ISBN: 978-628-95207-4-3. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2023.1.1.1355

Segmentation of University Students using Clustering and considering a Virtual Cycle

Edson Nicks Lazaro-Camasca, Estudiante de Ciencia de la Computación[1], Yuri Nuñez-Medrano, MSc. en Ingeniería de Sistemas[2]
[1,2] *Universidad Nacional de Ingeniería, Perú*, [1]*elazaroc@uni.pe*, [2]*ynunezm@uni.edu.pe*

***Abstract*– The present work had as purpose the creation of Grouping or Clustering Models for each specialty that the National University of Engineering has, with the purpose of carrying out tutorials focused on students who obtained low academic performance both in the face-to-face period 2019-1 and in the virtual period 2020-1.**

In this research, the CRISP-DM methodology was used to create the models, in addition, the necessary theory of Clustering is developed. The data used was purely academic. During the data analysis process, some important findings were found, such as the significant improvement in qualifications after taking a virtual cycle. The models used were K-means, DBSCAN, Affinity Propagation and MeanShift.

To find the optimal parameters of each model in each specialty, parameter searches were carried out using the Silhouette Coefficient as the Performance Metric. Finally, to choose the best model, the Success Metric was defined as the number of groups generated by each model among students with low academic performance, then it was compared between the models, and it was found that the Affinity Propagation model is the one that performs better for the application of focused tutorials.

***Keywords*— Clustering, CRISP-DM, K-means, Academic Qualifications, DBSCAN, Affinity Propagation, MeanShift, Virtual Cycle.**

Digital Object Identifier: (only for full papers, inserted by LACCEI).
ISSN, ISBN: (to be inserted by LACCEI).
DO NOT REMOVE

Segmentación de Estudiantes Universitarios usando Clustering y considerando un Ciclo Virtual

Edson Nicks Lazaro-Camasca, Estudiante de Ciencia de la Computación[1], Yuri Nuñez-Medrano, MSc. en Ingeniería de Sistemas[2]

[1,2] *Universidad Nacional de Ingeniería, Perú*, [1]*elazaroc@uni.pe*, [2]*ynunezm@uni.edu.pe*

Resumen– ***El presente trabajo tuvo como propósito la creación de Modelos de Agrupamiento o Clustering por cada especialidad que tiene la Universidad Nacional de Ingeniería, con el propósito de realizar tutorías focalizadas en alumnos que obtuvieron bajo rendimiento académico tanto en el periodo presencial 2019-1 como en el periodo virtual 2020-1.***

En esta investigación se usó la metodología CRISP-DM para la creación de los modelos, además se desarrolla la teoría necesaria de Clustering. Los datos que se usaron fueron netamente académicos. Durante el proceso de análisis de datos se encontraron algunos hallazgos importantes tales como la mejora significativa de las calificaciones luego de llevar un ciclo virtual. Los modelos que se usaron fueron K-means, DBSCAN, Propagación de Afinidad y MeanShift.

Para encontrar los parámetros óptimos de cada modelo en cada especialidad se realizó búsquedas de parámetros donde se usó el Coeficiente de Silueta como la Métrica de Rendimiento. Finalmente, para elegir el mejor modelo se definió como Métrica de Éxito a la cantidad de grupos que genera cada modelo dentro de los alumnos con bajo rendimiento académico, luego se comparó entre los modelos y se encontró que el modelo de Propagación de Afinidad es el que se desempeña mejor para la aplicación de tutorías focalizadas.

Keywords-- Clustering, CRISP-DM, K-means, Calificaciones Académicas, DBSCAN, Propagación de Afinidad, MeanShift, Ciclo Virtual.

I. Introducción

La presente investigación tiene una temática aplicada al análisis del rendimiento académico de los alumnos de la Universidad Nacional de Ingeniería y como este rendimiento ha variado en la transición de una educación presencial a una educación virtual a causa de la pandemia del COVID 19. Comprender este cambio haciendo uso de los datos y análisis nos ayuda a mejorar la calidad educativa tanto en ámbito virtual y presencial.

En diciembre del año 2019 apareció un tipo de coronavirus llamado COVID-19 en la provincia de Wuhan en China el cual en un corto lapso de tiempo este se volvió una pandemia, debido a esto muchos países han optado por realizar cuarentenas estrictas como medida de prevención. A finales de febrero del 2020 se registra el primer caso de COVID-19 en América, en el sector educativo se canceló las clases presenciales en todos los niveles ya que es sabido que los centros educativos congregan una gran cantidad de personas y estos podrían ser un foco viral, los países tomaron diferentes respuestas a esta problemática [1]. Entonces la Educación universitaria paso a ser Virtual el 2020.

II. Objetivo

Desarrollar diversos modelos de Aprendizaje No Supervisado específicamente Modelos de Agrupamiento o Clustering capaces de agrupar a los alumnos de la Universidad Nacional de Ingeniería, con estas agrupaciones se pretende identificar alumnos con bajo rendimiento académico.

A. Objetivos Específicos

- Recolectar un dataset de los alumnos a estudiar.
- Identificar y transformas características que serán usados por los modelos.
- Realizar el pre-procesamiento de los datos.
- Establecer los Modelos de Agrupamiento.
- Realizar una búsqueda de los parámetros óptimos de cada Modelo.
- Establecer las Métricas de Éxito.
- Comparar los modelos y obtener la mejor agrupación según la aplicación.

B. Alcances

A continuación, se presentan los alcances de la investigación:

- Se hace uso de modelos de agrupamiento muy usados en la literatura y que se encuentran en la librería Sklearn.
- El desarrollo de scripts para todo el proceso de agrupamiento esta desarrollado usando el lenguaje python.
- El dataset solo contempla récord de calificaciones de los periodos 2019-2 y 2020-1.
- El estudio está enfocado en los alumnos de cada especialidad de la Universidad Nacional de Ingeniería.

III. Marco Teórico

A. Métodos de Agrupamiento

Hay muchos tipos de métodos o algoritmos de agrupación en clústeres. Muchos algoritmos utilizan medidas de similitud o distancia entre ejemplos en el espacio de entidades en un esfuerzo por descubrir regiones densas de observaciones.

Como tal, a menudo es una buena práctica escalar datos antes de usar algoritmos de agrupación en clústeres.

"La noción de la similitud (o diferencia) entre los objetos individuales que se agrupan es fundamental para todos los objetivos del análisis de clústeres. Un método de agrupación en clústeres intenta agrupar los objetos en función de la definición de similitud que se le proporciona "[2].

B. K-means

K-Means Clustering puede ser el algoritmo de agrupación en clústeres más conocido e implica la asignación de ejemplos a clústeres en un esfuerzo por minimizar la varianza dentro de cada clúster.

"El propósito principal de este documento es describir un proceso para dividir una población de N dimensiones en conjuntos k sobre la base de una muestra. El proceso, que se llama 'k-means', parece dar particiones que son razonablemente eficientes en el sentido de varianza dentro de la clase "[3].

