman y Rousseeuw, 1990).

La mayoría de los autores recomiendan utilizar diversos procedimientos y comparar los resultados (Sharma, 1996; Johnson,1998). Si los distintos métodos aportan agrupaciones similares, será razonable suponer que existe una agrupación natural objetiva (Aldás, 2017).

La principal limitación de estos métodos es desconocer el número de *clústers* que se van a crear. La Figura 3 muestra las agrupaciones para valores de k arbitrariamente elegidos posibilitando ya una primera aproximación al número de grupos a definir. Se puede observar en que la segmentación en dos o tres grupos parece ser la adecuada para lograr agrupaciones no solapadas.

Figura 3

K-means para diferentes valores de k



Fuente: Elaboración propia realizada en R Studio a partir de la Base de Datos generada para este estudio

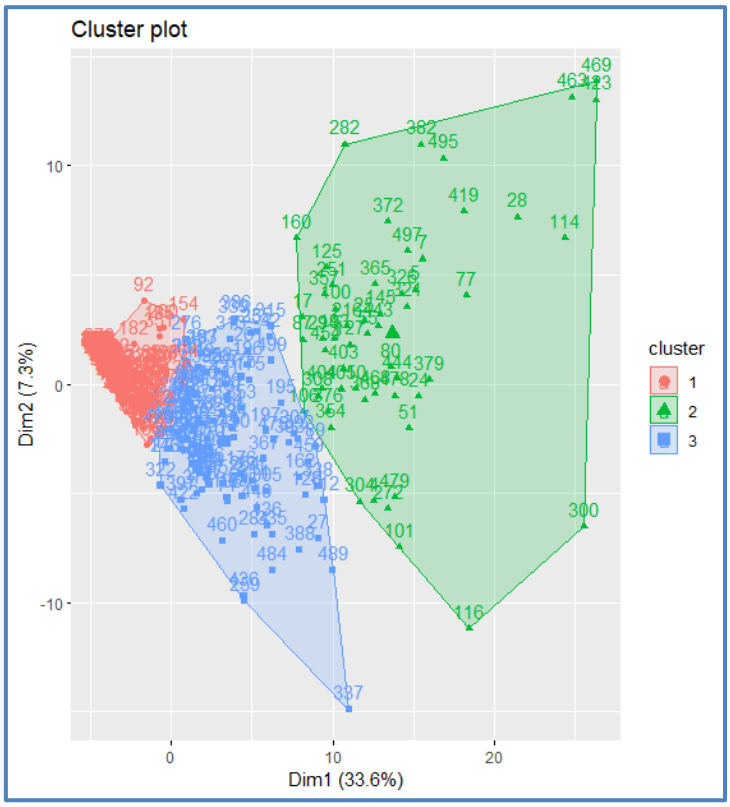
Para definir el número de clústers se aplican criterios que muestran el valor óptimo de k. El primero de ellos, el método del Codo (Within Sum of Square) que utiliza la distancia media de las observaciones a su centroide, buscando minimizarla para obtener clústeres más compactos. El segundo, Estadística de brecha (Gap Statistic), que busca encontrar la mayor distancia entre los distintos grupos. Y el tercero, el de la Silueta (Average Silhouttes width) que evalúa cuan cerca está cada punto de un clúster a puntos de los clústers vecinos.

Al ejecutar en R Studio los códigos correspondientes a cada criterio mencionado para la determinación del número óptimo de clústers (ver código en la Figura A del Apéndice) los métodos del Codo y Estadística de brecha proponen trabajar con tres clústers como número óptimo de grupos, mientras que el método de la silueta propone hacerlo con dos.

Con los resultados obtenidos se determinó la segmentación de los datos en tres grupos. Ejecutado el código para el método K-means en R Studio (ver script en Figura B del Apéndice) para k=3. La Figura 4 muestra la conformación de los clústeres. Los 499 alumnos quedan divididos en tres grupos, el primer clúster formado por 323, el segundo por 56 y el último por 120 alumnos.

Figura 4

K-means para k=3



Fuente: Elaboración propia realizada en R Studio.

Una desventaja de este método de agrupación es su sensibilidad a valores atípicos. Frente a esta situación se prueba una nueva clasificación empleando otro método de clustering más robusto. el K-medoids (PAM). Este enfoque es menos sensible a los valores atípicos y proporciona una alternativa más sólida a K-means.

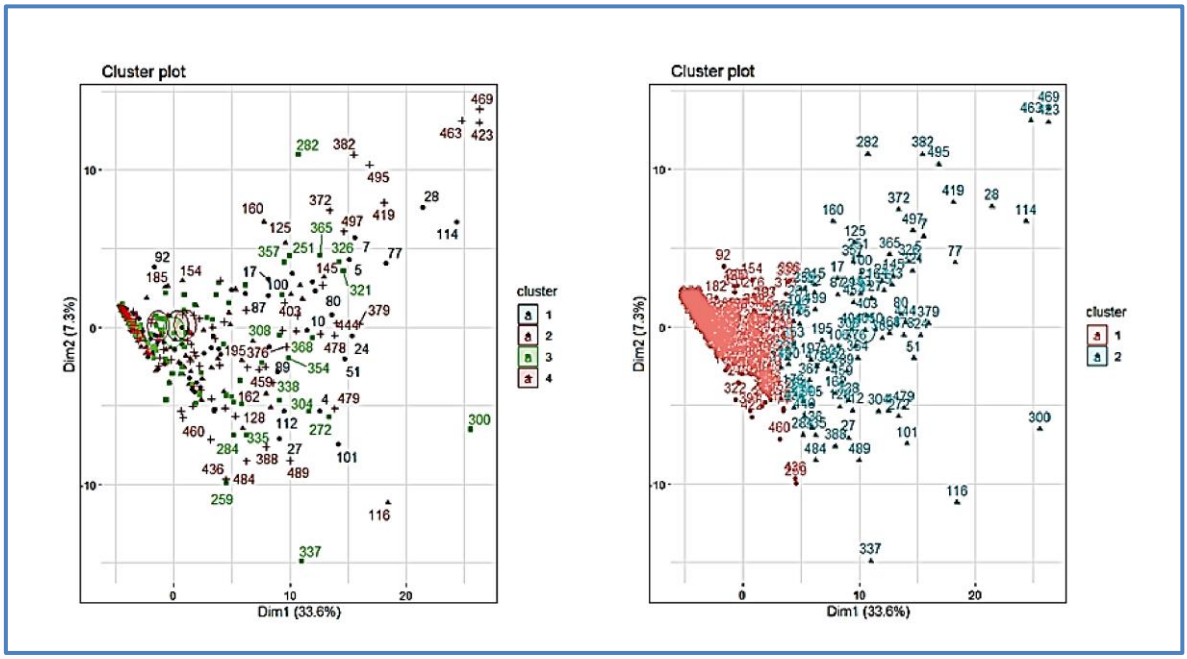
A diferencia del algoritmo K-means, en el que se minimiza la suma total de cuadrados intra-clúster (suma de las distancias al cuadrado de cada observación respecto a su centroide), el algoritmo PAM minimiza la suma de las diferencias de cada observación respecto a su medoid.

Al igual que K-means, necesita que se especifique de antemano el número de clústers que se van a crear. Los criterios empleados en K-means para identificar el numero óptimo de k, pueden aplicarse en K-medoids.

En función que el conjunto de datos es el mismo no es necesario analizar el número óptimo de clústers, se analizó la clusterización para k=2 y k=4. La Figura 5 muestra el resultado del agrupamiento para los valores de k indicados.

Figura 5

K-Medoids (PAM) para k=4 y k= 2



Fuente: Elaboración propia realizada en R Studio.

Finalmente se aplicó el método CLARA que combina la idea de K-medoids con el resampling y que soporta grandes volúmenes de información.

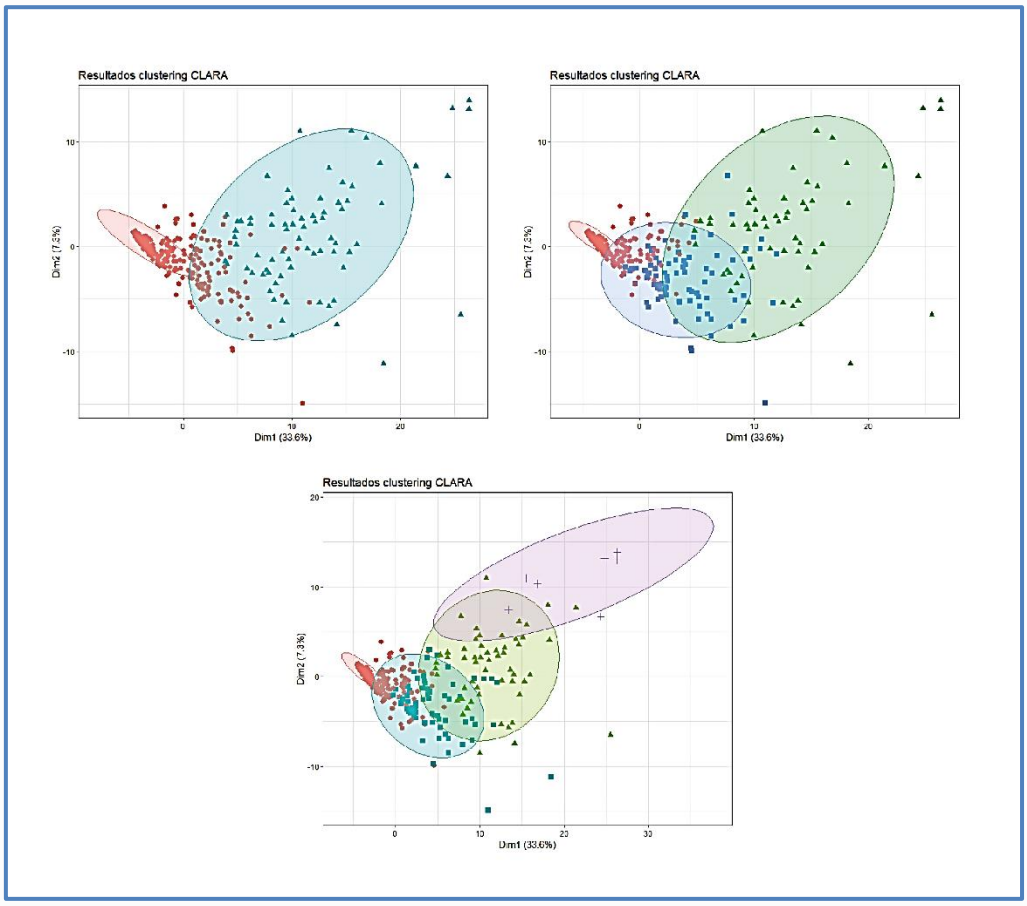
CLARA selecciona una muestra aleatoria de un tamaño determinado y le aplica el algoritmo PAM para encontrar los clústers óptimos acorde a esa muestra. Utilizando esos medoids se agrupan las observaciones de todo el set de datos. La calidad de los medoids resultantes se cuantifica con la suma total de las distancias entre cada observación del set de datos y su correspondiente medoid (suma total de distancias intra-clúster).

