ESTIMACIÓN DE LAS PRINCIPALES CAUSAS DE LA DESERCIÓN UNIVERSITARIA MEDIANTE EL USO DE TÉCNICAS DE MACHINE LEARNING

ESTIMATION OF THE MAIN CAUSES OF UNIVERSITY DROPOUTS THROUGH THE USE OF MACHINE LEARNING TECHNIQUES

Juber Orlando Gutiérrez Villarreal[1]
Lida Rubiela Fonseca Gómez[2]
Wilmer Pineda-Ríos [3]

Resumen

Según el reporte del banco mundial para el año 2015, la deserción estudiantil universitaria en Colombia había llegado a una tasa del 42%, ocupando el segundo lugar en Latinoamérica, lo que refleja una problemática social y económica de gran impacto. De acuerdo con lo anterior, el presente estudio realiza un análisis para identificar las causas que influyeron en la deserción estudiantil de una Institución de Educación Superior en la ciudad de Bogotá, aplicando técnicas computacionales como el Machine Learning. Para la identificación de las causas se emplearon técnicas del aprendizaje no supervisado como el análisis de componentes principales (PCA) y la descomposición en valores singulares (SVD) para la reducción de la dimensionalidad a 24 componentes con el 71% de la variabilidad explicada y el algoritmo de k-means posibilitó el agrupamiento en tres grupos con información recolectada de 207 estudiantes que desertaron durante el 2020. Este análisis permite establecer patrones de asociación entre algunas variables como los aspectos institucionales, encontrando semejanzas y diferencias, además de la identificación de factores de deserción relevantes en la población, siendo las dificultades económicas la principal causa de deserción en hombres (67.6%) de los programas de ingeniería mecánica y ambiental.

Palabras clave: Deserción, muestreo, multivariado, Machine Learning, componentes principales, agrupamiento, k-means.

Abstract

According to the World Bank report for 2015, university student dropout in Colombia had reached a rate of 42%, ranking second in Latin America, reflecting a social and economic problem of great impact. In accordance with the above, the present study carries out an analysis to identify the causes that influenced student desertion from a Higher Education Institution in the city of Bogotá, applying computational techniques such as Machine Learning.To identify the causes, unsupervised

Recepción: 20 de septiembre / Evaluación: 05 de octubre / Aprobado: 15 noviembre de 2021

[1]Estudiante de Maestría en Estadística aplicada. Especialista en Estadística Aplicada. Ingeniero Industrial. Licenciado en Matemáticas y Física. Analista de datos. Docente de Matemáticas y Física en secretaría de educación de Cundinamarca . jubergutierrez@usantotomas.edu.co. ORCID: https://orcid.org/0000-0003-4476-1445

[2] Estudiante de Doctorado en Educación Universidad Pedagógica, Docente Estadística, Universidad Santo Tomás, Fundación Universitaria Los Libertadores. Magíster en Educación, Especialista en Diseño de Ambientes de Aprendizaje lidafonseca@usantotomas.edu.co. ORCID: https://orcid.org/0000-0002-3597-728X.

[3] Estudiante de Doctorado en Estadística Universidad Nacional De Colombia. Magister en matemáticas de la Universidad Nacional de Colombia. Conocimientos en Diseño de Experimentos, Modelamiento Estadístico, Procesos Estocásticos y Datos Funcionales. Experiencia como Consultor Estadístico en proyectos de investigación académicos. wpinedar@unal.edu.co. ORCID: https://orcid.org/0000-0001-7774-951X

learning techniques were used, such as principal component analysis (PCA) and singular value decomposition (SVD) to reduce the dimensionality to 24 components with 71% of the variability explained and the k-means algorithm made it possible to group into three groups with information collected from 207 students who dropped out during 2020. This analysis allows establishing patterns of association between some variables such as institutional aspects, finding similarities and differences, in addition to identifying factors of relevant desertion in the population, with economic difficulties being the main cause of desertion in men (67.6%) of the mechanical and environmental engineering programs.

