Análisis de clasificación de aspirantes en procesos de selección de la compañía Galletas Noel

Sara Lucía Echeverry Agudelo

Universidad de Antioquia
Facultad de Ciencias Exactas y Naturales
Instituto de Matemáticas
Medellín, Colombia
2021

Análisis de clasificación de aspirantes en procesos de selección de la compañía Galletas Noel

Sara Lucía Echeverry Agudelo

Trabajo de grado presentado como requisito parcial para optar al título de:
Estadístico

María Eugenia Castañeda López
Orientador Interno, Instituto de Matemáticas

Ángela María Acosta Sierra
Orientador externo, Galletas Noel

Universidad de Antioquia
Facultad de Ciencias Exactas y Naturales
Instituto de Matemáticas
Medellín, Colombia
2021

Resumen

La compañía Galletas Noel S.A.S cumple en el mercado 105 años y en la actualidad busca introducir el análisis de datos en sus procesos de selección de colaboradores. El área de gestión humana se encuentra interesada en identificar características clave que pertenezcan a sus aspirantes, con el objetivo de reconocer personas con tendencia a ser seleccionadas y formar parte de la compañía. Para la solución de este planteamiento se considera un problema de clasificación de aspirantes, con lo que se lleva a cabo el desarrollo de un algoritmo de reducción de dimensionalidad no lineal llamado UMAP y un algoritmo de agrupación jerárquico llamado HDBSCAN. Además, se realiza una estadística descriptiva para comparar los grupos.

Palabras clave: Clúster; Agrupación jerárquica; Selección; HDBSCAN; UMAP.

Abstract

The company Galletas Noel S.A.S has been in the market for 105 years and is currently seeking to introduce data analysis in its employee selection processes. The human resources area is interested in studying key characteristics that belong to its applicants, with the objective of identifying people with a tendency to be selected and become part of the company. For the solution of this approach, an applicant classification problem is considered, with which the development of a nonlinear dimensionality reduction algorithm called UMAP and a hierarchical clustering algorithm called HDBSCAN is carried out. In addition, descriptive statistics is performed to compare the groups.

Keywords: Clustering; Hierarchical clustering; Selection; HDBSCAN; UMAP.

Análisis de clasificación de aspirantes a trabajar en la Compañía Galletas Noel

Sara Lucía Echeverry Agudelo [*]

28 de julio de 2021

Contenido

[*]E-mail: lucia.echeverry@udea.edu.co, Instituto de Matemáticas, Universidad de Antioquia, Medellín, Colombia.

1. Introducción

En el tiempo actual, con el gran desarrollo de las tecnologías y estudios relacionados con el análisis de datos, las compañías tienen la necesidad de estar a la vanguardia. Una de estas compañías es Galletas Noel, quien cumple en el mercado 105 años y donde sus integrantes buscan constantemente la innovación y el desarrollo de nuevas ideas que transformen, actualicen y mejoren sus diferentes procesos, ya sean administrativos u operativos. Es por ello que el área de gestión humana perteneciente a esta compañía busca renovar sus procesos de selección de colaboradores.

Los procesos de gestión humana se focalizan en temas de formación, desarrollo y atracción de talento humano. Esta área busca una nueva forma de reconocer aquellos aspirantes que poseen características clave, que hacen del individuo un buen prospecto y que tienda a formar parte de la compañía. Un aporte positivo de la buena elección de personal es una baja rotación de empleados, ya que los integrantes nuevos demandan un buen tiempo y gasto en capacitación y acoplación al rol que desempeñarán, con lo que mejoran también los diferentes procesos e indicadores internos relacionados con estas variables.

Para la solución de este problema se desarrollarán metodologías estadísticas de clasificación, con lo que se obtendrá una segmentación de los individuos en dos grupos, donde entre ellos compartirán características similares. En los grupos se busca identificar esas características que hacen que los individuos sean una buena opción para el ingreso a la organización. Estas características que se tendrán en cuenta son características evaluadas y medidas en los procesos de selección de personal. También hay un interés por encontrar entre las variables algún tipo de correlación que nos indique la existencia de relación lineal.