El algoritmo KMeans agrupa los datos intentando separar muestras en n grupos de igual varianza, minimizando un criterio conocido como inercia o suma de cuadrados dentro del clúster. Este algoritmo requiere que se especifique el número de clústeres. Se escala bien a un gran número de muestras y se ha utilizado en una amplia gama de áreas de aplicación en muchos campos diferentes.

C. DBSCAN

DBSCAN Clustering (donde DBSCAN es la abreviatura de Clustering espacial basado en densidad de aplicaciones con ruido) implica la búsqueda de áreas de alta densidad en el dominio y la expansión de esas áreas del espacio de entidades a su alrededor como clústeres.

"... presentamos el nuevo algoritmo de agrupación en clústeres DBSCAN que se basa en una noción basada en la densidad de clústeres que está diseñada para descubrir clústeres de forma arbitraria. DBSCAN requiere sólo un parámetro de entrada y apoya al usuario en la determinación de un valor adecuado para él "[3].

El algoritmo DBSCAN ve los clústeres como áreas de alta densidad separadas por áreas de baja densidad. Debido a esta vista bastante genérica, los clústeres encontrados por DBSCAN pueden tener cualquier forma, a diferencia de k-means, lo que supone que los clústeres tienen forma convexa. El componente central del DBSCAN es el concepto de muestras de núcleo, que son muestras que se encuentran en áreas de alta densidad. Por lo tanto, un clúster es un conjunto de muestras de núcleo, cada una cercana entre sí (medida por alguna medida de distancia) y un conjunto de muestras no principales que están cerca de una muestra de núcleo (pero no son muestras de núcleo). Hay dos parámetros para el algoritmo, y que definen formalmente lo que queremos decir cuando decimos denso. Más alto o más bajo indican una mayor densidad necesaria para formar un clúster.

D. Propagación de Afinidad

Propagación de afinidad crea clústeres enviando mensajes entre pares de muestras hasta la convergencia. Los mensajes enviados entre pares representan la idoneidad de que una muestra sea el ejemplo del otro, que se actualiza en respuesta a los valores de otros pares. Esta actualización se realiza de forma iterativa hasta la convergencia, momento en el que se eligen los ejemplos finales y, por lo tanto, se da la agrupación en clúster final [4].

La propagación de afinidad puede ser interesante, ya que elige el número de clústeres en función de los datos proporcionados. Para ello, los dos parámetros importantes son la preferencia, que controla cuántos ejemplares se utilizan, y el factor de amortiguación que amortigua la responsabilidad y los mensajes de disponibilidad para evitar oscilaciones numéricas al actualizar estos mensajes.

El principal inconveniente de Affinity Propagation es su complejidad. El algoritmo tiene una complejidad de tiempo de la orden, donde está el número de muestras y es el número de iteraciones hasta la convergencia. Además, la complejidad de la memoria es del orden si se utiliza una matriz de similitud densa, pero reducible si se utiliza una matriz de similitud dispersa. Esto hace que Affinity Propagation sea más adecuado para conjuntos de datos de tamaño pequeño a mediano.

E. MeanShift clustering

MeanShift clustering tiene como objetivo descubrir blobs en una densidad suave de muestras. Es un algoritmo basado en centroide, que funciona actualizando a los candidatos para que los centroides sean la media de los puntos dentro de una región determinada. Estos candidatos se filtran en una etapa de postprocesamiento para eliminar los casi duplicados para formar el conjunto final de centroides [4].

Dado un centroide candidato para la iteración, el candidato se actualiza de acuerdo con la siguiente ecuación:

$$x_i^{t+1} = m(x_i^t)$$

Dónde está la vecindad de las muestras dentro de una distancia dada alrededor y es el vector de desplazamiento medio que se calcula para cada centroide que apunta hacia una región del aumento máximo en la densidad de puntos. Esto se calcula utilizando la siguiente ecuación, actualizando efectivamente un centroide para que sea la media de las muestras dentro de su vecindario:

$$m(x_i) = \frac{\sum_{x_j \in N(x_i)} k(x_j - x_i) x_j}{\sum_{x_j \in N(x_i)} k(x_j - x_i)}$$

El algoritmo establece automáticamente el número de clústeres, en lugar de depender de un parámetro, que dicta el tamaño de la región que se va a buscar. Este parámetro se puede establecer manualmente, pero se puede estimar

mediante la función proporcionada, a la que se llama si no se establece el ancho de banda [5].

F. Medidas de Rendimiento.

La Medida de Rendimiento es la forma en la que se desea evaluar una solución al problema. Se cubrirán dos métricas:

- Método de Codo.
- Análisis de Silueta.

IV. METODOLOGÍA

Se describe la creación de un Modelo de Aprendizaje No Supervisado que será capaz de agrupar estudiantes de acuerdo con un conjunto de características.

Se realizó la extracción de características, la transformación da datos, la creación de modelos, la validación de los modelos. Se usó la metodología planteada por Jhon Rollins basado en CRISP-DM.

La población en este proyecto está dada por un subconjunto de toda la población universitaria de aproximadamente 12,061 estudiantes, estos han cursado el periodo 2019-II o 2020-I. Por otro lado, se tiene que la muestra, son todos los alumnos que han cursado ambos periodos son aproximadamente 10,394 estudiantes.

Las librerías usadas para el desarrollo son del lenguaje Python; PyPDF2, Pandas, Numpy, Seaborn, Matplotlib, sklearn.preprocessing, sklearn.metrics, sklearn.cluster. KMeans, sklearn.cluster.AffinityPropagation.

V. DESARROLLO

A. Colección de Datos.

El conjunto de datos iniciales para los modelos fue coleccionados a través de los diferentes Departamentos de Estadística de cada Facultad de la Universidad Nacional de Ingeniería. El conjunto de datos consiste de una colección de calificaciones académicas de 12,061 estudiantes del periodo 2019-2 y 11,928 del periodo 2020-1, esta variación en la cantidad de estudiantes se debe a que algunos estudiantes no pudieron llevar el un ciclo 100% virtual por la falta de recursos y otro motivo es que no hubo ingreso de nuevos estudiantes. La tabla 5.1 describe los datos por cada estudiante.

Cuadro 5.1: Descripción del conjunto de datos sin procesar.

Característica	Descripción
Abrev. del Periodo	Es un identificador que consta de un año y un numero de temporada.
Facultad	Dentro de la data existen 11 facultades y cada una está abreviada empezando con una letra.
Especialidad	Estas derivan de las facultades, dentro de la data existen 28 especialidades y cada una de estas tiene la abreviatura de la facultad y termina con un número.
Nombre del curso	Identificador del curso.
Código del curso	Identificador único del curso, consta de un 2 caracteres seguido de 3 números.
Crédito del curso	Peso de un curso basado en el número de horas.
Nota del curso	Calificación del estudiante en un rango de 0 a 20.
Condición	Esta se refiere a una condición de matrícula, M si está matriculado, N si no está matriculado y T si fue retirado.
Situación	Existen diferentes situaciones tales como normal, egresado, suspendido, titulado, expulsado entre otras.
Característica	Descripción
Abrev. del Periodo	Es un identificador que consta de un año y un numero de temporada.
Facultad	Dentro de la data existen 11 facultades y cada una está abreviada empezando con una letra.