Key words: Dropout, sampling, multivariate, Machine Learning, principal components, Clustering, k-means.

Introducción

La educación es uno de los pilares del desarrollo de un país al involucrar directamente aspectos económicos, políticos, sociales, culturales, sobre todo aspectos productivos, generando más y mejores oportunidades laborales para la población, esto de acuerdo con Patiño y Cardona (2012). Por tanto, las Instituciones de Educación Superior (IES) tienen el reto de retener a sus estudiantes hasta lograr la graduación en los diferentes programas académicos, sin embargo, una de las problemáticas que han tenido las IES en Colombia y Latinoamérica durante las últimas décadas es la deserción estudiantil, provocando pérdida de la productividad laboral de calidad por la disminución de la acumulación de capital humano, lo cual afecta el desarrollo de la Nación. (Vélez et al., 2008).

La deserción estudiantil es un fenómeno que no solo involucra a los estudiantes, también incluye a las IES, las familias y a la nación en general, teniendo en cuenta que el desarrollo intelectual de sus ciudadanos es el potencial económico de un país. Es necesario identificar la deserción como un fenómeno que requiere atención inmediata y como lo menciona el Ministerio de Educación Nacional – MEN- Vélez et al. (2008) al citar a Tinto (1975), cuando se afirma que "el estudio de la deserción en la educación superior es extremadamente complejo, ya que implica no solo una variedad de perspectivas sino también una gama de diferentes tipos de deserción" (Vélez, 2008, p 20). Según Tinto (1975) y Giovagnoli (2002), citados en la Metodología de seguimiento, diagnóstico y elementos para su prevención en el estudio de deserción estudiantil en la educación superior colombiana, establece que:

> La deserción como una situación a la que se enfrenta un estudiante cuando aspira y no logra concluir su proyecto educativo, considerándose como desertor a aquel individuo que siendo estudiante de una institución de educación superior no presenta actividad académica durante dos semestres académicos consecutivos, lo cual equivale a un año de inactividad académica (Vélez, 2008, p. 22).

De acuerdo a las investigaciones realizadas por diferentes autores, las causas que influyen en la deserción se pueden caracterizar en cuatro grupos, causas individuales, académicas, institucionales y socio-económicas. En consecuencia, es indispensable promover proyectos de investigación que contribuyan a mitigar dicha problemática y permitan aumentar la retención de estudiantes. Por tal razón, las técnicas computacionales como la Machine Learning (ML) son herramientas bastante eficaces para identificar la vulnerabilidad de un estudiante respecto a la deserción educativa (Kuvcak, 2018).

Actualmente la Institución de Educación Superior donde se realizó el presente estudio, cuenta con un plan estratégico para el fortalecimiento de la permanencia estudiantil, el cual incluye una política de retención que contiene actividades pertinentes a la situación actual de los estudiantes, sin embargo, la deserción continua (Castiblanco, 2020). Según el Sistema de prevención y análisis de la deserción en las instituciones de educación superior (SPADIES) en Colombia, las IES tienen una deserción promedio del 13%, lo que requiere una intervención urgente con el fin de disminuir esta tasa. Es importante considerar que existen tecnologías emergentes en el entorno educativo que permiten modelar el comportamiento humano (Bacos, 2019), y que aplicarlas podría aportar a disminuir la deserción y focalizar recursos a los programas de retención y permanencia estudiantil.

Vale la pena destacar que no se encontraron estudios en la IES donde se identifique el motivo de deserción de los estudiantes, por tanto, no fue posible identificar causas de la problemática en estos espacios educativos.

De esta manera, con el fin de identificar las causas por la cuales un gran número de estudiantes desertó, se inició con la identificación de la población desertora durante el año 2020 a través de las bases de datos suministradas con las que se construyó un diseño muestral, luego se aplicó un instrumento para conocer la percepción de estos estudiantes respecto a los motivos que los llevaron a tomar esta decisión.