El esquema del documento es el siguiente. En la sección 2 se describen los fundamentos teóricos de los métodos usados para darle solución al problema, estos son: Algoritmo UMAP y HDBSCAN. En la sección 3 se expone la metodología desarrollada, donde se evidencia con más detalle el uso de los algoritmos. En la sección 4 se exponen los resultados obtenidos. Para finalizar, en la sección 5 se presentan las conclusiones y sugerencias a las que se llegaron con el desarrollo de las metodologías para tenerlos en cuenta en futuros trabajos.

2. Marco Teórico

2.1. UMAP

UMAP (*Uniform Manifold Approximation and Projection for Dimension Reduction*) es un algoritmo para la reducción de dimensionalidad, basado en técnicas de aprendizaje y análisis de datos topológicos. A continuación, se explicará de forma sencilla el funcionamiento de este método y cómo soluciona algunos problemas que se encuentran en el camino para su desarrollo. Esta explicación está basada en *How umap works* [3], *Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study* [4] y *Understanding UMAP* [1]. Lo primero que se explica es cómo a través de nuestros datos, se puede crear un espacio topológico.

Los *Simplicial complexes* son una herramienta para crear espacios topológicos a partir de componentes combinatorios simples llamados *simplices*. Un *simplex* es una forma sencilla de crear un objeto $k-dimensional$. Es llamado *K-simplex* y se forma tomando el casco convexo entre $K+1$ puntos independientes.

Los *simplices*, entonces, son bloques de construcción, los cuales se unirán para crear espacios topológicos interesantes. A la unión de estos bloques se les llama *simplicial complex*. En palabras sencillas un *simplicial complex* es un conjunto de *simplices* unidos a lo largo de caras. Dada una cubierta abier-

ta de un espacio topológico, se crea un *simplicial complex*. Haciendo cada conjunto de la cubierta un *0-simplex*, y luego un *1-simplex* entre dos *0-simplex*, luego se unen tres de estos *1-simplex* obteniendo un *2-simplex* y así sucesivamente. Aunque este procedimiento parte de una idea sencilla, logra capturar gran parte de la topología.

Ahora, considerando que nuestra muestra de datos viene de un espacio topológico, entonces se crea un cubrimiento abierto para estudiar este espacio. Si se puede medir la distancia entre los puntos de datos, la forma en que se crea el cubrimiento abierto consiste en crear bolas de radio fijo con centro en cada punto[3].

Generalmente el *simplicial complex* está formado la mayoría por *0-simplex* y *1-simplex*, la ventaja es que computacionalmente estos son más fáciles de manejar. De esta forma se representa el espacio topológico, con lo que se reduce su dimensionalidad encontrando un espacio topológico similar de baja dimensión.

Cuando se usa este algoritmo en datos reales surgen ciertas dificultades. Por ejemplo, designar el radio de las bolas con las que se crea el cubrimiento abierto. Si se elige un radio muy pequeño entonces el espacio estará formado por muchos *simplicial complex* aislados, lo que produce una falta de cobertura y si se toma un radio muy grande entonces se obtienen pocos *simplicial complex* de dimensión grande, esto hace que se pierda el detalle.

La justificación de que este proceso capture la topología del espacio se basa en el teorema del Nervio[3]. Si los datos se distribuyeran de manera uniforme esto no representaría un problema, tomando el radio como la distancia promedio de los puntos se esperaría que todo marche bien, pero cuando no existe una distribución uniforme esta no es una forma adecuada para escoger la distancia.

Considerar entonces una distancia local para cada punto que dependa de sus *K-vecinos más cercanos* puede ser una solución. Localmente la bola unitaria se extenderá hasta su *k–vecino* más cercano. Cada punto tendrá su propia función de distancia única[1].

Escoger un número k es más sencillo, aunque igual depende de los datos y se debe observar las distancias de los mismos para elegir un buen valor, sin embargo, $k = 10$ debería funcionar para la mayoría de los datos. Un tamaño de k pequeño o grande tiene la misma analogía que la del tamaño del radio.