Se seleccionan como clústers finales los obtenidos con aquellos medoids que han conseguido menor suma total de distancias (Kaufman y Rousseeuw, 1990).

Se analizó el algoritmo CLARA con k=2, k=3 y k=4 (ver código en Figura C del Apéndice). En la Figura 6 se muestran los resultados, donde puede concluirse que la segmentación más adecuada es la que propone 4 grupos.

Figura 6

CLARA para *para k=2, k= 3 y k=4*



Fuente: Elaboración propia realizada en R Studio.

A modo de resumen, la segmentación definida por los algoritmos y que se utilizó como base para la interpretación de los perfiles, es de tres grupos para el método *K-means*, dos para el algoritmo PAM y cuatro para el CLARA.

Para cerrar este apartado, resta analizar el patrón de comportamiento de los alumnos en cada grupo en función del uso de los recursos en el aula. De manera de poder establecer perfiles que puedan ser empleados como insumo para determinar modelos predictores del rendimiento académico. 1.3 Evaluación de los clústers. Determinación de perfiles

Para poder analizar los resultados de la clusterización realizada se le asignó a cada alumno su grupo de pertenencia, agregando a la base de datos una columna con los resultados de la clasificación del algoritmo K-means, del algoritmo PAM y del algoritmo CLARA.

La Tabla 4 muestra el tamaño de los clústers según el método.

Tabla 4

Tamaño de clúster según método de clasificación

Método	Clúster 1	Clúster 2	Clúster 3	Clúster 4
k-means	323	56	120	-
PAM	400	99	,	,
CLARA	369	52	71	7

Fuente: Elaboración propia

Una característica que se observa en la tabla 4 es que en los tres métodos uno de los clústeres agrupa un alto porcentaje de alumnos, en K-means (65%), en PAM (80%) y en CLARA (74%) del total de los participantes, infiriendo un posible patrón de comportamiento común.

Para poder evaluar la existencia de características similares, es decir un perfil de alumno por clúster se tomaron en cuenta junto con los resultados de la clusterización, las dos variables categóricas Condición de Regularización y Rendimiento académico que no formaron parte de la segmentación por no ser variables numéricas.

La Tabla 5 resume la información en términos de cantidad de integrantes por clúster, según los resultados arrojados por el algoritmo K-means para k=3.

Tabla 5

Alumnos por clúster según Regularización y Rendimiento académico (K-means)

Cond_Regularización	Rendimiento_Acad	Clúster 1	Clúster 2	Clúster 3
AUSENTE	ABANDONÓ	182	9	27
INSUFICIENTE	NO SATISFACTORIO*	52	10	17
	NO SATISFACTORIO	33	10	27
	POCO SATISFACTORIO	35	15	37
REGULAR	SATISFACTORIO	17	7	8
	MUY SATISFACTORIO	4	3	4
	EXCELENTE	0	2	0
TOTAL		323	56	120

Nota: La categoría NO SATISFACTORIO* representa los alumnos que no pudieron regularizar la asignatura, mientras que la categoría NO SATISFACTORIO los que habiendo regularizado no aprobaron el examen final.
Fuente: Elaboración Propia

A partir de la información de la Tabla 5, y reagrupando los alumnos en tres subgrupos dentro de cada clúster. Las dos primeras categorías ABANDONÓ Y NO SATISFACTORIO* como alumnos con un rendimiento malo, las categorías NO SATISFACTORIO Y POCO SATISFACTORIO como alumnos con un rendimiento regular y finalmente las tres últimas SATISFACTORIO, MUY SATISFACTORIO Y EXCELENTE con rendimiento bueno.

A partir de lo anterior se pudo realizar una primera lectura , respecto al método K-means. Dentro del primer clúster el 72% de los alumnos que forman parte de él abandonaron la materia o no pudieron regularizarla, el segundo clúster comparado con los otros dos es el que presenta mayor porcentaje (21%) de los alumnos con un nivel satisfactorio o superior y en el tercer clúster el 53% de los alumnos que pertenecen a él se caracterizan por un nivel regular, es decir, no pudieron aprobar el examen o lo hicieron con la mínima calificación.

La Tabla 6 resume información según los resultados de arrojados por el algoritmo PAM para k=2, dónde se pueden observar la cantidad de alumnos en cada grupo.

Con relación al uso de los recursos por parte de los alumnos de cada clúster, cabe realizar algunas consideraciones, por ejemplo, para los alumnos del primer clúster el número de lecturas por alumno del material práctico fue muy bajo, menos 2 vistas por recurso. Los alumnos del segundo clúster tuvieron una mayor proporción de vistas comprendidas entre 3 y 10 vistas por material y en el tercer clúster nuevamente bajó el número de vistas a 3 o menos por recurso. De igual modo fue el comportamiento vinculado al uso de los recursos multimedia.

Tabla 6

Alumnos por clúster según Regularización y Rendimiento académico (PAM)

Cond_Regularización	Rendimiento_Acad	Clúster 1	Clúster 2
AUSENTE	ABANDONÓ	203	15
INSUFICIENTE	NO SATISFACTORIO*	62	17
	NO SATISFACTORIO	47	23
	POCO SATISFACTORIO	57	30
REGULAR	SATISFACTORIO	23	9
	MUY SATISFACTORIO	8	3
	EXCELENTE	0	2
TOTAL		400	99

Nota: La categoría NO SATISFACTORIO* representa los alumnos que no pudieron regularizar la asignatura, mientras que la categoría NO SATISFACTORIO los que habiendo regularizado no aprobaron el examen final.
Fuente: Elaboración Propia

Compilando la información de la Tabla 6 como se mencionó para K-means, la lectura de la misma permite concluir que el agrupamiento que realiza el método PAM el 66% de los alumnos que forman parte del primer clúster abandonaron la materia o no pudieron regularizarla. En el segundo clúster el 53% de los alumnos que forman parte de él tienen regularizada la materia. No se aconseja está separación dado que no permite establecer conclusiones sólidas sobre el comportamiento de los estudiantes.

Con relación al uso de los recursos por parte de los alumnos de cada clúster, por ejemplo, para los alumnos del primer clúster el número de lecturas por alumno del material práctico fue muy bajo, 3 o menos vistas por recurso. Los alumnos del segundo clúster tuvieron una proporción mucho mayor de vistas comprendidas más de 3 por material, llegando en muchos casos superar las 20 vistas. De igual modo fue el comportamiento vinculado al uso de los recursos multimedia.

Finalmente, la Tabla 7 resume según los resultados de arrojados por el algoritmo CLARA para k=4.

Tabla 7

Alumnos por clúster según Regularización y Rendimiento académico(CLARA)

Cond_Regularización	Rendimiento_Acad	Clúster 1	Clúster 2	Clúster 3	Clúster 4
AUSENTE	ABANDONÓ	196	7	14	1
INSUFICIENTE	NO SATISFACTORIO*	60	6	12	1
	NO SATISFACTORIO	39	10	18	3
	POCO SATISFACTORIO	50	17	19	1
REGULAR	SATISFACTORIO	19	8	5	0
	MUY SATISFACTORIO	5	2	3	1
	EXCELENTE	0	2	0	0
TOTAL		369	52	71	7

Nota: La categoría NO SATISFACTORIO* representa los alumnos que no pudieron regularizar la asignatura, mientras que la categoría NO SATISFACTORIO los que habiendo regularizado no aprobaron el examen final.
Fuente: Elaboración Propia según resultados obtenidos en R Studio.

De la Tabla 7 se puede interpretar que el método CLARA reconoce del mismo que los dos métodos anteriores en el primer clúster a los alumnos que en su mayoría abandonaron la cursada o no pudieron cumplir con los requisitos de regularización. o si regularizaron su rendimiento es insuficiente, en este caso son el 69% de los alumnos que forman parte del primer clúster. En el segundo y tercer grupo hay predominio de alumnos con rendimiento

regular. Sin embargo, el segundo clúster presenta la mayor proporción de alumnos de buen rendimiento con relación a los otros tres grupos. En términos de regularización es el clúster que reúne al mayor porcentaje de regularizados.

En el tercer clúster, hay menor porcentaje de alumnos con buen rendimiento y mayor porcentaje de alumnos que abandonaron o no pudieron regularizar. En comparación el rendimiento de los estudiantes del segundo clúster es superior a los del tercero.

Los alumnos seleccionados en el cuarto y último clúster se destacan por un nivel de rendimiento regular. El porcentaje de alumnos con este rendimiento constituye el 57% de sus integrantes. Es el mayor comparado con la proporción de ese nivel de rendimiento en los restantes clústers definidos.

Con relación al uso de los recursos por parte de los alumnos de los clústers determinados por el método, cabe mencionar, por ejemplo, para los alumnos del primer clúster el número de lecturas por alumno del material práctico fue muy bajo, menos 2 vistas por recurso. Los alumnos del segundo clúster tuvieron una mayor proporción de vistas comprendidas pero inferior a 5 por material , en el tercer clúster el número de vistas en promedio fue superior a 6 por recurso y en el cuarto clúster descendió en promedio a menos de 4 vistas.. De igual modo fue el comportamiento vinculado al uso de los recursos multimedia.

Hasta aquí, el análisis respecto a la segmentación de los alumnos en relación con su nivel de accesos a los recursos y el rendimiento, en el próximo apartado se complementa el análisis haciendo foco en los niveles de interacción en términos de la comunicación asincrónica.