Después de recopilar los datos, se realizó una extracción de ciertos campos del formato PDF 5.1 a CSV 5.2, para esto se desarrolló un script en python en donde se realizó lo siguiente:

Primero se extrajo la data en crudo sin ninguna estructura usando la librería PyPDF2.

Se Aplicó expresiones regulares para captar los patrones que tiene la estructura del PDF, para esto se usó la librería re.

La investigación está enfocada a analizar el cambio de rendimiento de los estudiantes al llevar un ciclo virtual es necesario tener datos más consolidados que pueden resumir el estado académico de un estudiante, que se observa en la Tabla 5.2.

Cuadro 5.2: Resumen de Características Académicas

Nueva Característica	Abreviatura
Situación Actual	SA
Condición Actual	CA
Último Periodo Académico	UPA
Facultad	FAC
Especialidad	ESP
Créditos Curriculares Aprobados	CCA
Promedio Ponderado Acumulado	PPA
Ultimo Ciclo Promedio Ponderado	UCLL-PP
Ultimo Ciclo Créditos Llevados	UCLL-CRDL

Se realizaron procesos de depuración de registros que sean considerados como outliers y se eliminan registros en donde de algunas características tengan una cantidad pequeña de datos.

Dentro del análisis se toma a la variable Último Periodo Académico como la característica discriminante, ya que esta nos va ayudar a diferenciar entre los periodos 19-2 y 20-1.

Cuadro 5.3: Tabla comparativa entra la cantidad de alumnos CA por periodo, especialidad y facultad

Id	Abrev.	Facultad	CA 2019-2	CA 2020-1	Id	Especialidad	CA 2019-2	CA 2020-1
A	FAUA	Arquitectura, Urb. y A.	673	753	A1	Arquitectura	673	753
C	FIC	Ing. Civil	1197	1332	C1	Ing. Civil	1197	1332
E	FIEECS	Ing. Económica, Estadística y Ciencias Sociales	720	795	E1	Ing. Económica	505	548
					E3	Ing. Estadística	215	247
					G1	Ing. Geológica	251	285
G	FIGMM	Ing. Geológica, Minera y Metalúrgica	790	884	G2	Ing. Metalúrgica	242	265
					G3	Ing. De Minas	297	334
I	FIIS	Ing. Industrial y de Sistemas	1167	1326	I1	Ing. Industrial	577	653
					I2	Ing. De Sistemas	590	673
L	FIEE	Ing. Eléctrica y Electrónica	1212	1366	L1	Ing. Eléctrica	412	465
					L2	Ing. Electrónica	427	481
					L3	Ing. de Telecom.	373	420
M	FIM	Ing. Mecánica	1368	1507	M3	Ing. Mecánica	397	442
					M4	Ing. Mecánica y Eléctr.	424	466

					M5	Ing. Naval	175	187
					M6	Ing. Mecatrónica	372	412
N	FC	Ciencias	816	933	N1	Física	189	215
					N2	Matemática	140	161
					N3	Química	101	120
					N5	Ing. Física	166	192
					N6	Ciencia de la Comput.	220	245
P	FIPGNP	Ing. de Petróleo, Gas Natural y Petroquímica	238	260	P2	Ing. Petroquímica	130	146
					P3	Ing. de Petróleo y Gas N.	108	114
Q	FIQT	Ing. Química y Textil	691	758	Q1	Ing. Química	573	622
					Q2	Ing. Textil	118	136
S	FIA	Ing. Ambiental	737	842	S1	Ing. Sanitaria	267	301
					S2	Ing. de Higi. y Seg. Ind.	205	242
					S3	Ing. Ambiental	265	299

.

En la figura 5.3 se muestra las abreviaturas por Facultad y Especialidad y se muestra cómo se distribuye la cantidad de alumnos por los periodos de estudio, existen varias especialidades que algunas tienen escasa cantidad de alumnos en comparación con la mayoría, entre estas se tiene a N2, N3, P2, P3, Q2 donde sus valores son menores a 150.

B. Selección de Características.

En esta sección vamos a describir cuales son las características finales que se van usar por los modelos de agrupamiento, según al análisis de características realizados en la sección anterior se obtuvo la tabla 5.4 en donde se menciona las cuatro características extraídas en base a transformación sobre las características base de la tabla 5.2.

Cuadro 5.4: Tabla de características usadas por los modelos de agrupamiento

Índice	Abreviatura	Característica
F1	CRE	Ciclo Relativo hasta el periodo 19-1
F2	PPA	Promedio Ponderado Acumulado hasta periodo el 19-1
F3	SCORE-19	Score del periodo 19-2 (presencial)
F4	SCORE-20	Score del periodo 20-1 (virtual)

Es importante mencionar que las características CRE y PPA nos proporcionan el rendimiento histórico que tiene cada estudiante, esta no contempla los periodos 19-2 y 20-1, mientras que el SCORE nos da el rendimiento de cada periodo sea 19-2 o 20-1.

C. Entrenamiento de los Modelos de Agrupación

El entrenamiento de un modelo consiste en encontrar los parámetros más óptimos de un modelo a través de una búsqueda de parámetros. Los Modelo de Agrupamiento o Clustering tienen sus propios parámetros estos al ser variados generan diferentes tipos de grupos, pero como saber cuál de estos tipos de agrupamiento es el mejor. Además, como saber cuáles son los valores óptimos de esos parámetros. En esta investigación se ha planteado utilizar una Medida de Rendimiento llamada Coeficiente de Silueta, esta medida nos da un valor en cuan mejor se han agrupado los datos, siendo el valor óptimo cuando los grupos se encuentran lo más lejos posible y cada grupo este bien condensado. Desde esta premisa se estable que los valores óptimos de los parámetros son aquellos que generaron el mayor valor del Coeficiente de Silueta.

Dentro de la investigación se realizó una búsqueda de parámetros óptimos para cada tipo de modelo, esta consiste en establecer un espacio de búsqueda discreto y pequeño para cada parámetro, luego se prueba con cada combinación posible y se obtiene el comportamiento de cada parámetro respecto al Coeficiente de Silueta. La búsqueda de los parámetros óptimos se realizó para cada especialidad y el comportamiento de cada parámetro se puede ver más adelante mediante gráficas de líneas de tendencia.

A continuación, se describen la búsqueda de parámetros que se realizó en cada modelo de agrupación y cuales fueron su comportamiento.

K-Means

En este modelo se tiene un parámetro variable llamado “n cluster“ este representa el número de grupos que se debe tener al final, el espacio de búsqueda que tiene este parámetro es el conjunto [2,3,4,5,6,7,8,9]. El comportamiento de este parámetro frente al Coeficiente de Silueta se puede ver la figura 5.1, en estas gráficas el número de grupos se encuentra en el eje “x“, mientras que el eje “y“ está representado por el Coeficiente de Silueta. Analizando estas gráficas se pudo encontrar que a medida que se aumenta los grupos el Coeficiente de Silueta disminuye, también se puede ver que los valores óptimos del parámetro se encuentran entre 2 o 3 grupos. Sin embargo, existen algunas especialidades como I1 y N3 que tiene un comportamiento diferente, a medida que se aumenta el número de grupos el coeficiente de silueta aumenta y posteriormente disminuye generando un parámetro óptimo con valor 5 o 6.