Al tener una bola de distancia local para cada punto se tiene la opción de medir significativamente esa distancia, y pasar de tener puntos dentro o fuera de ese espacio a tener probabilidades de pertenecer al espacio, estas probabilidades disminuyen al alejarse del centro de la bola. Para evitar puntos totalmente aislados (conectividad local) esa probabilidad decaerá a partir del punto más cercano. UMAP usa la *distribución exponencial* para calcular estas probabilidades[4]

$$p_{i|j} \propto exp\left(-\frac{d(x_i,x_j)-\rho_i}{\sigma_i}\right) \tag{1}$$

Donde $d(x_i,x_j)$ es la distancia entre los puntos x_i,x_j y ρ_i es la distancia entre x_i y su punto más cercano. Ahora un nuevo problema consiste en que las distancias locales no son compatibles. Esto es, la distancia del punto x_i al punto x_j es diferente a la distancia del punto x_j al punto x_i. Para esto lo que se hará es combinar las probabilidades, obteniendo la probabilidad de que exista un *1-simplex* entre los dos puntos, obteniendo entonces la probabilidad [3]

$$p_{ij} = p_{i|j} + p_{j|i} - p_{i|j}p_{j|i} \tag{2}$$

Ahora con la construcción del espacio topológico a partir de los datos, se busca encontrar un espacio lo más similar posible, pero en baja dimensión.

Se construye un espacio topológico de baja dimensión (espacio euclidiano) de la misma manera que el espacio topológico de alta dimensión. Hay que tener en cuenta que en el espacio euclidiano no cabe la definición de distancia local, UMAP para calcular la significancia de las distancias en este espacio, usa una distribución con forma similar a la *t-student*[3]

$$q_{ij} \propto (1 + a(y_i - y_j)^{2b})^{-1} \tag{3}$$

Donde $a \approx 1.93$ y $b \approx 0.79$. Este ya es un problema de optimización, donde se busca encontrar el espacio topológico de baja dimensión más parecido posible al espacio de alta dimensión, ya que los dos espacios cuentan con los mismos puntos, o sea los mismo *0-simplex*, entonces se comparan los dos *1-simplex* de los diferentes espacios. Como esto se comporta como variables *Bernoulli*, ya que el *1-simplex* existe o no con una probabilidad p_{ij} entonces la idea es optimizar la entropía cruzada.

Consideremos el conjunto de todos los *1-simplex* posibles, entonces se tienen funciones de probabilidad p_{ij} para los *1-simplex* en el espacio de alta dimensión y para el espacio de baja dimensión q_{ij}, entonces la entropía cruzada es[3]

$$CE(P,Q) = \sum_i \sum_j \left(p_{ij} log\left(\frac{p_{ij}}{q_{ij}}\right) + (1 - p_{ij}) log\left(\frac{1 - p_{ij}}{1 - q_{ij}}\right)\right) \tag{4}$$

La primera parte de la ecuación proporciona un tipo de fuerza de atracción, se hace grande cuando p_{ij} es grande y se minimizará cuando q_{ij} se haga lo más grande posible, la segunda parte de la ecuación proporciona una fuerza de repulsión, se hace grande cuando p_{ij} es pequeño y se minimiza cuando q_{ij} se haga lo más pequeño posible. Para el proceso de optimización UMAP usa el descenso del gradiente estocástico.

2.2. HDBSCAN

HDBSCAN es un algoritmo de agrupación jerárquico basado en la densidad de los datos, para encontrar clústers es necesario identificar las zonas con mayor densidad de puntos en medio de una zona de puntos de ruido, la suposición de ruido es importante pues en el mundo real existen puntos atípicos y confusos. El algoritmo se basa en los enlaces únicos [2] y esto puede ser sensible al ruido, por lo que primero se realiza una transformación del espacio donde por medio de la "distancia de accesibilidad mutua"[2] se alejan aun más los posibles puntos de ruido.