Dentro del proceso de análisis estadístico se estableció un primer acercamiento descriptivo, que incluyó un análisis multivariado para identificar el comportamiento de la población, al igual que la aplicación de técnicas comunes y funcionales del aprendizaje no supervisado, como la reducción de la dimensionalidad y la agrupación de estudiantes con características que permitieron explorar detalladamente las causas que llevaron a no continuar con sus estudios de tecnología y/o pregrado.

Metodología

El tipo de estudio corresponde a una investigación no experimental aplicada, la cual busca identificar estrategias adecuadas para definir las causas que influyeron en la deserción estudiantil en la IES durante el año 2020. La investigación tiene un componente explicativo y de asociación con un enfoque metodológico cuantitativo de nivel descriptivo. (Hernández y Mendoza, 2018).

Dentro del proceso investigativo, se tiene información primaria de una Institución de Educación Superior de la ciudad de Bogotá, con registros socioeconómicos, académicos, financieros, políticas de deserción, informe SPADIES, entre otros.

Población desertora

En primera instancia, se solicitó información a la IES, donde se le realizó un tratamiento y depuración a las bases de datos y se clasificaron las variables definidas, de esta manera, se determinó la población que desertó durante el año 2020.

Se tomó información de estudiantes matriculados y graduados en el año 2020, identificando una población de 1.463 estudiantes de diferentes programas académicos, con lo que se calculó una muestra EST_MAS con remplazo de 207 estudiantes, con un nivel de confianza del 94% y un margen de error del 6%, una probabilidad de éxito y fracaso del 50% y un factor de expansión de 6.96.

Instrumento y encuesta

Para la recolección de la información se contactó telefónicamente a los estudiantes que por dos periodos consecutivos no renovaron su matrícula y se les solicitó responder un instrumento de 115 preguntas elaborado en formato "*Google Forms*", respecto al motivo del retiro de la institución. La validación del instrumento se realizó a partir de pruebas piloto desarrolladas con 10 estudiantes de otra IES, y con análisis de expertos de la institución donde se realizó el estudio. Los datos fueron recolectados durante el primer semestre del 2021.

Análisis descriptivo

El análisis descriptivo y multivariado de la información recolectada permitió determinar asociaciones entre las diferentes variables a partir del análisis de correspondencia y de correlaciones. Se encontró que, de los 207 estudiantes encuestados, 67 fueron mujeres (32.4%) y 160 hombres (67.6%). La edad de los estudiantes oscila entre 18 y 46 años con media de 25.5 años y una desviación estándar (DE) de 4.9. La variable "problemas económicos" fue el factor que más influyó en la deserción, principalmente en estudiantes de estrato 2 (46%), como se muestra en la tabla 1.

Tabla 1.

Variables sociodemográficas y principal razón de abandono de la universidad

Análisis Descriptivo			
		Frecuencia	Porcentaje
Sexo	Femenino	67	32.4%
	Masculino	140	67.6%
Razón de abandono	Dificultades académicas	7	3%
	Otra	48	23%
	Problemas económicos	135	65%
	Problemas familiares	9	4%
	Programa no expectativas	8	4%
Actualidad del estudiante	Desempleado(a)	47	23%
	Estudia	11	5%
	Trabaja	89	43%
	Trabaja y estudia	60	29%
Estrato socio Económico	1	21	10%
	2	96	46%
	3	77	37%
	4	12	6%
	5	1	0.5%

Fuente: elaboración propia.

En cuanto al promedio académico de los estudiantes desertores, se tiene una media general de 3.37, para las mujeres de 3.55 (DE=0.71) y para los hombres de 3.28 (DE=0.68), indicando que las mujeres tienen mejor promedio académico. El máximo número de materias reprobadas por un estudiante desertor fue de 7 como se observa en la tabla 2.

Tabla 2.
Análisis descriptivo de las variables numéricas

	Edad	Materias Inscritas	Materias Reprobadas	Promedio	Numero hijos	Personas cargo
Count	207	207	207	207	207	207
mean	25,497585	5,014493	1,410628	3,375604	0,323671	0,908213
std	4,968712	2,343092	1,824984	0.705398	0,643542	1,073186
min	18	1	0	0.390000	0	0
25%	22	3	0	2,995	0	0
50%	25	5	1	3,46	0	1
75%	28	7	2	3,855	0	2
max	46	11	7	4,73	3	5

Fuente: elaboración propia.