DBSCAN

En este modelo se tiene dos parámetros llamados “eps“ y “min samples“, se muestra a detalles estos parámetros:

- eps: este parámetro es la distancia máxima entre dos muestras para que una se considere próxima a la otra. Este no es un límite máximo en las distancias de los puntos dentro de un grupo. Este es el parámetro DBSCAN más importante para elegir adecuadamente para su conjunto de datos y función de distancia. El espacio de búsqueda de este parámetro es el intervalo de [0,2 - 1>con distancia de 0,01.
- min samples: ese parámetro es el número de muestras (o peso total) en una vecindad para que un punto se considere como un punto central. Esto incluye el punto en sí. El espacio de búsqueda de este parámetro es [5, 7, 9, 11, 13].

El comportamiento de estos parámetros se puede ver en la figura 5.2, en estas gráficas el parámetro “min samples“ está representando por las líneas de colores y que el parámetro “eps“ se encuentra en el eje “x“, mientras que el eje “y“ está representado por el Coeficiente de Silueta. Dentro de la

gráfica de cada especialidad se puede ver un comportamiento similar de cada línea, estos tienen un aumento progresivo a medida que se aumenta el valor del “eps“, todas estas líneas convergen a un determinado valor del coeficiente de silueta, estos valores óptimos se encuentran en la parte de resultados.

Propagación de Afinidad

En este modelo se tiene un parámetro variable llamado “preference” este representa se define como la preferencia para cada punto en donde es más probable que los puntos con valores de preferencias más altos se elijan como ejemplos. El número de ejemplos, es decir, de agrupaciones, está influenciado por el valor de las preferencias de entrada. El espacio de búsqueda que tiene este parámetro esta dado por el intervalo $< -20,-1]$ con una distancia de 1. El comportamiento de este parámetro frente al Coeficiente de Silueta se puede ver en la figura 5.3, en estas gráficas el parámetro “preference“ se encuentra en el eje “x“, mientras que el eje “y“ está representado por el Coeficiente de Silueta. Analizando estas gráficas se pudo encontrar que el comportamiento del parámetro es muy diferente en cada especialidad, por ejemplo, en las especialidades E3, N6, P2, S3 se encontró que el valor del coeficiente de silueta va aumentando cuando el parámetro va disminuyendo. Sin embargo, en la especialidad N3 sucede todo lo contrario a medida que el parámetro disminuye el Coeficiente de Silueta disminuye. Otro ejemplo sobre el comportamiento del parámetro es que algunas especialidades tienen un crecimiento abrupto en el Coeficiente de Silueta, estas se pueden ver en G2, Q2, S3.

Tener diferentes tipos de comportamiento en el parámetro demuestra que el rendimiento de los estudiantes es diferente en cada especialidad.

MeanShift

En este modelo se tiene un parámetro llamado “bandwidth“ este se define como el ancho de banda utilizado en el kernel RBF, este consta de otros 2 parámetros llamados “quantile“ y “n samples“, se muestra a detalles estos parámetros:

- quantile: este parámetro debe estar entre [0, 1] por ejemplo 0,5 significa que se utiliza la mediana de todas las distancias por pares. El espacio de búsqueda de este parámetro es el intervalo de [0 - 1>con distancia de 0,1.
- n samples: ese parámetro es el número de muestras (o peso total) que se utilizarán. El espacio de búsqueda de este parámetro es el intervalo de [5 - 31>con distancia de 5.

El comportamiento de estos parámetros se puede ver en la figura 5.4, en estas gráficas el parámetro “n samples“ está representando por las líneas de colores y que el parámetro “quantile“ se encuentra en el eje “x“, mientras que el eje “y“ está representado por el Coeficiente de Silueta. Dentro de la gráfica de cada especialidad se puede ver un comportamiento similar de cada línea, estos tienen un aumento progresivo a medida que se aumenta el valor del “quantile“. Sin embargo, llega un punto en que el Coeficiente de Silueta empieza a descender. También, se puede notar que cuando el valor de “n samples“ es mayor este crece más rápido hacia el valor máximo, este comportamiento se puede ver claramente en la especialidad de A1 y L2.

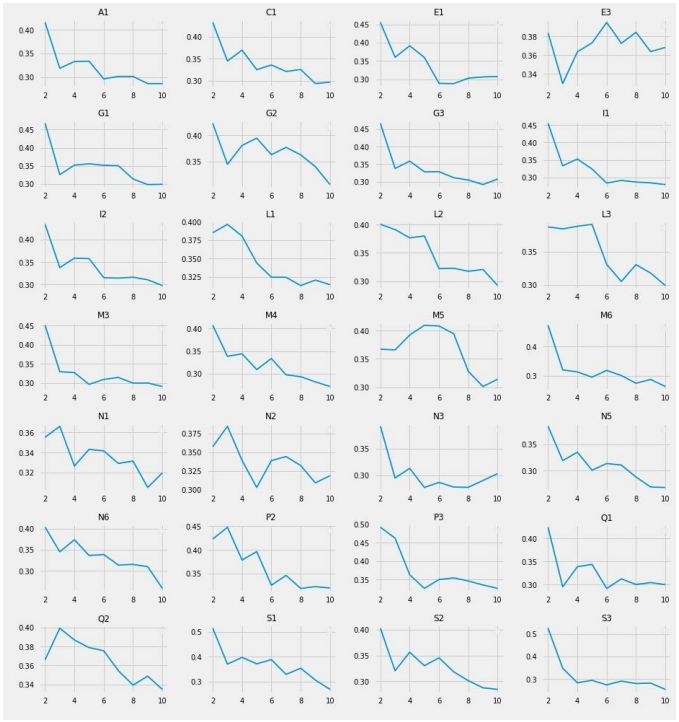

Figura 5.1: Comportamiento de los parámetros del modelo K-means frente al Coeficiente de Silueta, en el eje “x“ está el parámetro de numero de grupos y en el eje “y“ está el Coeficiente de Silueta.