2. Análisis de la comunicación asincrónica en el Aula Virtual

La interacción es el aspecto central de toda experiencia educativa, sobre todo cuando se intenta promover el desarrollo del pensamiento crítico y reflexivo mediante diversas estrategias, con el fin de que la comunicación sea sistemática y estructurada. Barberà, Badia y Mominó (2001) la definen por como un conjunto de reacciones interconectadas entre los miembros que participan en un determinado contexto educativo, en el que la actividad cognitiva humana se desarrolla en función de los elementos que determina la naturaleza de ese contexto educativo, en nuestro caso virtual. La interacción es entendida como un discurso que facilita los procesos de enseñanza-aprendizaje, con una orientación hacia la construcción social del conocimiento.

Medir, examinar la interacción es una tarea compleja, porque se trata de una variable latente, integrada por varios constructos y no puede siempre observarse de manera directa. Sin embargo, se tratará de cuantificarla a partir del número de vistas y mensajes de los alumnos según los registros de la Plataforma Moodle.

En este apartado se analiza el grado de participación de los estudiantes en el aula virtual en los foros de comunicación asincrónica, entendidos los mensajes como intervenciones en un marco académico, para la construcción y difusión del conocimiento. Intentando determinar el impacto que tiene el uso de los foros en el rendimiento académico.

Para el logro del objetivo de este apartado nuevamente se trabajará con técnicas de la minería de datos educacional, el preprocesando los datos para poder emplear las distintas técnicas de clustering.

Técnicas de clustering para el análisis de la comunicación en foros

Los datos utilizados forman parte de la base que ya se describió en el apartado 1.1.

En esta oportunidad del total de variables definidas en la Tabla 1, se consideran solo las que describen el nivel de participación en los 22 foros de comunicación abiertos, el número de accesos al archivo correspondiente a Política de Foros, dónde se establecen las reglas de participación en ellos, y las que refieren al número de mensajes enviados por cada estudiante.

Cabe destacar que de los 22 foros generales dos de ellos, el Foro de Novedades y el de Orientaciones Administrativas, su fin primordial es informar noticias sobre aspectos importantes durante el desarrollo de la asignatura, no constituyen un espacio en el cuál la interacción con los alumnos se destaque, pero si es importante su visualización. Los restantes 20 foros están vinculados a los contenidos teórico/prácticos del bloque temático en el que fueron abiertos. para el planteo de consultas.

2.1 Preprocesado de los datos. Selección de variables

Consecuencia de las actividades de los estudiantes en los espacios virtuales de comunicación, y a partir de las funcionalidades del sistema Moodle, se obtuvieron por cada estudiante datos relativos al número de mensajes leídos, mensajes enviados y número de accesos al archivo Política de Foros.

A efectos de reducir la dimensionalidad de los datos y tomando en consideración que en muchos de los foros el número de vistas y/o mensajes es nulo. Se derivaron los siguientes atributos, en cada nuevo atributo el valor representa la suma del número de vistas a los foros según el siguiente detalle: VF_N (Foro Novedades y Orientaciones Administrativas),

VF_ACM (Foros de Acompañamiento académico) y VF_U1 a VF_U5 (Foros temáticos de las unidades del programa).

De igual modo, pero atendiendo al número de mensajes enviados a cada uno de los foros ya mencionados se generaron los siguientes atributos: MF_N, MF_ACM, MF_U1 a MF_U5.

Se agrega por último al conjunto de datos dos nuevas variables, VT_F (número total de vistas) y MT_F (número total de mensajes enviados).

La base sobre la que se efectuó el análisis está compuesta por las 39 variables definidas en los dos párrafos anteriores.

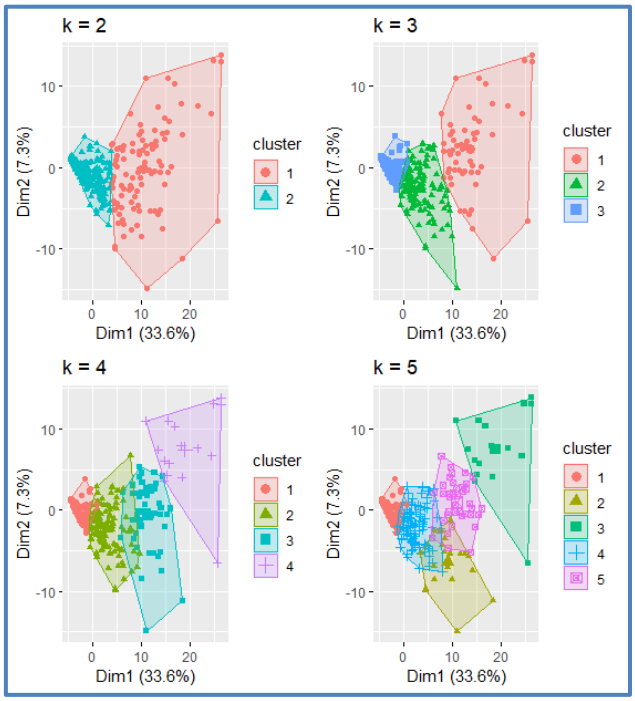
Para poder aplicar las técnicas de clustering y como ya se refirió en el apartado 1.1 se realizaron procesos de transformación de las variables, al ser todas las que intervienen en esta parte del trabajo numéricas se normalizaron para poder aplicar K-means.

2.2 Segmentación en función a la comunicación en foros

La Figura 7 muestra la segmentación inicial para valores de k arbitrariamente elegidos, permitiendo una primera aproximación visual al número de grupos a determinar. De la misma, puede advertirse que la segmentación en dos o tres grupos pareciera ser la adecuada para lograr grupos no solapados.

Figura 7

Clusterización mediante K-means para distintos valores de k



Fuente: Elaboración propia realizada en R Studio.

Durante el análisis del nuevo conjunto de datos se conservó la estructuración realizada en el apartado 1.2. Luego de la ejecución del código correspondiente a los criterios de optimización del número de clústers ya mencionados, ninguno de los tres arrojó un valor de k común, por lo que se realizó el análisis con el valor de k sugerido por el método del Codo que es k=3 , y por el de la Silueta que propone como óptimo k=2.

Elegidos los valores de k a utilizar para la segmentación de los datos se aplican los algoritmos para poder determinar cuál de los tres logra un agrupamiento que permita detectar comportamientos similares entre los alumnos. La Figura 8 muestra la segmentación del conjunto de datos en dos grupos aplicando el Método de K-means, el de PAM y el método CLARA, respectivamente.

Figura 8

Segmentación de los datos en dos clústeres. Métodos K-means-PAM y CLARA

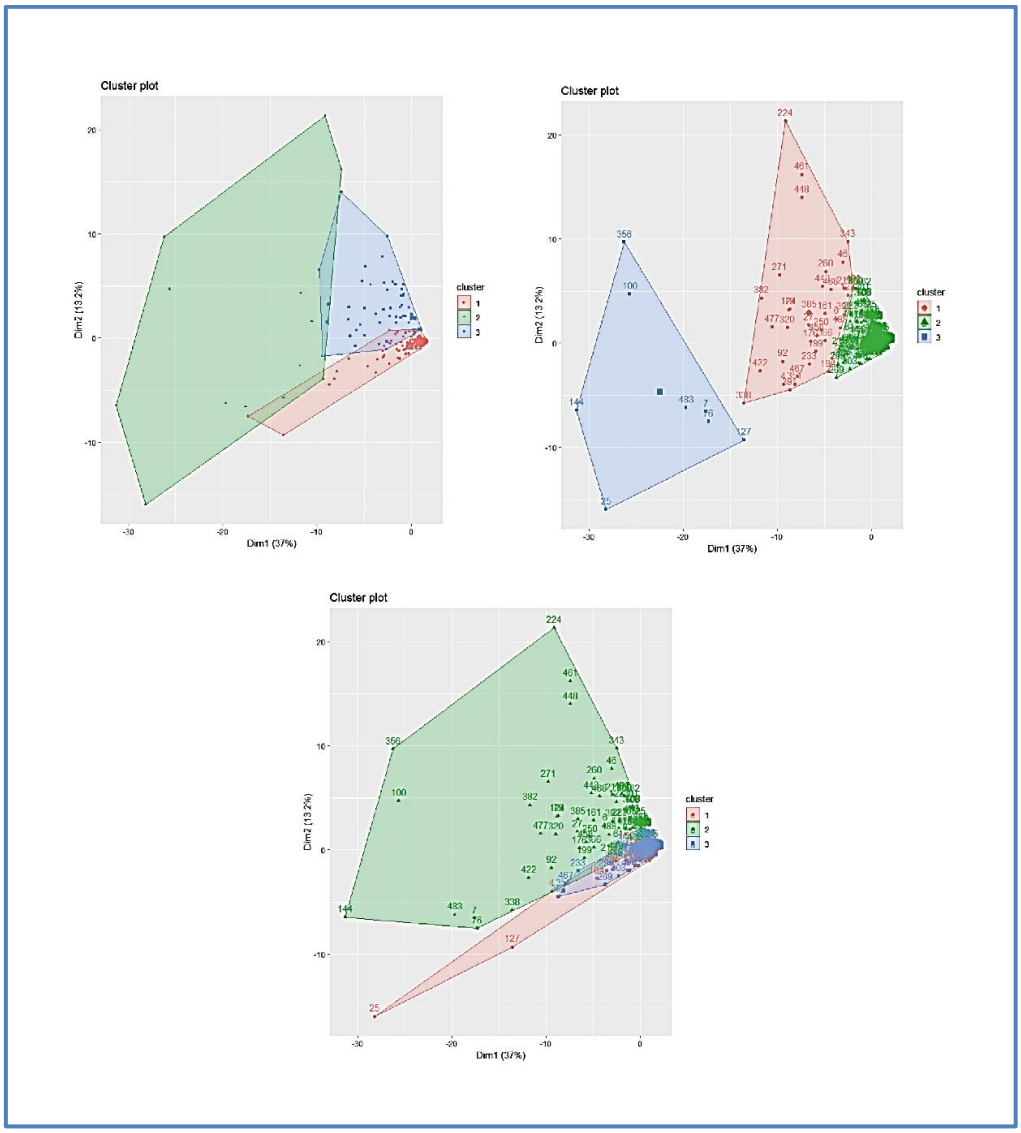
Fuente: Elaboración propia realizada en R Studio.