Reducción de dimensionalidad

Para el análisis y procesamiento de la información se emplearon los softwares estadísticos R y Python, empezando por la limpieza y depuración de la base, pasando de 115 a 70 variables, de las cuales 6 variables fueron numéricas y 64 categóricas. Esta primera reducción de la información se realizó a partir de la unificación variables, la generación de nuevas, y reconociendo variables socioeconómicas, personales, académicas e institucionales.

A partir de los algoritmos de Machine Learning, se llevó a cabo el análisis de los factores asociados a la deserción, empleando específicamente, el aprendizaje no supervisado. Esta técnica inicia con la reducción de la dimensionalidad con el propósito de reducir el número de variables a un grupo específico (Toledo y Pereira, 2015).

Por tal motivo, se exploraron técnicas que emplean estrategias diferentes, como el análisis de componentes principales (ACP), ACP incremental, ACP disperso, Kernel PCA, descomposición en valores singulares (SVD), proyección aleatoria, isomap e incrustación de vecinos estocásticos distribuidos en t (t-SNE) (Bravo, 2020), con la finalidad de determinar cuál era la más adecuada con los datos obtenidos. Es así como, algunas de ellas emplean la reducción de dimensionalidad lineal, donde el algoritmo utilizado permite encontrar una representación de baja dimensión de los datos mientras retiene la mayor cantidad de variación posible (Sánchez, et al., 2008). Cabe resaltar que la maldición de la dimensionalidad hace que se afecten los datos al incrementar la cantidad de dimensiones, donde el volumen de espacio se incrementa exponencialmente, haciendo que dichos datos se vuelvan más dispersos.

Otras técnicas basadas en la reducción no lineal, calculan una medida de similitud basada en la distancia entre puntos en lugar de intentar maximizar la varianza, en donde su análisis tiene un carácter más gráfico y visual. En este caso en particular, las técnicas que se basaron en una combinación lineal, de fácil interpretación y uso, y que además se ajustaron al comportamiento de los datos, fueron el ACP y la SVD. Estas técnicas permitieron reducir a 24 componentes principales, recogiendo un **71%** de la varianza explicada. Se decidió emplear las dos ya que el APC tiene una amplia aplicación en el análisis exploratorio de los datos que simplifica la complejidad de espacios muestrales con muchas dimensiones y proporciona una interpretación probabilística de los datos en función de la cantidad de varianza que explica.

El análisis de componentes principales (PCA) busca generar nuevos ejes que se dirigen en las direcciones de la máxima varianza y la minimización del error cuadrático (Bravo, 2020). Este mismo autor afirma que los PCA se pueden usar para la reducción de dimensión si no se utilizan

todos los componentes principales para representar los datos, tal como se plantea en el presente estudio. Según Jolliffe (2002), citado por Sánchez, et al., 2008.

El ACP es una técnica que emplea la simplicidad y capacidad de reducción de dimensión, minimizando el error cuadrático de reconstrucción producido por una combinación lineal de variables latentes, las cuales se obtienen a partir de una combinación lineal de los datos originales. Los parámetros del modelo pueden calcularse directamente de la matriz de datos centralizada $\boldsymbol{X}$, bien sea por descomposición en valores singulares o por la diagonalización de la matriz de covarianza Sea xi el i-ésimo vector de observación (vector columna) de tamaño $\boldsymbol{c}$, $\boldsymbol{X} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n})^{\boldsymbol{T}}$. La matriz de rotación $\boldsymbol{U}$ permite calcular las p componentes principales **z** que mejor representan **x**.