Cuadro 6.1: Parámetros óptimos y resultados del modelo K-means para cada especialidad

Esp	num clusters	silhouette score
A1	2	0.41
C1	2	0.43
E1	2	0.45
E3	6	0.39
G1	2	0.46
G2	2	0.42
G3	2	0.46
I1	2	0.45
I2	2	0.43
L1	3	0.39
L2	2	0.40
L3	5	0.39
M3	2	0.45
M4	2	0.40

Esp	num clusters	silhouette score
M5	5	0.40
M6	2	0.47
N1	3	0.36
N2	3	0.38
N3	2	0.39
N5	2	0.38
N6	2	0.40
P2	3	0.44
P3	2	0.49
Q1	2	0.42
Q2	3	0.39
S1	2	0.51
S2	2	0.40
S3	2	0.52

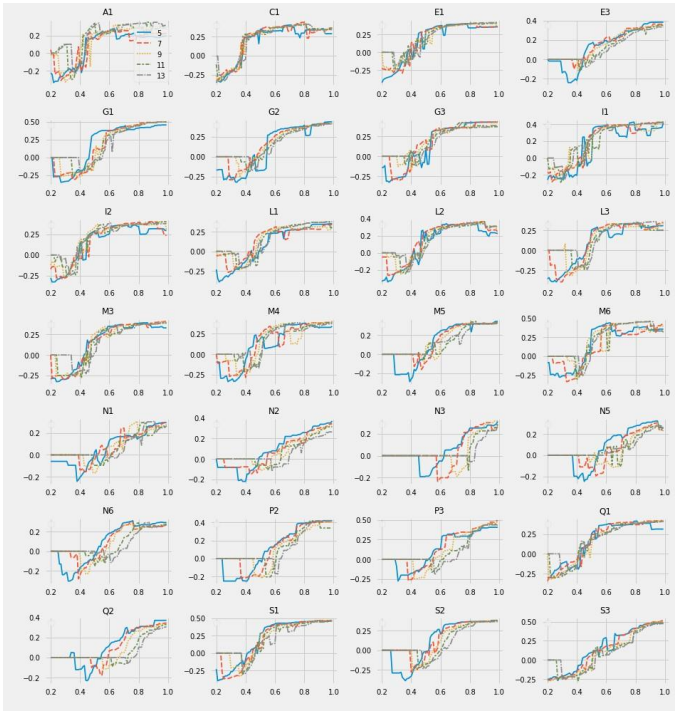

Figura 5.2: Comportamiento de los parámetros del modelo BSCAN frente al Coeficiente de Silueta, las líneas están representadas por el min samples“ y en el eje “x“ está el parámetro de eps, mientras en el eje “y“ está el Coeficiente de Silueta.

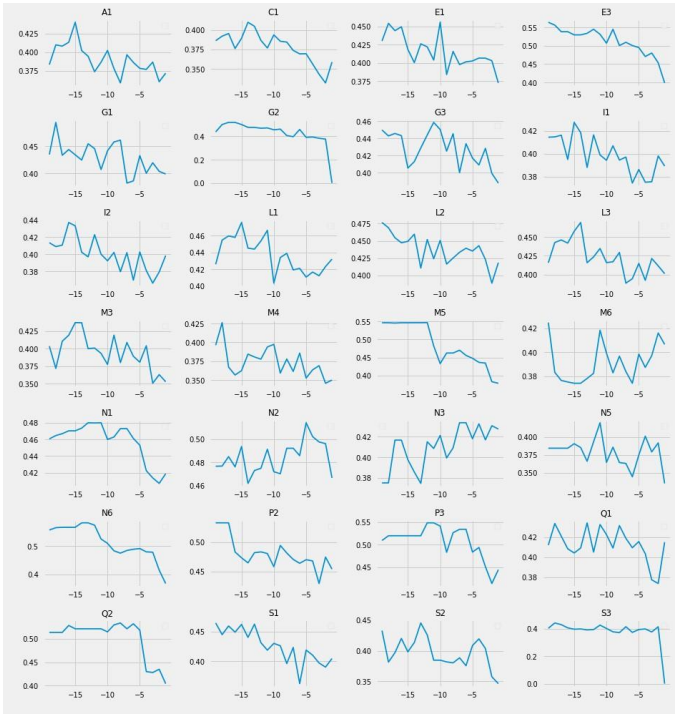

Figura 5.3: Comportamiento de los parámetros del modelo propagación de Afinidad frente al Coeficiente de Silueta, en el eje “x” está el parámetro “preference“ y en el eje “y“ está el Coeficiente de Silueta.

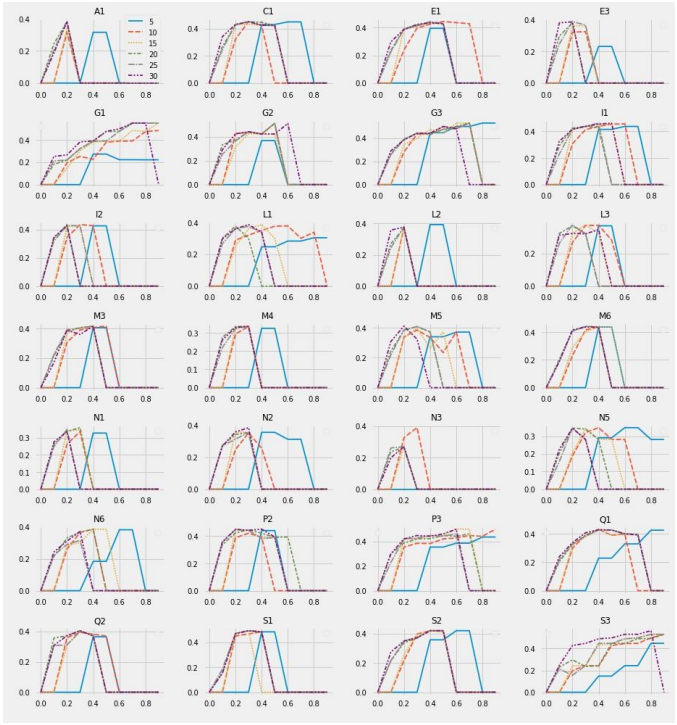

Figura 5.4: Comportamiento de los parámetros del modelo MeanShift frente al Coeficiente de Silueta, las líneas están representadas por el “n samples“ y en el eje “x“ está el parámetro de “quantile“, mientras en el eje “y“ está el Coeficiente de Silueta.

VI. Resultados

Se mostrará los resultados de los modelos con los parámetros óptimos encontrados en la sección anterior. Posteriormente se realizará una comparación entre los modelos para decidir cuál es el mejor modelo según la aplicación plantea.

A. Selección de la Métrica de Éxito.

Antes de presentar los resultados de cada modelo es necesario definir la métrica de éxito que nos dará la seguridad de que cierto modelo es el mejor para la aplicación de focalizar alumnos para realizar tutorías. En esta investigación se plantea al número de clusters o grupos como la métrica de éxito, es decir que aquel modelo que genere una mayor cantidad de grupos será el mejor, se plantea esta métrica porque cuando se tiene una mayor cantidad de grupos estos suelen ser más pequeños y tienen una menor cantidad de alumnos, entonces realizar tutorías a grupos pequeños de alumnos es más sencillo. Además, si se realizarían tutorías a grupos muy grandes de 200 a más alumnos sería muy difícil, ya que las tutorías requieren de un seguimiento continuo.

Si nos enfocamos en la aplicación que se le va dar a los clusters, que consiste en encontrar grupos de alumnos que requieren tutorías estos serían alumnos se deben encontrar en el cuadrante III, ya que estos tienen un rendimiento baja en ambos periodos.

B. Resultados de los Modelos.

Los modelos para cada especialidad fueron ejecutados usando los mejores parámetros donde estos tenían el mayor valor en el Coeficiente de Silueta, asegurando de que el número de grupos que se generaban sería el óptimo según el modelo, a excepción del modelo de K-means donde el número de grupos era el parámetro.