Después de esto se crea un espacio difuso, obteniendo componentes totalmente conectados y desconectados, esto basado en "el árbol de expansión mínimo del grafo"[2]. Luego de la construcción de este árbol se evalúa la jerarquía de los clúster, y se llega a un árbol más pequeño con un poco más de datos en cada nodo, se evalúa también la estabilidad de los nodos para elegir los clústers, se puede obtener más información sobre este algoritmo en [2].

3. Metodología

La metodología que se desarrolló para este trabajo se divide en varias etapas. Primero se presenta la base de datos: en que consiste, donde se lleva registrada y la forma en que se adecuo para su uso; luego la forma en que se aplicaron los algoritmos UMAP y HDBSCAN descritos en la sección anterior. La limpieza de los datos y la aplicación de los algoritmos se desarrolló en Excel y Python [5].

3.1. Base de datos

Para el procedimiento de selección de operarios el área de gestión humana ha implementado diferentes pruebas que demuestran la destreza de los aspirantes. El resultado de estas pruebas está registrado en hojas de cálculo de Google, en este trabajo se van a considerar los datos desde el año 2019 hasta el año 2020.

Las variables son puntuaciones que obtienen los aspirantes en pruebas de concentración, reacción, agilidad y relacionadas con la destreza física. Existen variables categóricas que describen la personalidad de los aspirantes, pero estas variables poseen gran cantidad de datos faltantes y no se guardan de una manera adecuada para hacer uso de ellas. Respecto a las demás variables, son variables cuantitativas discretas donde se eliminan los registros con valores faltantes, al encontrar variables con poca información se decide no tenerlas en cuenta para el estudio.

Primero se realiza un análisis gráfico para cada variable con el fin de identificar datos atípicos. Luego se realiza un análisis de correlación lineal y dados los resultados se hace un cambio en los datos, donde se eliminan 6 variables con correlación lineal mayor a 0.7, con lo que se tiene una base de datos constituida por 1348 registros y 7 variables. Además, se tiene una variable binaria que indica las etiquetas de los datos, si el aspirante paso las pruebas de selección o no. Para fines prácticos y de confidencialidad estas variables se llamarán MC, MG, RB, DS, PC, LS, LA.

3.2. Aplicación UMAP

Se aplica a los datos el algoritmo UMAP, con el objetivo de llevar los datos a un espacio de dos dimensiones. Luego del uso de varios valores y sabiendo que un valor pequeño para los *K-vecinos más cercanos* se centra en una estructura muy detallada, se decide entonces aumentar el parámetro del valor determinado a $k = 20$.

Para evaluar el comportamiento del algoritmo en individuos nuevos, se separa la base de datos en dos grupos: grupo de entrenamiento y grupo de testeo. Primero para establecer la transformación del algoritmo, se da como argumento el grupo de datos de entrenamiento y luego esta transformación se aplica a los datos del grupo de testeo.

El algoritmo se aplicó de dos formas: supervisado y no supervisado.

3.2.1. Supervisado

En este caso se entrega al algoritmo la variable de etiquetas. El algoritmo tiene entonces en cuenta este etiquetado para la reducción de dimensionalidad y la distancia entre los puntos. Por medio de un gráfico se observa el conjunto de datos de entrenamiento en el espacio de dimensión reducido y luego un gráfico con el conjunto de testeo transformado. En este caso los datos por el etiquetado pertenecen ya a un grupo donde se realiza una estadística descriptiva para compararlos.

3.2.2. No supervisado

En este caso no se hace entrega del etiquetado al algoritmo, por lo que realiza la reducción de dimensionalidad basado totalmente en las variables del conjunto de datos. En este tipo de aplicación se hace uso del algoritmo HDBSCAN.