$$\boldsymbol{z} = \boldsymbol{U^T x}$$

$\boldsymbol{U}$ puede obtenerse al solucionar un problema de valores propios, y está definida como los p mayores vectores propios de $\boldsymbol{X^T X}$, esto es,

$$\boldsymbol{X^T X U} = \boldsymbol{n U \Lambda}$$

La matriz $\boldsymbol{X^T X}$ está asociada a la matriz de covarianza $\boldsymbol{C} = \frac{\boldsymbol{1}}{\boldsymbol{n}} \boldsymbol{X^T X}$; además, puede calcularse como el estimado de la matriz de covarianza de los datos

$$\boldsymbol{C} = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{x_i})(\boldsymbol{x_i})^{\boldsymbol{T}}$$

El problema de valores propios $\boldsymbol{Cu} = \boldsymbol{\lambda u}$ implica que todas las soluciones de $\boldsymbol{u}$ deben estar en el espacio generado por el conjunto de vectores $x_1, x_2, \ldots, x_n$; por lo cual (Schölkopf et al. 1996),

$$\boldsymbol{\lambda}(\boldsymbol{x_i}, \boldsymbol{u}) = (\boldsymbol{x_i}, \boldsymbol{Cu}), \quad \forall \boldsymbol{i} = \boldsymbol{1}, \ldots, \boldsymbol{n}.$$

La aplicación de la técnica SVD, permite reducir los conjuntos de datos que contienen un gran número de valores y genera una matriz de rango más pequeño, adicionalmente, posibilitó una mejor interpretación gráfica de los clusters. Para Steward, (1993) citado por Giraldo (2021) la SVD es,

> "una reducción de forma cuadrática a forma diagonal mediante cambios de base ortogonales, siendo $\boldsymbol{A} \in \mathbb{R}^{\boldsymbol{nxp}}$, entonces existe una matriz ortogonal $U \in \mathbb{R}^{\boldsymbol{nxn}}$ y una matriz ortogonal $\boldsymbol{V} \in \mathbb{R}^{\boldsymbol{pxp}}$ tal que
>
> $$\boldsymbol{A} = \boldsymbol{U} \sum',$$
>
> donde $\sum$ en este caso es llamado matriz de valores singulares y es una matriz diagonal de entradas $\boldsymbol{\sigma_j} \geq \boldsymbol{0}\ \boldsymbol{con}\ \boldsymbol{j} = \boldsymbol{1}, \boldsymbol{2}, \ldots, \mathbf{min}(\boldsymbol{n}, \boldsymbol{p})$, y las constantes $\boldsymbol{\sigma_j}$ son los valores singulares de $\boldsymbol{A}$.
>
> Se define el rango de la matriz rank$(\boldsymbol{A})$ como el número de valores singulares no nulos de la matriz $\boldsymbol{A} \in \mathbb{R}^{\boldsymbol{nxp}}$ Si todas las filas y columnas de la matriz $\boldsymbol{A}$ son linealmente independientes entonces $\boldsymbol{r} = \boldsymbol{rank}(\boldsymbol{A}) = \mathbf{min}\{\boldsymbol{n}, \boldsymbol{p}\}$.; La descomposición en valores singulares también puede ser expresada como una expansión de la matriz $\boldsymbol{A}$ que depende

del rango r de la matriz. Esto es expuesto en el teorema de Eckart - Young, empleado en el análisis de componentes principales tanto lineales como no lineales".

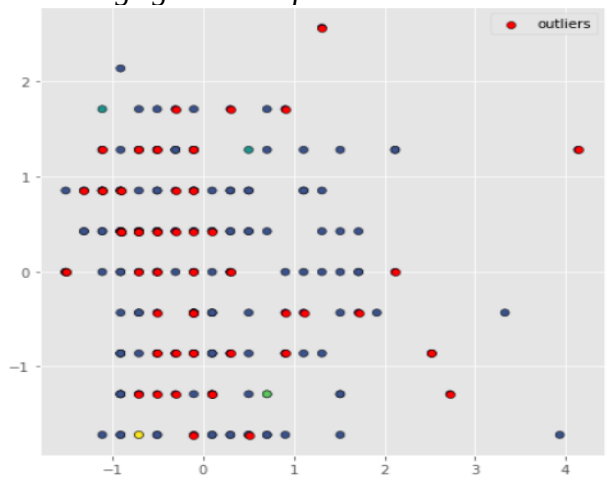
Clusterings (Agrupación)