Es importante mencionar que en la gráfica de los clusters solo se considera dos dimensiones en el eje "x" se encuentra el score estandarizado del periodo 19-2 y en el eje "y" el score del periodo 20-1. Sin embargo, la segmentación se realizó en cuatro dimensiones estas dimensiones se muestran en la tabla 5.3.

K-Means

El resultado del Coeficiente de Silueta con los parámetros óptimos está en la tabla 6.1, mientras que las agrupaciones se encuentran en la figura 6.1.

Según la tabla se tiene que la mayoría de especialidades posee de 2 a 3 grupos a excepción de 3 especialidades las cuales son E3 con 6 grupos, L3 con 5 grupos y M5 con 5 grupos. Cuando se visualiza los grupos de cada especialidad se puede notar que la mayoría tiene un grupo en el primer cuadrante estos son alumnos que han tenido un alto rendimiento en los periodos 19-2 y 20-1. Sin embargo, también se encuentran 3 especialidades que tienen más de un grupo en esa zona, estos son E3, L3 y M5 los mismos que tenían una mayor cantidad de grupos.

Si nos enfocamos en el tercer cuadrante que son los alumnos objetivos para las tutorías, el modelo de K-mean tan solo genera un grupo en la mayoría de las especialidades a excepción de las 3 especialidades mencionadas anteriormente.

DBSCAN

El resultado del Coeficiente de silueta con los parámetros óptimos está en la tabla 6.2, mientras que las agrupaciones se encuentran en la figura 6.2.

Según la tabla se tiene que algunas especialidades solo contienen un grupo en este caso este modelo no es el más adecuado, otros tienen un resultado similar al K-means donde contiene de 2 a 3 grupos, a excepción de E3, I2, N6 que contienen 4 grupos y L2 que contiene 5 grupos. Si pasamos a contrastar estos resultados en la gráfica podemos notar que las especialidades que tienen de 1 a 2 grupos, la mayoría de estos se concentran en los dos primeros cuadrantes y consideran al cuadrante III como ruido. Si ahora vemos a las especialidades con más grupos podemos ver en comportamiento similar en donde la mayoría de los grupos están en el I cuadrante un claro ejemplo es la especialidad N6.

El resultado de tener grupos en el primer cuadrante se debe a que los datos están más próximos en comparación con los otros cuadrantes.

Propagación de Afinidad

El resultado del Coeficiente de silueta con los parámetros óptimos está en la tabla 6.3, mientras que las agrupaciones se encuentran en la figura 6.3.

Según la tabla se tiene que la cantidad de grupos que genera este modelo es mayor en comparación que los otros, las especialidades con menor cantidad de grupos son P2 Y P3 que tiene 6 grupos, por otro lado, la especialidad con mayor cantidad de grupos es C1 con 24 grupos. En la gráfica se puede ver que existen grupos en todos los cuadrantes de todas las especialidades, si nos enfocamos en el tercer cuadrante de algunas facultades como A1, C1, E1, entre otras podemos ver que existe más de un grupo esto es muy beneficioso para la aplicación de la investigación.

Algo que notar de este modelo es que la cantidad de grupos es proporcional a la cantidad de alumnos.

MeanShift

El resultado del Coeficiente de silueta con los parámetros óptimos está en la tabla 6.4, mientras que las agrupaciones se encuentran en la figura 6.4.

Según la tabla se tiene que la mayoría de especialidades posee de 2 a 5 grupos a excepción de 2 especialidades las cuales son E3 con 6 grupos, M5 con 9 grupos. Cuando se visualiza los grupos de cada especialidad se puede notar que la cantidad de estos es mayor en este modelo a comparación de K-means, si nos enfocamos en el III cuadrante podemos notar que grupos que se generan son tan solo de uno y abarcan otros cuadrantes. Un buen resultado de este modelo se puede ver en la especialidad M5 en que tiene 5 grupos en el tercer cuadrante. Sin embargo, 2 de estos grupos abarcan otros cuadrantes.

Algo que notar de este modelo es que es útil si se quiere encontrar grupos de grandes de alumnos, algo que para la aplicación de la investigación no es beneficioso.

Cuadro 6.2: Parámetros óptimos y resultados del modelo DBSCAN para cada especialidad

Esp.	eps	min samples	silhouette score	n clusters	num noise
A1	0.93	13	0.35	1	77
C1	0.81	7	0.44	3	65
E1	0.99	13	0.42	2	70
E3	0.95	5	0.38	4	13
G1	0.93	7	0.50	1	41
G2	0.96	5	0.44	3	12
G3	0.81	13	0.43	1	71
I1	0.99	11	0.43	2	68
I2	0.92	9	0.41	4	53
L1	0.96	11	0.38	1	52
L2	0.9	7	0.37	5	35
L3	0.92	13	0.36	1	64
M3	0.99	7	0.41	2	43
M4	0.99	7	0.40	1	45
M5	0.81	9	0.35	3	39
M6	0.89	11	0.46	1	82
N1	0.91	13	0.31	1	81
N2	0.99	5	0.37	2	25
N3	0.97	9	0.32	1	42
N5	0.94	5	0.32	2	30
N6	0.76	5	0.31	4	54
P2	0.93	5	0.42	2	20
P3	0.98	7	0.49	1	27
Q1	0.96	7	0.42	4	39

Q2	0.98	5	0.37	2	16
S1	0.98	5	0.47	3	18
S2	0.98	5	0.38	2	19
S3	0.99	7	0.52	1	34

Cuadro 6.3: Parámetros óptimos y resultados del modelo Propagación de Afinidad para cada especialidad.

Esp	preference	silhouette score	Num clusters
A1	-15	0.44	16
C1	-14	0.41	24
E1	-10	0.46	18
E3	-19	0.56	9
G1	-18	0.49	9
G2	-16	0.52	11
G3	-11	0.46	15
I1	-15	0.43	16
I2	-16	0.44	16
L1	-15	0.48	14
L2	-19	0.48	12
L3	-14	0.47	14
M3	-15	0.44	12
M4	-18	0.43	12
M5	-12	0.55	8
M6	-19	0.42	11
N1	-11	0.48	9
N2	-5	0.51	14
N3	-6	0.43	10
N5	-11	0.42	9
N6	-13	0.58	10
P2	-17	0.53	6
P3	-11	0.55	6
Q1	-13	0.43	19
Q2	-8	0.53	8
S1	-19	0.46	9
S2	-13	0.45	10
S3	-18	0.44	9

Cuadro 6.4: Parámetros óptimos y resultados del modelo MeanShift para cada especialidad.