Para k = 2 la segmentación más adecuada por la distancia de separación de los clústers es la que ofrece el Método K-means. Este método funciona mejor en conjuntos de datos pequeños porque itera sobre todos los datos.

En la Figura 9 se muestra la segmentación con los mismos métodos, pero para tres grupos. Igual conclusión que para de k=2 se pudo extraer respecto al algoritmo que mejor separa al conjunto de datos. De los tres métodos de clustering se optó para analizar el perfil de interacción de los estudiantes exclusivamente según la agrupación determinada por el algoritmo K-means.

Figura 9

Segmentación de los datos en tres clústers. Métodos CLARA- K-means y PAM



Fuente: Elaboración propia realizada en R Studio.

2.3 Evaluación de los clústers. Determinación de perfiles.

Para realizar el análisis de los resultados de la segmentación se agrega a la base de datos una columna con la asignación al clúster de cada alumno, según corresponda. El tamaño de los clústeres determinados en cada caso se detalla en la Tabla 8.

Tabla 8

Tamaño de clúster según método de clasificación

Método	Clúster 1	Clúster 2	Clúster 3
k-means	31	468	-
k-means	8	454	37

Fuente: Elaboración propia realizada según los resultados arrojados por R Studio.

En los grupos definidos, nuevamente hay una tendencia de comportamiento similar que el algoritmo detecta como sucedió cuando se analizó el uso de los recursos, es notorio que en un clúster se agrupe por similitud de comportamiento al 94% de los alumnos, cuando se trabajó sólo con dos grupos y al 91% en el caso de tres grupos. Nuevamente es posible detectar en los alumnos que forman parte de esos grupos un tipo de comportamiento común con respecto al nivel de interacción.

A efectos de poder detectar esas similitudes dentro de cada clúster, y siguiendo el mismo criterio del Apartado 1.3 se tomó en cuenta junto con los resultados de la clusterización, las dos variables categóricas Condición de Regularización y Rendimiento académico que no formaron parte de la segmentación por no ser variables numéricas.

La Tabla 9 resume la información en términos de cantidad de integrantes por clúster, según los resultados de arrojados por el algoritmo k-means para k=2

Tabla 9

Alumnos por clúster según Regularización y Rendimiento académico (K-means)

Cond_Regularización	Rendimiento_Acad	Clúster 1	Clúster 2
AUSENTE	ABANDONÓ	8	210
INSUFICIENTE	NO SATISFACTORRIO	4	75
	NO SATISFACTORIO	5	80
	POCO SATISFACTORIO	7	65
REGULAR	SATISFACTORIO	5	27
	MUY SATISFACTORIO	1	10
	EXCELENTE	1	1
TOTAL		31	468

Fuente: Elaboración propia realizada según los resultados arrojados por R Studio.

De la Tabla 9, se puede comprobar la marcada diferencia en la dimensión de un clúster y el otro, está división tan evidente permite inferir conductas sobre el grado de participación en los foros que repercute en el rendimiento.

Refuerza esto, el porcentaje de integrantes por categorías. Manteniendo el criterio de agrupación de las categorías del Rendimiento académico ya mencionado en la Tabla 6.

El 61% de los integrantes del segundo clúster abandonaron o no pudieron regularizar mientras que el 61% de los integrantes del primer clúster regularizaron la asignatura.

Con respecto a la interacción con los foros temáticos por unidad, los alumnos del primer clúster tuvieron casi nula participación, a diferencia del segundo clúster. Sin embargo, ello no se reflejó en el rendimiento final de la materia.

La Tabla 10 resume la información en términos de cantidad de integrantes por clúster, según los resultados de arrojados por el algoritmo k-means para k=3. Como se desprende de su lectura el algoritmo detecta un grado de similitud fuerte en la forma de interacción de los alumnos que conforman el segundo clúster.

Tabla 10

Alumnos por clúster según Regularización y Rendimiento académico (K-means)

Cond_Regularización	Rendimiento_Acad	Clúster 1	Clúster 2	Clúster 3
AUSENTE	ABANDONÓ	2	204	12
INSUFICIENTE	NO SATISFACTORRIO	1	73	5
	NO SATISFACTORIO	1	76	4
	POCO SATISFACTORIO	1	65	9
REGULAR	SATISFACTORIO	2	25	6
	MUY SATISFACTORIO	1	10	1
	EXCELENTE	0	1	1
TOTAL		8	454	38

Fuente: Elaboración propia realizada según los resultados arrojados por R Studio.

Como en los métodos anteriores hay un comportamiento generalizado que fue detectado por K-means para la conformación del segundo clúster, el 61% de sus integrantes de este abandonaron o no pudieron regularizar. mientras que el 37,5% de los integrantes del primer clúster se caracterizan por un rendimiento regular y el 21% de los integrantes del tercero lo hacen por tener un rendimiento mejor que los alumnos que forman los otros grupos en esa categoría.

2.3 Interpretación de los resultados

Una primera conclusión que podría extraerse de la segmentación en dos grupos es que en el primero, si bien el número de integrantes es pequeño, su participación en foros provocó una mejora en su rendimiento académico, el 61% de los alumnos de este grupo pudieron regularizar la materia cursada en forma virtual, una diferencia porcentual mucho mayor al 39% de los integrantes del segundo grupo.

El segundo clúster presenta como característica más destacada el porcentaje de abandono o no regularización de la materia.

Respecto a la segmentación en tres grupos la disparidad de tamaños hace que no sea posible detectar más que una tendencia respecto a la escasa o nula participación en la comunicación y que ello se vea reflejado en el alto porcentaje de abandono o falta de regularización del clúster con más integrantes.

Con respecto a la interacción con los foros temáticos por unidad, los alumnos del primer clúster tuvieron casi nula participación, a excepción de tres integrantes con un alto grado de visualización en cada uno de ellos foro, en el segundo clúster el nivel de participación fue mucho mayor, con un promedio entre 3 y 10 vistas por foro, pero aquí tampoco esa participación se reflejó en el rendimiento final de la materia. Respecto a la participación de los alumnos del tercer grupo, hay un dispar nivel de accesos, cuatro integrantes con un alto grado de participación en todos los foros y por el otro lado, los restantes sólo presentaron al menos 1 visita a algunos de los foros abiertos por unidad.

Hasta aquí, el análisis respecto a la segmentación de los alumnos en relación con su nivel de interacción en los foros de comunicación.

3. Factores predictivos del rendimiento académico

El rendimiento académico, por ser multicausal, envuelve una enorme capacidad explicativa de los distintos factores y espacios temporales que intervienen en el proceso de aprendizaje. Existen diferentes aspectos que se asocian al rendimiento académico, entre los que intervienen componentes tanto internos como externos al individuo.

A partir de la información obtenida en los dos apartados anteriores se trata de determinar qué modelo predictivo de la Minería de Datos educacional permite identificar aquellos factores que resultan mejores predictores del rendimiento académico en este particular contexto educativo de emergencia académica de virtualidad obligatoria.

Se aplicó al conjunto de datos, el filtrado de variables, Las técnicas de selección de atributos ayudan a identificar los más significativos, es decir aquellos que aporten el mayor poder predictivo en el modelo. Además, cuando existen muchas características, ayudan a reducir la complejidad de los modelos mejorando su comprensión . Disminuyen la necesidad de más espacio dealmacenamiento y acortan el tiempo para el entrenamiento y el procesamiento. De igual forma, la recopilación de nuevos casos se simplifica al ignorar aquellos atributos que no tienen potencial predictivo (Márquez-Vera, Romero y Ventura, 2012).

Luego del proceso de selección de atributos, la base que originalmente contaba con 149 variables quedo reducida a 131, se eliminaron atributos que por contener mayoría de registros nulos no aportaban capacidad predictiva. Definida la base de trabajo se aplicaron diferentes modelos predictivos en RapidMiner a fin de detectar cual de ellos ofrece mejores resultados.

Se ha optado por aplicar algoritmos de Regresión Logística Multinomial, Árboles de Decisión y el ensamble Gradient Boosted Tree, que permiten variables categóricas tanto en las variables independientes (atributos) como en la dependiente (variable objetivo).

De este modo se busca cumplir el objetivo general planteado para poder encontrar un modelo predictor del Rendimiento académico a partir de los factores relacionados a patrones de comportamiento en la interacción en el aula virtual.

3.1 Evaluación de Modelos Predictivos: Regresión Logística Multinomial

De las técnicas de la minería de datos, las supervisadas o predictivas, utilizan todas o algunas de variables o campos en una base de datos para predecir valores desconocidos o futuros de otra. Para la aplicación de los modelos predictivos se empleó la herramienta de minería de datos RapidMiner. La que posibilita el desarrollo de procesos de análisis de datos mediante el encadenamiento de operadores a través de un entorno gráfico.

RapidMiner contiene técnicas de preprocesamiento de datos, modelación predictiva y descriptiva, métodos de entrenamiento y prueba de modelos, visualización de datos y aprendizaje automático (Beltrán, 2010).

Para evaluar el rendimiento en clasificación de los modelos, se consideraron las métricas Confusion Matrix (Matriz de Confusión) , Precision (Precisión), Recall (Exhaustividad), Accuracy (Exactitud) y Especificidad.

Matriz de Confusión (MC): Una matriz cuadrada de orden igual al número de categorías de

la variable a predecir. Las columnas muestran los valores reales (positivos y negativos) y las filas los valores estimados (positivos y negativos). De la lectura de la matriz se pueden conocer el número de casos que han sido correctamente clasificados (Verdaderos Positivos), ubicados en la diagonal principal de la matriz. Por fuera de ella, los falsos negativos ,la predicción es negativa cuando el valor tendría que ser positivo y los falsos positivos , la predicción es positiva cuando realmente el valor tendría que ser negativo.

La Sensibilidad o Exhaustividad (recall), la Precisión (precision) , la Especificidad y la Exactitud (Accuracy) del modelo son los indicadores utilizados a la hora de evaluar el modelo predictivo. La exhaustividad mide la cantidad que el modelo es capaz de identificar, la precisión mide la calidad del modelo , la especificidad mide la capacidad del modelo para detectar verdaderos negativos y la exactitud mide las predicciones correctas respecto del total (Moya, 2020).