Se procedió a la aplicación de las técnicas de agrupamiento por medio de los algoritmos de DBSCAN y K-Means, para establecer algunos factores comunes de agrupación. Después de realizar un análisis multivariado para determinar el comportamiento de los datos, se procedió a la aplicación de estas metodologías, las cuales precisaron la conformación de 3 grupos de acuerdo al gran número de variables categóricas existentes. El algoritmo de agrupamiento espacial de aplicaciones con ruido basado en la densidad DBSCAN, intenta agrupar puntos de datos similares en grupos o agrupaciones artificiales (Mysiak, 2020).

Este algoritmo generó una agrupación que no permitía identificar las características de deserción, construyendo 4 grupos mal representados. El agrupamiento por DBSCAN identificó un primer grupo de 141 estudiantes, el segundo con 3, el tercero y cuarto grupo con dos estudiantes cada uno. También identificó 59 datos atípicos (puntos de color rojo de la figura 1) los cuales son puntos que se encuentran solos en regiones de baja densidad, lo que es una particularidad del algoritmo. Esta agrupación se generó con los parámetros de épsilon (eps = 8.8), puntos mínimos (min_samples = 1.8), empleando la métrica euclidiana y fueron con los que mejor funcionó ya que al entrenarlo con otros valores en los parámetros, se generaba un solo grupo o una gran cantidad de grupos. Esta organización se pudo haber producido tal vez, por la alta dimensionalidad y el gran número de variables categóricas que tenía la base o la sensibilidad en los parámetros. Ver figura 1.

Figura 1.

Clusterings generados por DBSCAN



Fuente: elaboración propia.

En consecuencia, debido a la débil agrupación generada por DBSCAN, se procedió a emplear el algoritmo K-means el cual fue el que mejor agrupó los datos (Suárez, 2015). Este algoritmo agrupa los datos tratando de separar muestras en n grupos de igual varianza y minimizando la inercia o suma de cuadrados dentro del grupo:

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

La inercia se puede reconocer como una medida de la coherencia interna de los clústeres.

Este algoritmo realiza el proceso en tres pasos:

1. Inicia escogiendo el número de grupos K, generando k centroides en el grupo de datos.

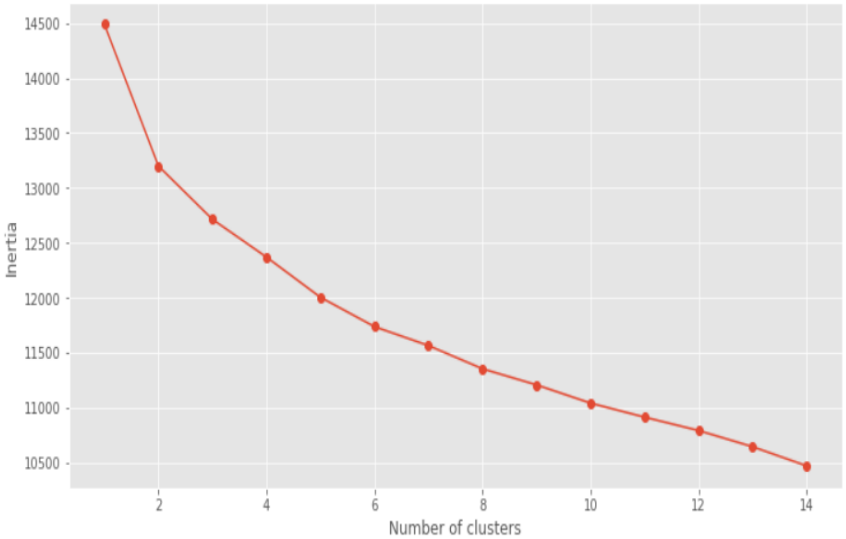
2. Asigna objetos a los centroides más cercanos.