Esp	quantile	num samples	silhouette score	Num clusters
A1	0.2	20	0.38	3
C1	0.3	10	0.45	3
E1	0.5	10	0.44	4
E3	0.2	25	0.39	6
G1	0.7	20	0.55	2
G2	0.5	25	0.51	2
G3	0.6	15	0.53	2
I1	0.4	25	0.46	2
I2	0.2	20	0.43	3
L1	0.3	30	0.39	4
L2	0.4	5	0.39	3
L3	0.3	15	0.38	5
M3	0.4	15	0.42	2
M4	0.3	25	0.34	5
M5	0.2	30	0.41	9
M6	0.3	20	0.44	3
N1	0.3	20	0.36	3
N2	0.3	30	0.39	4
N3	0.3	10	0.39	2
N5	0.4	10	0.35	5
N6	0.4	10	0.38	3
P2	0.4	30	0.45	4
P3	0.6	15	0.50	2
Q1	0.4	15	0.43	5
Q2	0.3	20	0.40	3
S1	0.3	15	0.49	5
S2	0.4	10	0.42	3
S3	0.8	30	0.56	2

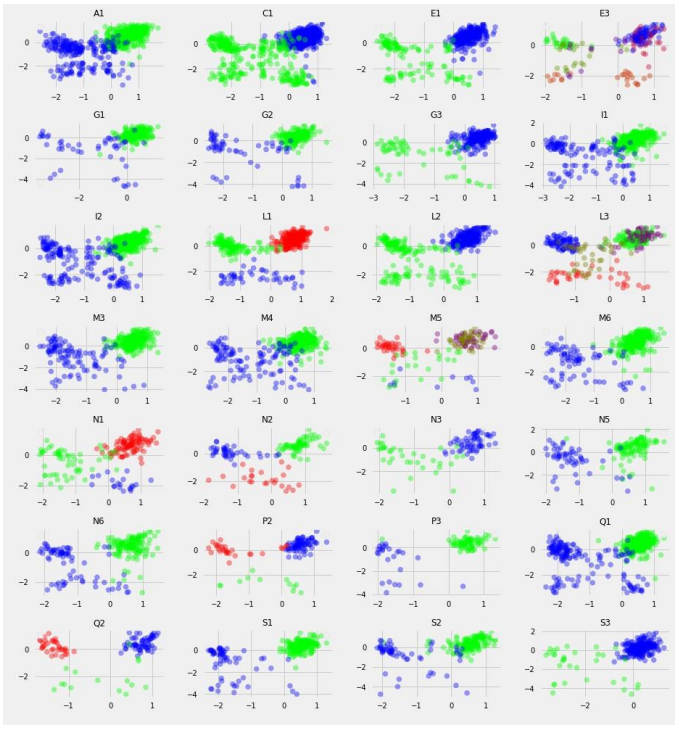

Figura 6.1: Agrupaciones óptimas generadas por el modelo K-means en cada especialidad, en cada eje se encuentra el SCORE estandarizado en el eje "x" está el 19-2 y en el eje "y" está el 20-1.

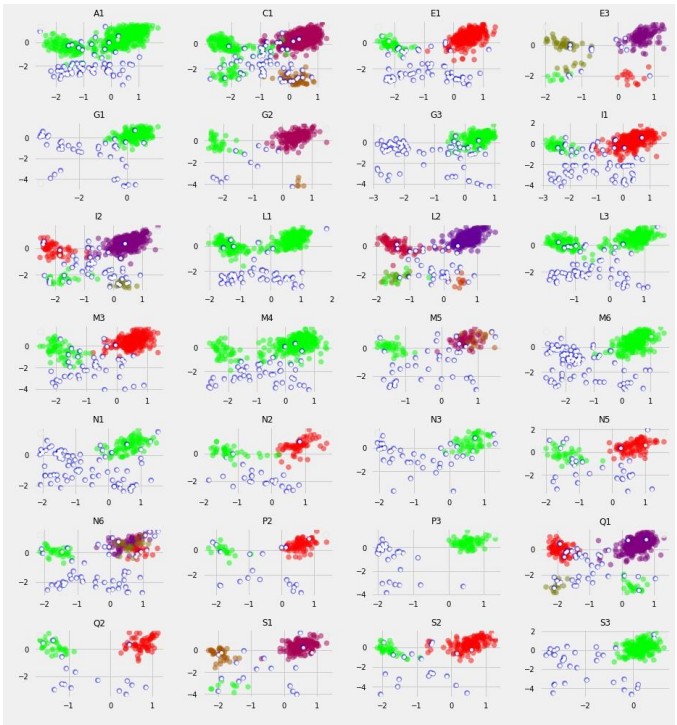

Figura 6.2: Agrupaciones óptimas generadas por el modelo DBSCAN en cada especialidad, en cada eje se encuentra el SCORE estandarizado en el eje "x" está el 19-2 y en el eje "y" está el 20-1.

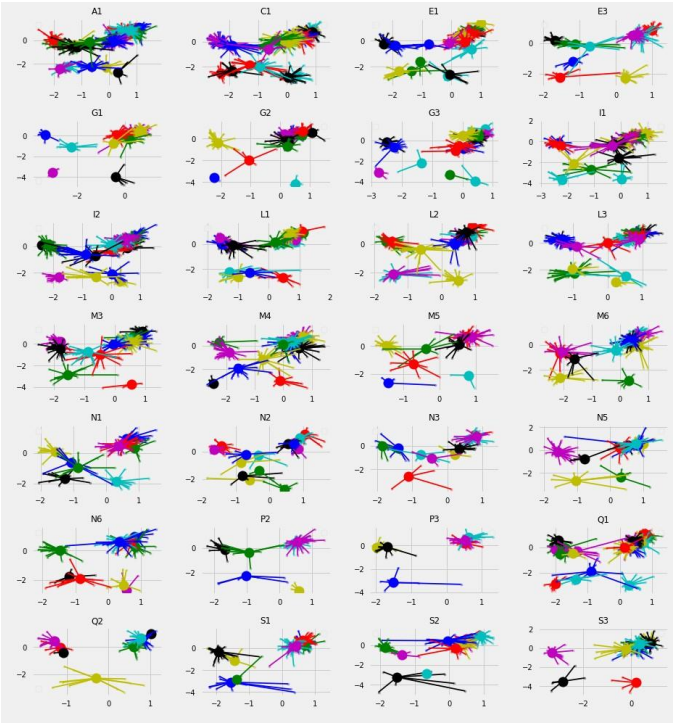

Figura 6.3: Agrupaciones óptimas generadas por el modelo Propagación de Afinidad en cada especialidad, en cada eje se encuentra el SCORE estandarizado en el eje "x" está el 19-2 y en el eje "y" está el 20-1.