Si lo que interesa es identificar los verdaderos negativos, (evitar falsos positivos) se debe elegir especificidad alta. Si lo que interesa es evitar falsos negativos, se debe elegir sensibilidad alta.

Considerando que el modelo de clasificación presenta seis categorías (N=6) , la MC es una matriz cuadrada de orden 6. Para calcular las métricas correspondientes para su evaluación, se calcularon las siguientes medidas.

La ecuación 1 permite calcular la precisión (Precision) para una etiqueta que es el cociente entre los casos positivos bien clasificados por el modelo y el total de predicciones positivas.

$$\text{Precisión}_{\text{clase k}}\% = \frac{\text{MC}_{(\text{k;k})}}{\sum_{i=1,j=k}^{N} MC(i;j)} \times 100 \qquad (1)$$

Dónde el numerador de la expresión MC(k; k) representa la casilla de la matriz en donde converge la estimación del modelo con la realidad para esa etiqueta *k*. El denominador, es la sumatoria total de elementos identificados como positivos por el clasificador para esa etiqueta k, que incluye todos los Falsos Positivos (FP) y su Verdadero Positivo (TP) respectivamente.

La ecuación 2 permite calcular la exhaustividad o sensibilidad (*Recall*) que representa la tasa de verdaderos positivos. Proporción entre los casos positivos bien clasificados por el modelo, respecto al total de positivos

$$\text{Exhaustividad}_{\text{clase k}}\% = \frac{\text{MC}_{(k;k)}}{\sum_{j=1,i=k}^{N} MC(i;j)} \times 100 \qquad (2)$$

La ecuación 3 permite calcular la exactitud (*Accuracy*) que es el cociente entre los casos bien clasificados por el modelo (verdaderos positivos y verdaderos negativos, es decir, los valores en la diagonal de la matriz de confusión), y la suma de todos los casos.

$$Exactitud_{modelo}\% = \frac{\sum_{i=1}^{N} \text{MC}_{(i;i)}}{\sum_{\substack{1\leq i \leq N \\ 1\leq j\leq N}} MC(i,j)} \times 100 \qquad (3)$$

Modelo de Regresión Logística Multinomial

El primer modelo predictivo aplicado al conjunto de datos es el de regresión logística multinomial, ya que se desea determinar la relación entre una variable dependiente cualitativa con seis categorías, y las variables explicativas independientes.

Este modelo se aplica sobre el conjunto de datos preprocesado. El modelo se construyó usando el operador de validación cruzada, que es una de las formas más consistentes de evaluar clasificadores, ya que los conjuntos se determinan de manera aleatoria y el error de la evaluación es muy bajo. Comparado con los demás métodos, es la validación más estricta para verificar la precisión de los modelos, esto asegura su capacidad de generalización. Con un grupo de entrenamiento y otro de testeo. Se dividió al conjunto de datos en training y testing usando Split Data al 80 y 20%. Para este modelo se emplea la extensión WEKA para RapidMiner, ya que la herramienta no ofrece por sí misma el algoritmo de regresión logística multinomial (W-Logistic). (Ver Figura D y E del Apéndice con el proceso del modelo). La Tabla 11 muestra la Matriz de Confusión del W Logistic , la que permite calcular la Precisión, la Exhaustividad por clase y el Accuracy del modelo según las ecuaciones 1, 2 y 3.

Tabla 11

Matriz de Confusión del Modelo W Logistic

ACCURACY 72,17%

	True A	True S	True E	True PS	True NS	True MS	CLASS PRECISION
Pred. A	198	3	0	4	10	0	91,71%
Pred. S	3	14	0	6	5	5	42,42%
Pred. E	1	0	0	0	1	0	0,00%
Pred. PS	14	12	1	47	23	4	46;53%
Pred. NS	10	1	0	27	109	0	74,15%
Pred. MS	2	2	1	3	1	2	18,18%
CLASS RECALL	86,24%	43,75%	0,00%	54,02%	73,15%	18,18%	

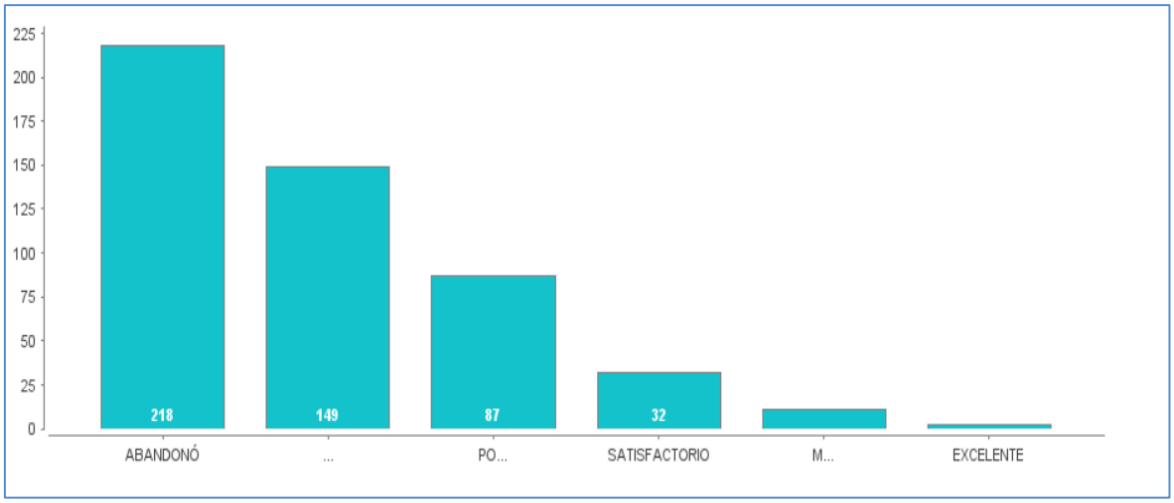
Nota: A: ABANDONÓ, S:SATISFACTORIO, E:EXCELENTE, PS:POCO SATISFACTORIO, NS: NO SATISFACTORIO, MS: MUY SATISFACTORIO
Fuente: Elaboración propia realizada según los resultados arrojados por RapidMiner.

En la Figura 10 se muestran las frecuencias de observación de cada categoría de la variable dependiente, las categorías ABANDONÓ y NO SATISFACTORIO son las que presentan la mayor frecuencia de observación y ello se evidencia en el alto porcentaje de precisión del modelo para esas categorías (91,71% y 74,15%, respectivamente).Se puede concluir que se debe a que el conjunto de atributos independientes permite discernir de una forma clara el tipo de clasificación. El entrenamiento del algoritmo en esas categorías es mejor que en las demás dada su mayor frecuencia.

En cuanto a las restantes categorías, el número de casos no permite al algoritmo el entrenamiento necesario para determinar cómo se comporta un estudiante, fundamentalmente en las categorías del máximo nivel de rendimiento como lo son las tres últimas. De hecho, puede observarse en la Tabla 11 que el modelo no pudo clasificar correctamente a alumnos en la categoría Excelente.

Figura 10

Histograma variable Rendimiento académico



Fuente: Elaboración propia realizada según los resultados arrojados por RapidMiner.

Con relación a la performance del modelo, el Accuracy o Exactitud (72,17%) es aceptable, pero debe ser analizado con precaución ya que las clases se encuentran muy desbalanceadas como se ve en la Figura 13. Esta situación hace que se haya considerado evaluar otros modelos predictivos.

3.2 Evaluación de Modelos Predictivos: Decision Tree y Gradient Boosted Tree

A través de la aplicación de árboles de decisión se busca mejorar la capacidad predictiva del modelo anterior. Los árboles son útiles para explorar un conjunto de datos y entender cómo ciertas variables de las interacciones de los estudiantes con el entorno virtual de aprendizaje inciden sobre otra.

Permiten una organización eficiente del conjunto de datos, debido a que los árboles son construidos a partir de la evaluación del primer nodo raíz y de acuerdo con su evaluación o valor tomado se va descendiendo en las ramas hasta llegar al final del camino u hojas del árbol (Jaramillo,2015). La calidad del árbol depende de la precisión de la clasificación y del tamaño del árbol (Chen, Han y Yu, 1996).

El modelo se construye nuevamente con el operador de validación cruzada, manteniendo el grupo de entrenamiento, y el operador Decision Tree con parámetros optimizados en máxima profundidad en 7 y el criterio gain_ratio (Ver Figura D y F del Apéndice) .

La Tabla 12 muestra la Matriz de Confusión del modelo de la que se puede leer las métricas Precision , Recall para cada una de las categorías. El Accuracy del modelo fue de 96,39% mejorando la exactitud del modelo de regresión logística que fue de 72,17%.

Tabla 12

Matriz de Confusión del Modelo Decision Tree

ACCURACY 96,39%

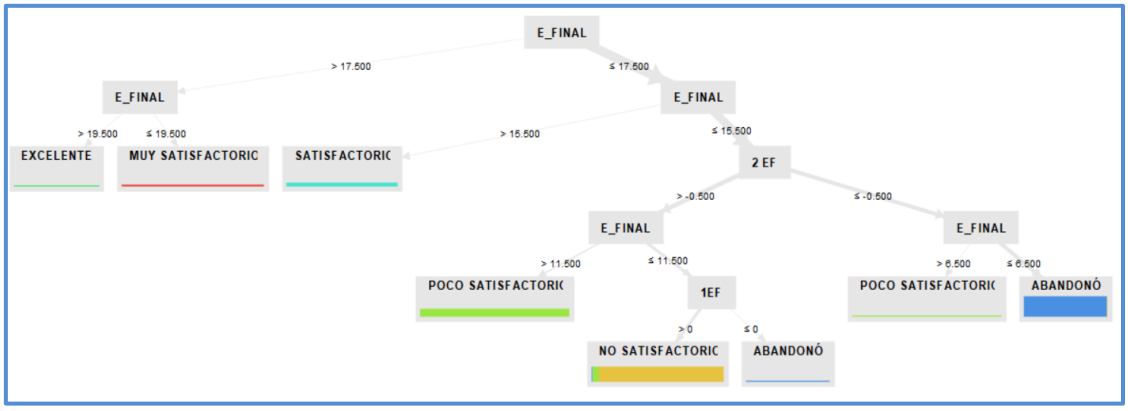
	True A	True S	True E	True PS	True NS	True MS	CLASS PRECISION
Pred. A	216	0	0	2	3	0	97,74%
Pred. S	0	32	0	0	0	0	100%
Pred. E	0	0	0	0	0	1	0,00%
Pred. PS	0	0	0	77	0	0	100%
Pred. NS	2	0	0	8	146	0	93,59%
Pred. MS	0	0	2	0	0	10	83,33%
CLASS RECALL	99,08%	100%	0,00%	88,51%	97,99%	90,91%	

Nota: A: ABANDONÓ, S:SATISFACTORIO, E:EXCELENTE, PS:POCO SATISFACTORIO, NS: NO SATISFACTORIO, MS: MUY SATISFACTORIO
Fuente: Elaboración propia realizada según los resultados arrojados por RapidMiner.