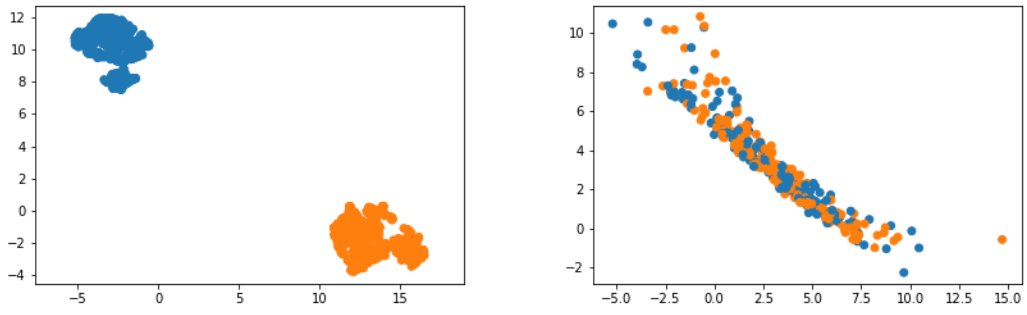
3.3. Aplicación HDBSCAN

Se aplica el algoritmo HDBSCAN a los datos transformados por el UMAP no supervisado para establecer clústers y asignar una etiqueta a cada punto, ya sea que el punto pertenezca a un clústers o se etiquete como un punto de ruido. Para los dos grupos encontrados se realiza también una estadística descriptiva con fines comparativos.

4. Resultados

4.1. UMAP supervisado

Después de la aplicación del UMAP supervisado se puede obtener el gráfico de baja dimensión, ver figura 1 (a); en este se puede observar como el algoritmo incrusta los datos de entrenamiento en el nuevo espacio, se evidencia que existe una gran separación entre los puntos de color naranja que están etiquetados con 1 (pasaron las pruebas de selección) y los puntos de color azul etiquetados con 0 (no pasaron las pruebas de selección), a estos grupos se les llamará grupo 1 y grupo 0 respectivamente.



(a) Entrenamiento

(b) Testeo

Figura 1: UMAP Supervisado

En la figura 1 (b) se observa como UMAP transforma en el espacio de baja dimensión los datos de testeo, que son datos nuevos para el algoritmo. En este caso no se logra realizar una separación adecuada como en el conjunto de entrenamiento, los puntos se ven juntos y la forma del gráfico es en gran medida diferente al anterior.

Este no es un resultado que se espere, pues UMAP no está logrando realizar una separación en los individuos nuevos, algunas razones para esto pueden ser que el algoritmo se está sobre entrenando o la falta de variables en los datos que proporcionen mayor información o relación con las etiquetas. En

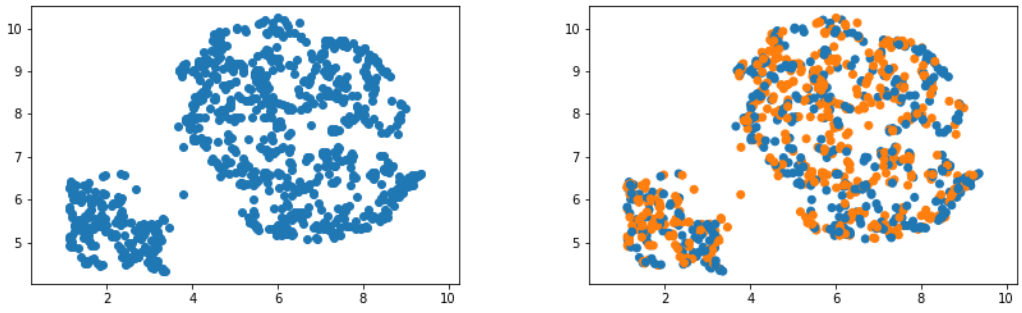
la estadística descriptiva se encontraron los siguientes resultados para los dos grupos en cuanto a los cuantiles de cada una de las variables.