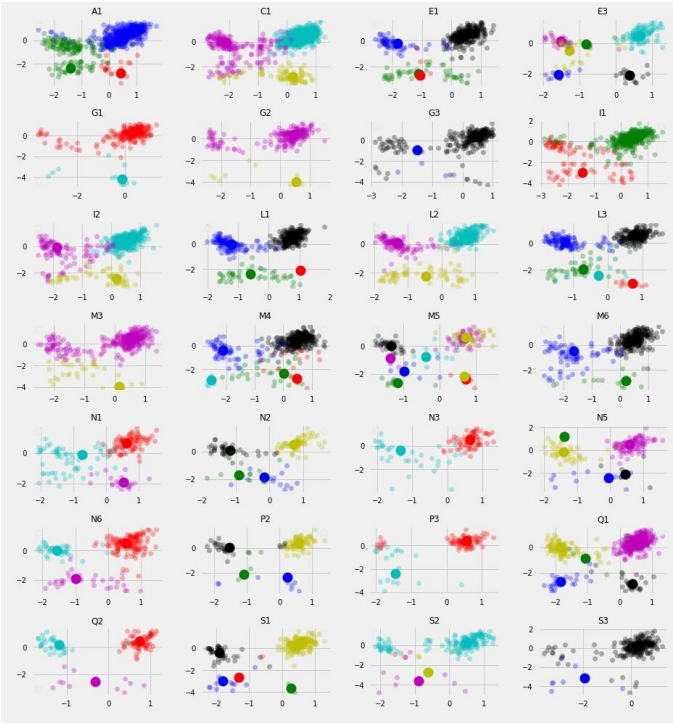

Figura 6.4: Agrupaciones óptimas generadas por el modelo MeanShift en cada especialidad, en cada eje se encuentra el SCORE estandarizado en el eje "x" está el 19-2 y en el eje "y" está el 20-1.

VII. Conclusiones

La presentación de los grupos encontrados en cada modelo se basa en el análisis de los datos de los alumnos en relación con las dimensiones consideradas (score estandarizado del periodo 19-2 y score del periodo 20-1 en el eje x e y respectivamente). Cada modelo utiliza diferentes algoritmos y parámetros para agrupar a los alumnos de manera óptima, y los resultados reflejan la forma en que los datos se agrupan en función de esas características.

Las características de los grupos encontrados pueden variar según el modelo y la especialidad. En general, se observa una separación de los grupos en diferentes cuadrantes, lo que indica diferentes niveles de rendimiento académico en los dos periodos considerados. Los grupos en el primer cuadrante suelen representar a los alumnos con un alto rendimiento en ambos periodos, mientras que los grupos en el tercer cuadrante son los alumnos con un bajo rendimiento en ambos periodos, que son los que se consideran como objetivo para las tutorías.

Métrica de éxito: La métrica de éxito utilizada en este estudio es el número de grupos generados por cada modelo. Se considera que un mayor número de grupos es preferible, ya que esto implica la formación de grupos más pequeños de alumnos que requieren tutorías. La lógica detrás de esto es que los grupos más pequeños permiten un seguimiento más personalizado y eficiente en las tutorías.

Resultados de K-Means: El modelo de K-Means genera en su mayoría de 2 a 3 grupos por especialidad, excepto en 3 especialidades (E3, L3 y M5) que tienen más grupos. La mayoría de las especialidades tienen al menos un grupo en el primer cuadrante, lo que indica un alto rendimiento en ambos periodos. Sin embargo, solo se genera un grupo en el tercer cuadrante en la mayoría de las especialidades, que son los alumnos objetivo para las tutorías.

Resultados de DBSCAN: El modelo de DBSCAN muestra resultados variados. Algunas especialidades tienen solo 1 grupo, lo cual no es adecuado para la aplicación de tutorías. Sin embargo, otras especialidades muestran resultados similares a K-Means, con 2 a 3 grupos. Algunas especialidades (E3, I2, N6 y L2) tienen más grupos, y en general, los grupos tienden a concentrarse en los primeros cuadrantes.

Resultados de Propagación de Afinidad: Este modelo genera la mayor cantidad de grupos en comparación con los otros modelos. La cantidad de grupos varía según la especialidad, siendo la especialidad C1 la que tiene la mayor cantidad de grupos (24 grupos). En general, se encuentran grupos en todos los cuadrantes, lo cual es beneficioso para la aplicación de tutorías. Además, la cantidad de grupos es proporcional a la cantidad de alumnos.

Resultados de MeanShift: El modelo de MeanShift genera de 2 a 5 grupos en la mayoría de las especialidades, con

algunas excepciones que tienen más grupos. A diferencia de K-Means, se generan grupos en el tercer cuadrante, aunque algunos de estos grupos también abarcan otros cuadrantes. Este modelo es útil si se buscan grupos grandes de alumnos, pero no es beneficioso para la aplicación de tutorías.

En general, cada modelo tiene sus fortalezas y debilidades en relación con la aplicación de focalizar alumnos para realizar tutorías. Las acciones que se pueden proponer sobre los grupos encontrados dependen de los objetivos y las necesidades de la universidad. Algunas posibles acciones podrían ser:

Tutorías personalizadas: Los grupos identificados en el tercer cuadrante, que representan a los alumnos con bajo rendimiento en ambos periodos, pueden ser el foco principal de las tutorías. Se pueden diseñar programas de tutorías específicos para abordar las necesidades de estos grupos y proporcionarles un apoyo académico adicional.

Monitoreo continuo: Los grupos identificados en el primer cuadrante, que representan a los alumnos con alto rendimiento en ambos periodos, también pueden beneficiarse del seguimiento continuo. Aunque no necesiten tutorías intensivas, se les puede ofrecer orientación y apoyo para mantener su rendimiento académico y fomentar su desarrollo.

Identificación de necesidades específicas: Dentro de cada grupo, se pueden realizar análisis más detallados para identificar las necesidades y dificultades específicas de los alumnos. Esto permitirá adaptar las tutorías y los recursos educativos de manera individualizada, brindando un enfoque más efectivo y personalizado.

Intervenciones tempranas: Si se detectan patrones consistentes en ciertos grupos a lo largo del tiempo, es posible implementar intervenciones tempranas para prevenir un deterioro adicional en el rendimiento académico. Estas intervenciones podrían incluir programas de apoyo adicional, estrategias de motivación y seguimiento más cercano.

Es importante destacar que las acciones propuestas deben adaptarse a las características específicas de cada grupo y a las políticas y recursos disponibles en el contexto educativo. La aplicación de tutorías y medidas de apoyo efectivas requiere una planificación cuidadosa y una evaluación continua de su impacto en el rendimiento y bienestar de los alumnos.

AGRADECIMIENTO

Se agradece a la Universidad Nacional de Ingeniería UNI Perú, especialmente a la Facultad de Ciencias y al Vicerrectorado de Investigación de la UNI.

REFERENCIAS

[1] UNESCO. La educación en américa latina y elcaribe ante lacovid-19. https://es.unesco.org/fieldoffice/santiago/covid-19-education-alc/monitoreo.
[2] Trevor Hastie. The elements of statistical learning: data mining, inference, and prediction. Springer, New York, 2009.
[3] J. MacQueen. Some methods for classification and analysis of multivariate observa-tions. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, pages 281–297, Berkeley, Calif., 1967. University of California Press.
[4] Sklearn. Clustering sklearn. https://scikit-learn.org/stable/modules/clustering.html.
[5] Matt Nedrich. Mean shift clustering. https://spin.atomicobject.com/2015/05/26/meanshift-clustering/: :text=%20Mean%20Shift%20Clustering%20%201%20Kernel%20Density,mean%20shift%20algorithm%20iteratively%20shifts%20each...%20More%20.