El árbol de decisión encuentra reglas mucho más determinantes a la hora de predecir los ABANDONÓ con número muy bajo de falsos positivos (97,74% de precisión). En comparación con la regresión logística multinomial clasifica mejor a las restantes categorías de la variable. En la Figura 11 se observa el Árbol obtenido, se ver que la categoría E_ Final (número de respuestas correctas en el examen final) es el Nodo principal de todas las ramas.

Figura 11

Árbol de Decisión



Fuente: Elaboración propia realizada según los resultados arrojados por RapidMiner.

Finalmente se aplica un ensamble Gradient Boosted Tree, para evaluar la predicción del Rendimiento academico. Usando los parametros optimizados con 30 árboles, máxima profundidad 7 y learning rate en 0,01 (Ver Figura D y G del Apéndice).

El modelo está formado por un ensamble de árboles entrenados de forma secuencial, cada nuevo árbol emplea información del árbol anterior para aprender, mejorando iteración a iteración. Al tratarse de métodos no paramétricos, no es necesario que se cumpla ningún tipo de distribución específica.

Los árboles pueden, en teoría, manejar tanto predictores numéricos como categóricos sin tener que crear variables dummy.

Por lo general, requieren mucha menos limpieza y preprocesado de los datos en comparación a otros métodos de aprendizaje estadístico (por ejemplo, no requieren estandarización). No se ven muy influenciados por *outliers* (James, 2013).

La Tabla 13 muestra la Matriz de Confusión del modelo de la que se pueden ver las métricas Precision, Recall para cada categoría de la variable Rendimiento Académico y el Accuracy.

Tabla 13

Matriz de Confusión del Modelo Gradient Boosted Tree

ACCURACY 98,20%

	True A	True S	True E	True PS	True NS	True MS	CLASS PRECISION
Pred. A	218	0	0	1	0	0	99,54%
Pred. S	0	30	2	0	0	3	85,71%
Pred. E	0	0	0	1	0	0	0,00%
Pred. PS	0	0	0	85	0	0	100%
Pred. NS	2	0	0	0	149	0	100%
Pred. MS	0	2	0	0	0	8	80,0%
CLASS RECALL	100%	100%	0,00%	97,70%	100%	72,73%	

Nota: A: ABANDONÓ, S:SATISFACTORIO, E:EXCELENTE, PS:POCO SATISFACTORIO, NS: NO SATISFACTORIO, MS: MUY SATISFACTORIO
Fuente: Elaboración propia realizada según los resultados arrojados por RapidMiner.

La exactitud del modelo fue del 98,20% mejorando el Accuracy del modelo de regresión logística multinomial que era de 72,17% y del modelo de árbol de decisión 96,39%.

El ensamble Gradient Boosted Tree mejoró la predicción del rendimiento académico en comparación con los modelos anteriores.

3.3 Evaluación del mejor modelo para predecir el Rendimiento académico

La Tabla 14 muestra los resultados de las métricas que se tomaron en cuenta para evaluar los modelos generados: instancias clasificadas correctamente (Accuracy), instancias clasificadas incorrectamente (Clasification_error), estadística de Kappa que mide la coincidencia de la predicción con la clase real (Kappa).

Tabla 14

Métricas de los modelos predictivos empleados

	W Logistic	Decision Tree	Gradient Boost Tree(GBT)
Accuracy	72,17%	96,39%	98,20%
Clasification_error	27,83%	3,61%	1,8%
Kappa	0,599	0,947	0,973

Nota: Resultados obtenidos de los algoritmos usando RapidMiner
Fuente: Elaboración propia

De la tabla 14, se observa que el modelo *Gradient Boosted Tree* es el que tiene el mayor porcentaje de instancias correctamente clasificadas por sobre los dos modelos, lo que hace que el número de instancias clasificadas incorrectamente sea el menor. De todas las métricas analizados el modelo GBT es el que mejor resultados presenta.

Tomando como modelo predictivo el GBT los atributos que más peso tienen para la predicción del rendimiento académico son la variable Condición_Reg que refiere a la condición de regularización de la cursada, la variable 2EF que representa el número de respuestas correctas en la segunda evaluación formativa, la variable InfoE_FI que refiere al número de visitas al archivo información sobre Examen Final, dónde se explicitan las condiciones de aprobación y estructura del examen final , la variable E_Final que representa el número de respuestas correctas en el Examen Final y la variable VT_F que resume el número de vistas totales a los foros (Ver Figura H, Apéndice).

Los resultados empíricos indican que el modelo Gradient Boosted Trees, obtuvo altos valores de exactitud, por lo que el modelo predictivo es efectivo para su aplicación en el contexto académico. A partir de la definición de las variables importantes según se observa en la Figura H del Apéndice. Se empleó este conjunto de variables para generar un modelo confiable de predicción del rendimiento académico.

4. Conclusiones

Como se declaró en la Introducción del trabajo su objetivo principal fue la determinación de la existencia de variables relacionadas con la participación en el aula virtual que actuaran como predictoras del rendimiento académico, todo a partir de la información que brinda la plataforma virtual del curso.

Las plataformas virtuales multiplicaron los datos que se manejaban en el entorno académico. Las interacciones alumno–plataforma suponen una rica fuente de información para analizar el comportamiento de los estudiantes y la efectividad de la plataforma para aplicar acciones correctivas que mejoren su rendimiento. Para la realización de este trabajo se ha considerado la información disponible desde de la plataforma *Moodle* a través del diseño de un itinerario de análisis completo de la participación de los alumnos en el aula virtual de Álgebra durante el Primer cuatrimestre del 2021.

La tarea de recopilación y procesamiento de datos desarrollada en el primer apartado fue crucial para obtener información de calidad y atributos relevantes para ser utilizados en las tareas de la minería de datos educativa. El método empleado para seleccionar variables permitió reducir la cantidad de atributos significativos asociados a los alumnos en estudio, la reducción de complejidad y la mejora en la calidad del modelo predictivo.

Se pudo comprobar a lo largo del desarrollo del trabajo, que las diversas técnicas de la minería de datos pudieron aplicarse de manera flexible y práctica para seleccionar subconjuntos de atributos significativos y con ellos aplicar algoritmos de agrupamiento y clasificación. Los resultados en los niveles de precisión y confiabilidad de los algoritmos indican la posibilidad de emplearlos de manera efectiva, con el propósito de predecir el rendimiento académico de los estudiantes de la materia analizada.

Se sabe que la determinación del rendimiento académico es un problema multifactorial en donde confluyen una gran cantidad de variables, no sólo de orden académico, sino del contexto sociodemográfico y de aspectos cognitivos e interpersonales.

En esta oportunidad, solo se contemplaron las variables que tienen que ver con el nivel de participación en el aula virtual.

Los resultados de la segmentación en clústers, trabajada en los dos primeros apartados, permitieron contrastar la hipótesis general que establecía que a mayor participación mejor sería el rendimiento académico. No existe evidencia significativa que permita el rechazo de ésta. La evidencia marcó que un nivel bajo de interacción condujo al abandono de la cursada o a no poder cumplir con las condiciones de regularización y al mismo tiempo

evidenció que a mayor participación las condiciones de rendimiento mejoraron.

No obstante, y como ponen de manifestó otros estudios (Moral de la Rubia, 2006), hay factores adicionales que contribuyen a explicar los logros académicos que se relacionan con características aptitudinales o de personalidad, como pueden ser los estilos de aprendizaje (Shaw, 2012), que no se han considerado en este trabajo. La inclusión de variables adicionales en posteriores estudios contribuirá, probablemente, a un mayor poder predictivo.

El conocimiento adquirido con los modelos predictivos permite concluir, que desde las cátedras hay que atender aspectos relacionados con la participación de los estudiantes en la comunicación en foros y en mejorar la implicación en el cursado de la materia. La baja capacidad predictiva de las variables vinculadas con la interacción en foros y el uso de los recursos pedagógicos revela un aspecto a tomar en cuenta desde lo pedagógico. Se deben reforzar las acciones que deriven en un mayor grado de participación e involucramiento en el proceso de aprendizaje. Sin dejar de considerar que la situación de emergencia académica atravesada por los estudiantes repercutió también sobre esas acciones. En muchos casos los alumnos que formaron parte de esta investigación iniciaron su formación académica en tiempos de pandemia lo que les impidió una construcción adecuada de su rol de alumno universitario.

Si bien, para este trabajo se realizó un recorte temporal de un cuatrimestre y una Sede del Ciclo Básico Común (CBC) en el marco de un particular contexto sanitario. Se espera que lo actuado contribuya no sólo a la Cátedra a la que los alumnos pertenecen, sino que pueda en un futuro cercano extenderse a la totalidad de las cinco Cátedras a cargo de la materia y a las correspondientes Sedes en las que se dictan, incrementando el universo de alumnos y por consiguiente, el volumen de información para mejorar la precisión de los modelos predictivos y poder generalizar los resultados.

Se espera que los hallazgos del presente trabajo sirvan también al Departamento de Matemática de la Facultad de Ciencias Económicas para replicar el trabajo en otras asignaturas y conocer cómo actúan los alumnos en la plataforma y disponer de un modelo de predicción del rendimiento que permita orientar las acciones tendientes a aplicar las mejoras necesarias para mantener la calidad del proceso de enseñanza aprendizaje.