Variable	Grupo	Min.	Q1	Q2	Q3	Max.
MC	1	37	55	59	62	100
	0	42	55	58	60	90
MG	1	34	47	51	56	90
	0	35	46	50	55	109
RB	1	31	60	64	70	86
	0	32	61	65	70	85
DS	1	10	289	300	300	400
	0	5	209	300	300	400
PC	1	5	60	112	180	300
	0	8	54	100	160	252
LS	1	200	7050	11500	15100	42150
	0	550	8800	12750	16150	41000
LA	1	400	8500	10500	12000	18400
	0	800	9000	11000	12000	19300

Para las primeras 4 variables se observa que no existe alguna diferencia fuerte, pues los cuantiles están muy cerca entre los dos grupos. Para la quinta variable “PC” si existe una diferencia más pronunciada en sus cuantiles, se evidencia que para el grupo 1 los valores tienden a ser más altos. Para las últimas dos variables las cuales tienen un rango más amplio de valores, también se encuentra un poco de diferencia, pero en esta ocasión es el grupo 0 el que tiende a tener puntuaciones más altas.

4.2. UMAP No supervisado y HDBSCAN

En esta aplicación del algoritmo no supervisado, donde no se tiene en cuenta la variable de etiquetas se puede ver nuevamente el gráfico en baja dimensión figura 2 (a). Aunque este es diferente del gráfico obtenido en la aplicación anterior, también es posible observar la separación de los datos en dos grupos.



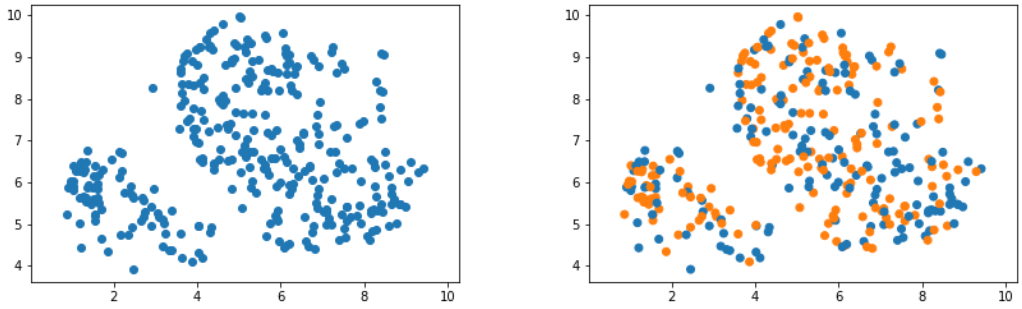
(a) Sin Etiqueta

(b) Con Etiqueta

Figura 2: UMAP No supervisado

Teniendo en cuenta esta gráfica, representando los grupos con color diferente de acuerdo a su eti-

queta, se observa en la figura 2 (b) que los colores están totalmente mezclados, lo que sugiere que el algoritmo está realizando una separación de los datos, pero no está relacionada con si los individuos pasan o no las pruebas de selección. Ahora aplicando esta transformación a los datos de testeo se observa en la figura 3 que los gráficos son similares a la figura 2, pues se evidencian los mismos grupos. Este resultado indica que el algoritmo tiene un buen comportamiento en los datos nuevos y encuentra una forma de separarlos.



(a) Sin Etiqueta

(b) Con Etiqueta

Figura 3: UMAP No supervisado

Con el objetivo de encontrar la base de la separación de los datos se hace uso del algoritmo HDBSCAN, por medio de este se logra etiquetar los individuos según el grupo al que pertenecen en la transformación realizada por el UMAP no supervisado.

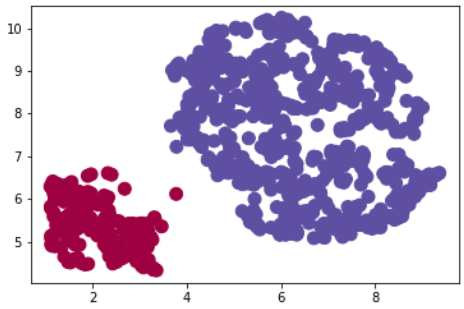


Figura 4: HDBSCAN

Se puede ver en la figura 4 que se logra realizar una buena clasificación sin puntos de ruido, en esta ocasión el grupo con color granate será llamado grupo A y el grupo con color violeta será llamado grupo B. Para estos dos grupos se realiza la misma estadística descriptiva con lo que se obtienen los siguientes valores.