5. Referencias Bibliográficas

Aldas Manzano, J., y Uriel Jiménez, E. (2017). *Análisis multivariante aplicado con R*. Ediciones Paraninfo, S.A.

Aldowah, H., Al-Samarraie, H., y Fauzy, W. M. (2019*). Educational data mining and learning analytics for 21st century higher education: A review and synthesis*. *Telematics and Informatics*, *37*, 13-49.

Anoopkumar, M., y Rahman, A. M. Z. (2016). *A Review on Data Mining techniques and factors used in Educational Data Mining to predict student amelioration.* In *2016* International Conference on Data Mining and Advanced Computing (SAPIENCE)

Ayala Franco, E., López Martínez, R.E. y Menéndez Domínguez, V.H. (2021). Modelos predictivos de riesgo académico en carreras de computación con minería de datos educativos. RED. Revista de Educación a Distancia, 21(66). https://doi.org/10.6018/red.463561

Barberà, E. (coord.), Badia, A. y Mominó, J. Ma. (2001), *La incógnita de la educación a distanci*a, Cuadernos de Educación, núm. 35, Horsori.

Beltrán, D. Poveda, D.: *RAPIDMINER*, Universidad Nacional de Colombia- Facultad de Ciencias Económicas - Unidad de Informática y Comunicaciones, http://www.fce.unal.edu.co/uifce/pdf/Rapid_Miner.pdf 2010

Campbell, J. P., DeBlois, P.B., y Oblinger, D. G. (2007). *Analítica académica: Una nueva herramienta para una nueva era*. Revisión EDUCAUSE, *42*(4), 40.

Chen, M. S., Han, J., y Yu, P. S. (1996). *Minería de datos: una visión general desde la perspectiva de la base de datos*. IEEE Transactions on Knowledge and data Engineering, *8*(6), 866-883.

James, G., Witten, D., Hastie, T., y Tibshirani, R. (2013). *Una introducción al aprendizaje estadístico* (Vol. 112, p. 18). Springer.

Jaramillo, A., y Arias, H. P. P. (2015). *Aplicación de Técnicas de Minería de Datos para Determinar las Interacciones de los Estudiantes en un Entorno Virtual de Aprendizaje*. Revista Tecnológica-ESPOL, *28*(1).

Johnson, D.F. (1998) *Applied multivariate methods for data analysts* Brookd/Cole.

MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observation.* 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press.

Márquez-Vera, C., Cano, A., Romero, C., y Ventura, S. (2013). *Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data*. Applied intelligence, *38*(3), 315-330.

Moral de la Rubia, J. (2006). *Predicción del rendimiento académico universitario*.

Moya Pérez, Carlos (2015). *Estudio de patrones de interacción entre los estudiantes y la Plataforma de Tele-Enseñanza en la UPM*. Proyecto Fin de Carrera / Trabajo Fin de Grado, E.T.S.I. y Sistemas de Telecomunicación (UPM),

Park, H. S., & Jun, C. H. (2009). *Un algoritmo simple y rápido para la agrupación de K-medoids*. Sistemas expertos con aplicaciones, *36*(2), 3336-3341. *Perfiles Educativos, 28*(113), 38-63.

Rodriguez Almeida, A.y da Silva Camargo, S . (2015) *Academic Analytics: Aplicando técnicas de Business Intelligence sobre datos de performance académica en enseñanza superior*. Interfaces Científicas - Exactas e Tecnológicas. ISSN ELETRÔNICO - 2359-4942, vol. 1, nº 2, pp. 35-46, 2015.

Rousseeuw, P. J., Kaufman, L., & Trauwaert, E. (1996). *Agrupamiento difuso mediante matrices de dispersión*. Estadística computacional y análisis de datos, *23*(1), 135-151.

Sharma, S. (1996). *Applied multivariate techniques*. John Wiley &Sons,1era Edition

Shaw, R. S. (2012). *A study of the relationships among learning styles, participation types, and performance in programming language learning supported by on- line forums.* Computers & Education, *58*(1), 111-120.

Apéndices

Figura A

Script R para la evaluación del número óptimo de k

```
library(openxlsx)
#leemos la Base de Datos
A3<-read.xlsx("Datos.xlsx",colNames=TRUE,sheet=1)
#Inspeccionamos los nombres de las variables
names(A3)
na.omit(A3)
A_3<-scale(A3)
summary(A3)
A3
install.packages("ggplot2")
install.packages("factoextra")

library(cluster)
library(factoextra)


fviz_nbclust(A3,kmeans,method = "wss")
fviz_nbclust(A3,kmeans,method = "gap_stat")
fviz_nbclust(A3,kmeans,method = "silhouette")

km.means <- kmeans (A3, 3, nstart = 50, algorithm = "Lloyd")
fviz_cluster(km.means, data = A3)
```

Fuente: Elaboración propia con referencia al Script brindado desde el Taller de Programación dictado en la Especialización (Del Rosso, 2020).

Figura B

Determinación de clúster para distintos valores de k usando el Método k-means

```
DF <- B3
# Remove any missing value (i.e, NA values for not available)
DF <- na.omit(DF)
# Scale variables
DF <- scale(DF)
# View the firt 3 rows
head(DF, n = 3)
View(DF)

df_1
summary(df_1)
df_1$size
df_1 <- na.omit(df_1)
# Scale variables
df_1<- scale(df_1)


library(cluster)
library(factoextra)

res.dist <- get_dist(df_1, stand = TRUE, method = "pearson")
fviz_dist(res.dist,
          gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07"))
k2 <- kmeans(df_1, centers = 2, nstart = 25)
str(k2)
k2
fviz_cluster(k2, data = df_1)
k3 <- kmeans(df_1, centers = 3, nstart = 25)
k4 <- kmeans(df_1, centers = 4, nstart = 25)
k5 <- kmeans(df_1, centers = 5, nstart = 25)
k2 <- kmeans(df_1, centers = 2, nstart = 25)


# plots to compare
p1 <- fviz_cluster(k2, geom = "point", data = DF) + ggtitle("k = 2")
p2 <- fviz_cluster(k3, geom = "point",  data = DF) + ggtitle("k = 3")
p3 <- fviz_cluster(k4, geom = "point",  data = DF) + ggtitle("k = 4")
p4 <- fviz_cluster(k5, geom = "point",  data = DF) + ggtitle("k = 5")
```

Fuente: Elaboración propia con referencia al Script de RPubs K-Means Clustering Tutorial(2020).

Figura C

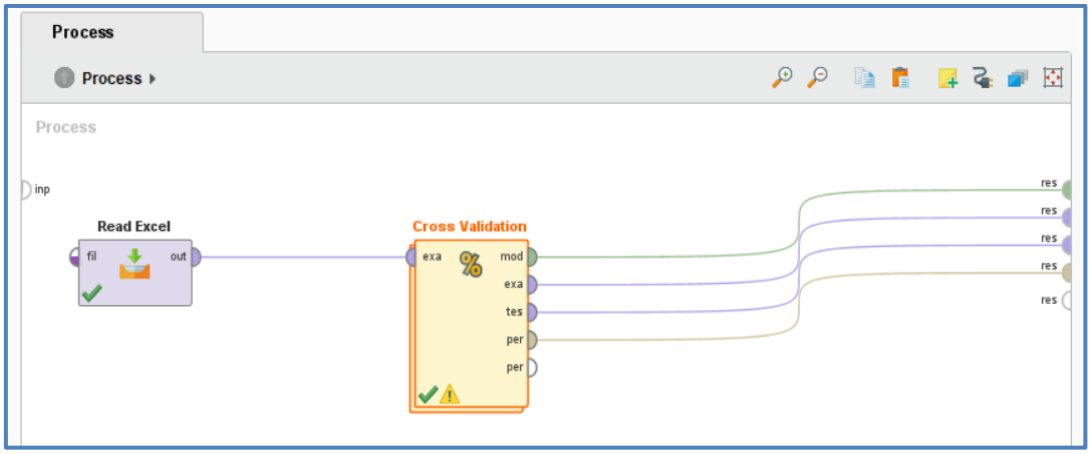
Script de R para el método PAM y CLARA

```
#Si no está instalada, instalo openxlsx
#install.packages("openxlsx")
library(openxlsx)
#leemos la Base de Datos
A_1_1<-read.xlsx("Datos.xlsx",colNames=TRUE,sheet=1)
#Inspeccionamos los nombres de las variables
names(A_1_1)

library(cluster)
library(factoextra)
fviz_nbclust(x = A_1_1, FUNcluster = pam, method = "wss", k.max = 15,
             diss = dist(A_1_1, method = "manhattan"))

set.seed(123)
pam_clusters <- pam(x = A_1_1, k = 2, metric = "manhattan")
pam_clusters
fviz_cluster(object = pam_clusters, data = A_1_1, ellipse.type = "t",
             repel = TRUE) +  theme_bw() +  labs(title = "Resultados clustering PAM") +  theme(legend.position = "none")

library(cluster)
library(factoextra)
clara_clusters <- clara(x = A_1_1, k = 2, metric = "manhattan", stand = TRUE,
                        samples = 50, pamLike = TRUE)
clara_clusters
fviz_cluster(object = clara_clusters, ellipse.type = "t", geom = "point",
             pointsize = 2.5) +  theme_bw() +  labs(title = "Resultados clustering CLARA") +  theme(legend.position = "none")
library(cluster)
library(factoextra)
clara_clusters <- clara(x = A_1_1, k = 3, metric = "manhattan", stand = TRUE,
                        samples = 50, pamLike = TRUE)
clara_clusters
fviz_cluster(object = clara_clusters, ellipse.type = "t", geom = "point",
             pointsize = 2.5) +  theme_bw() +  labs(title = "Resultados clustering CLARA") +  theme(legend.position = "none")

library(cluster)
```

Fuente: Elaboración propia con referencia al Script brindado desde el Taller de Programación dictado en la Especialización (Del Rosso, 2020).

Figura D

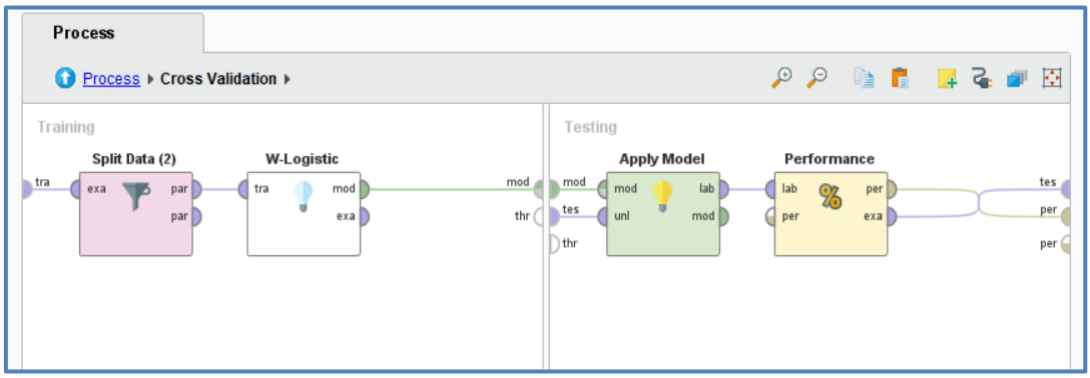
Proceso de RapidMiner para la Validación cruzada y evaluación de los Modelos



Fuente: Elaboración propia con referencia al modelo brindado desde la materia Fundamentos de Métodos Analíticos Predictivos dictada en la Especialización (Abalde, 2021).

Figura E

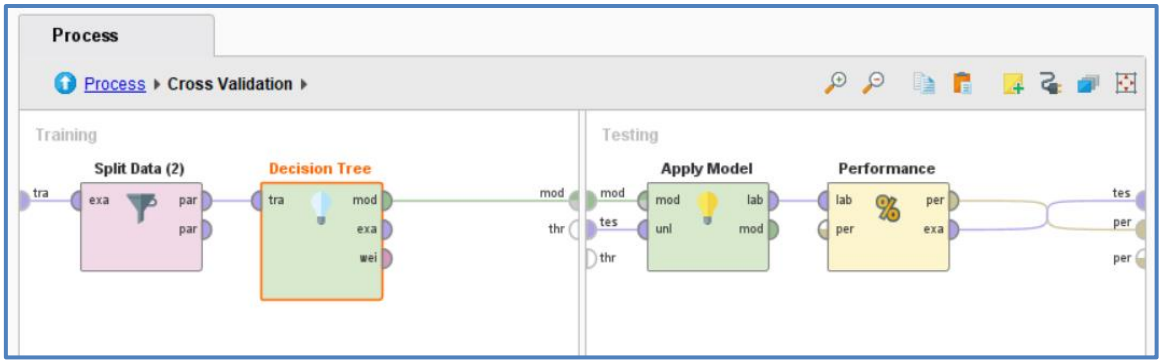
Proceso de RapidMiner para la Validación cruzada y evaluación W-logistic



Fuente: Elaboración propia con referencia al modelo brindado desde la materia Fundamentos de Métodos Analíticos Predictivos dictado en la Especialización (Abalde, 2021).

Figura F

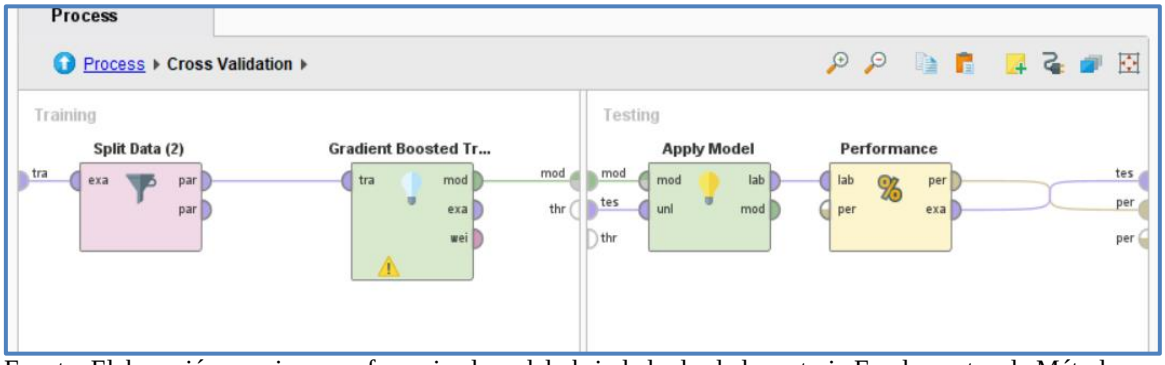
Proceso de RapidMiner para la Validación cruzada y evaluación Decision Tree



Fuente: Elaboración propia con referencia al modelo brindado desde la materia Fundamentos de Métodos Analíticos Predictivos dictada en la Especialización (Abalde, 2021).

Figura G

Proceso de RapidMiner para la Validación cruzada y evaluación Gradient Boosted Tree



Fuente: Elaboración propia con referencia al modelo brindado desde la materia Fundamentos de Métodos Analíticos Predictivos dictada en la Especialización (Abalde, 2021).

Figura H

Modelo Gradient Boosted Tree. Performance Vector

```
PerformanceVector:
accuracy: 98.20% +/- 2.57% (micro average: 98.20%)
ConfusionMatrix:
True:   ABANDONÓ        SATISFACTORIO   EXCELENTE       POCO SATISFACTORIO      NO SATISFACTORIO        MUY SATISFACTORIO
ABANDONÓ:       218     0       0       1       0       0
SATISFACTORIO:  0       30      2       0       0       3
EXCELENTE:      0       0       0       1       0       0
POCO SATISFACTORIO:     0       0       0       85      0       0
NO SATISFACTORIO:       0       0       0       0       149     0
MUY SATISFACTORIO:      0       2       0       0       0       8
classification_error: 1.80% +/- 2.57% (micro average: 1.80%)
ConfusionMatrix:
True:   ABANDONÓ        SATISFACTORIO   EXCELENTE       POCO SATISFACTORIO      NO SATISFACTORIO        MUY SATISFACTORIO
ABANDONÓ:       218     0       0       1       0       0
SATISFACTORIO:  0       30      2       0       0       3
EXCELENTE:      0       0       0       1       0       0
POCO SATISFACTORIO:     0       0       0       85      0       0
NO SATISFACTORIO:       0       0       0       0       149     0
MUY SATISFACTORIO:      0       2       0       0       0       8
kappa: 0.974 +/- 0.037 (micro average: 0.974)
```

Gradient Boosted Model

```
Model Metrics Type: Multinomial
 Description: N/A
 model id: rm-h2o-model-gradient_boosted_trees-1620
 frame id: rm-h2o-frame-gradient_boosted_trees-1620
 MSE: 0.22022614
 RMSE: 0.4692826
 R^2: 0.9345149
 logloss: 0.63474685
 mean_per_class_error: 0.16666667
 hit ratios: [0.995, 1.0, 1.0, 1.0, 1.0, 1.0]
 CM: Confusion Matrix (Row labels: Actual class; Column labels: Predicted class):
                ABANDONÓ  SATISFACTORIO  EXCELENTE  POCO SATISFACTORIO  NO SATISFACTORIO  MUY SATISFACTORIO   Error      Rate
      ABANDONÓ       174              0          0                   0                 0                  0  0.0000   0 / 174
 SATISFACTORIO         0             26          0                   0                 0                  0  0.0000    0 / 26
     EXCELENTE         0              0          0                   0                 0                  2  1.0000     2 / 2
POCO SATISFACTORIO     0              0          0                  70                 0                  0  0.0000    0 / 70
  NO SATISFACTORIO     0              0          0                   0               119                  0  0.0000   0 / 119
MUY SATISFACTORIO      0              0          0                   0                 0                  9  0.0000     0 / 9
        Totals       174             26          0                  70               119                 11  0.0050   2 / 400
Variable Importances:
  Variable Relative Importance Scaled Importance Percentage
   E_FINAL          4703.904297          1.000000   0.576059
   Cursada          3447.737305          0.732952   0.422224
      2 EF             5.318538          0.001131   0.000651
    VID24B             2.525497          0.000537   0.000309
      VE.3             1.542913          0.000328   0.000189
      VG.2             1.408873          0.000300   0.000173
     VF_U1             0.619791          0.000132   0.000076
       1.0             0.586942          0.000125   0.000072
      VJ.4             0.585810          0.000125   0.000072
      VA.3             0.292758          0.000062   0.000036
---
```

Fuente: Elaboración propia con referencia al modelo brindado desde la materia Fundamentos de Métodos Analíticos Predictivos dictada en la Especialización (Abalde, 2021).

Buenos Aires, 30 de noviembre de 2021

Reporte Trabajo Final Integrador de Especialización

"Modelos predictivos del rendimiento académico en Álgebra a través de la minería de datos educativa.

Estudio de Patrones de Interacción en la Plataforma Moodle del Campus Virtual CBC"

Autora: Lic. Andrea Leonor Gache

En mi carácter de mentora del Trabajo Final Integrador de Especialización de la Lic. Andrea Gache, me dirijo a ustedes con el objetivo de transmitirle mi opinión sobre la investigación realizada.

Destaco que el tema en análisis aborda una problemática relevante para la Especialización, siendo los objetivos, tanto el general como los específicos, claros y coherentes con la hipótesis propuesta. La bibliografía es pertinente y actual.

Desarrolla en su trabajo aspectos teóricos que dan marco a su investigación, analizando las estrategias de participación en el Campus Virtual asumidas por los alumnos de Algebra del Primer Tramo del Ciclo General de la Facultad de Ciencias Económicas en el contexto del dictado virtual de emergencia debido a la pandemia y su influencia en el rendimiento académico

Articula su investigación con contenidos importantes de las asignaturas Taller de Programación (uso de RStudio), Fundamentos de Métodos Analíticos Predictivos (Modelos predictivos) y Métodos de Análisis Multivariado (Clustering).

La coherencia del enfoque planteado, el uso de algoritmos de Minería de Datos, la pertinencia de las referencias bibliográficas y la correcta redacción permiten señalar a este trabajo como un aporte relevante del tema indicado en el título del trabajo, evidenciando una capacidad de emprender líneas de investigación futuras.

Por estas razones doy mi expresa conformidad a la entrega de este Trabajo Final Integrador de Especialización

Sin otro particular, saludo cordialmente.

Dra. María José Bianco