Variable	Grupo	Min.	Q1	Q2	Q3	Max.
MC	A	37	55	58	62	100
	B	42	55	58	60	72
MG	A	34	47	51	56	109
	B	36	45	50	55	75
RB	A	31	59	64	69	86
	B	45	62	66	70	83
DS	A	77	300	300	300	400
	B	5	96	134	180	245
PC	A	5	60	110	179	300
	B	9	50	94	150	290
LS	A	400	7500	12000	15500	42150
	B	200	10050	13250	16600	24200
LA	A	800	8500	10500	12000	19300
	B	400	9500	11000	12500	13500

Se observa que para las tres primeras variables ocurre lo mismo que en los grupos 1 y 0 descritos anteriormente, el valor de sus cuantiles es muy cercano; en la variable 4 “DS” se encuentra tal vez la mayor diferencia, el grupo A tiende a ser mayor notoriamente, por lo que esta variable puede ser la base que el algoritmo encuentra para hacer una diferencia entre los puntos. Para las últimas 3 variables se puede notar que los cuantiles son más distantes, donde para la variable 5 “PC” el grupo A tiende a tener valores más altos y para las últimas dos variables “LS” y “LA” el grupo B tiende a tener valores más altos.

4.3. Discusión

Para los dos modelos implementados se desea encontrar características que distingan los grupos.

Los grupos del primer modelo muestran poca diferencia entre ellos, lo que sugiere la falta de información en los datos relacionada con la etiqueta, y para los datos de testeo no se obtiene un buen resultado.

Respecto al segundo modelo, los grupos obtenidos muestran una diferencia respecto a la variable 4 "DS"donde gran parte de los datos tiene un valor de 300, todos los individuos con este puntaje o mayor pertenecen al grupo A. En esta ocasión se puede observar un buen comportamiento en el conjunto de datos de testeo, lo que sugiere que el modelo está aprendiendo bien del conjunto de entrenamiento.

Si bien el segundo modelo no nos ofrece una distinción entre los individuos que son seleccionados y los que no, nos permitió encontrar una diferencia entre ellos basada en el puntaje de la variable 4, es entonces esta variable la que nos genera una diferencia considerable entre los individuos.

5. Conclusiones y Recomendaciones

En el presente trabajo se usó el algoritmo UMAP, un algoritmo para la reducción de dimensionalidad no lineal y el algoritmo HDBSCAN, un algoritmo para establecer clúster.

Primero se aplicó el algoritmo UMAP supervisado, donde se obtiene una buena separación en los datos de entrenamiento, pero no una buena separación en el conjunto de testeo. No se logra evidenciar diferencias fuertes entre los grupos, por lo que se decide aplicar el algoritmo UMAP no supervisado y con esto el algoritmo HDBSCAN para establecer grupos y etiquetas.

En este segundo modelo se logra establecer dos grupos, diferenciados fuertemente por una variable y se logra un buen comportamiento en los datos de testeo. La variable que diferencia los grupos ha sido llamada “DS” donde los individuos con valores iguales o mayores a 300 pertenecen al grupo A, mientras que para los valores menores en su mayoría pertenecen al grupo B.

Para futuras investigaciones se recomienda el uso de más variables que también se tiene en cuenta en los procesos de selección de aspirantes, variables que nos den más información sobre los individuos, como por ejemplo las variables categóricas las cuales podrían aportar considerablemente al modelo, y con esto llegar a tener predicciones que apoyen la toma de decisiones.

Referencias

[1] Adam Pearce Andy Coenen. Understanding umap. https://pair-code.github.io/understanding-umap/.

[2] John Healy Leland McInnes. How hdbscan works. https://hdbscan-.readthedocs.io/en/latest/how_hdbscan_works.html, 2016.

[3] Leland McInnes. How umap works. https://umap-learn-.readthedocs.io/en/latest/how_umap_works.html, 2018.

[4] Abdelhakim Cheriet Mebarka Allaoui1, Mohammed Lamine Kherfi2. Considerably improving clustering algorithms using umap dimensionality reduction technique: A comparative study. *Springer Nature*, 2020.

[5] Guido van Rossum. Lenguaje de programación python versión 3.