3. Finalmente, actualiza la posición de los centroides empleando el promedio de cada de cada grupo, repitiendo los pasos anteriores hasta que los centroides no se muevan, resolviendo el problema de optimización (Martínez, 2022). Clustering (Agrupamiento), K-Means con ejemplos en Python. https://www.iartificial.net/

Luego, se procedió a la aplicación de la Técnica Estadística del Codo (Elbow Method), la cual calcula la suma de las distancias al cuadrado desde cada punto hasta su centroide asignado para cada iteración de K-Means (Nainggolan, 2019), indicando que se tomaran 3 grupos. También, se analizó la distancia de la silueta, la cual señala que entre mayor distancia mejor clasificación entre los diferentes grupos, encontrándose, de manera específica para este caso una distancia de la silueta de 0.07.

Figura 2.

Método de Elbow para determinar el número óptimo de grupos



Fuente: elaboración propia.

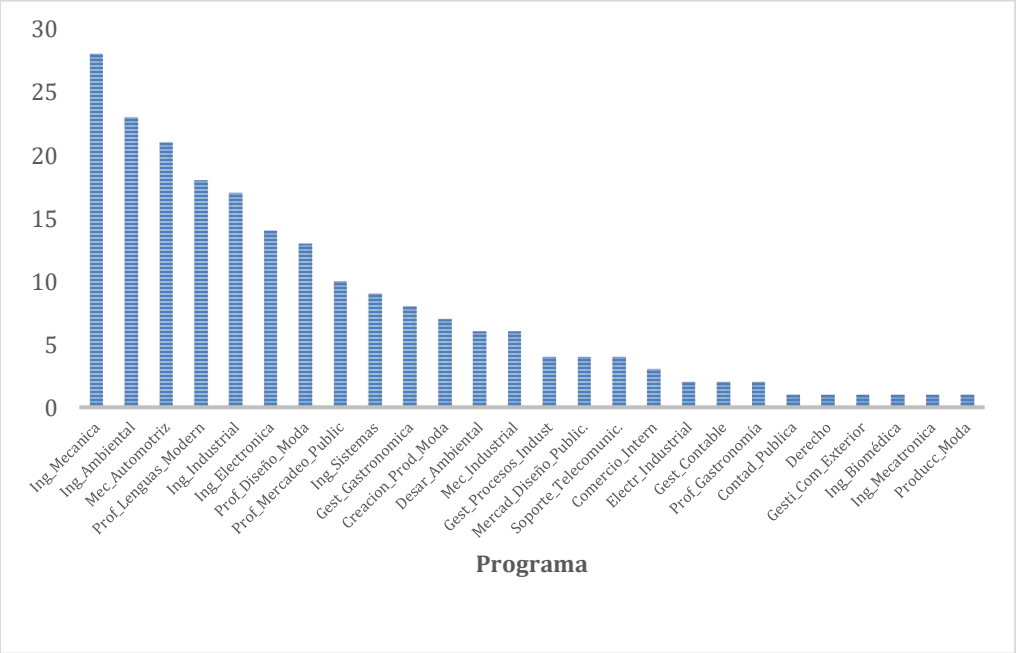
Posteriormente, una vez realizada la reducción de dimensionalidad a 24 componentes, se procedió de nuevo, a aplicar el método del codo para definir el número de grupos a conformar, el cual indicó nuevamente 3 grupos, con un comportamiento similar al conjunto de datos originales.

En cuanto a la distancia de la silueta, esta aumentó a 0.105, lo que determina que existe una mayor distancia entre los grupos con los componentes principales, por tanto, genera un mejor agrupamiento entre los mismos.

Resultados y Discusión

El análisis descriptivo posibilita un acercamiento al comportamiento de las variables que inciden en la deserción, teniendo como resultado 207 estudiantes que desertaron durante al año 2020, principalmente de los programas de ingeniería mecánica y ambiental, en su mayoría hombres (67.6%) entre los 18 y 25 años, de acuerdo con la figura 3.

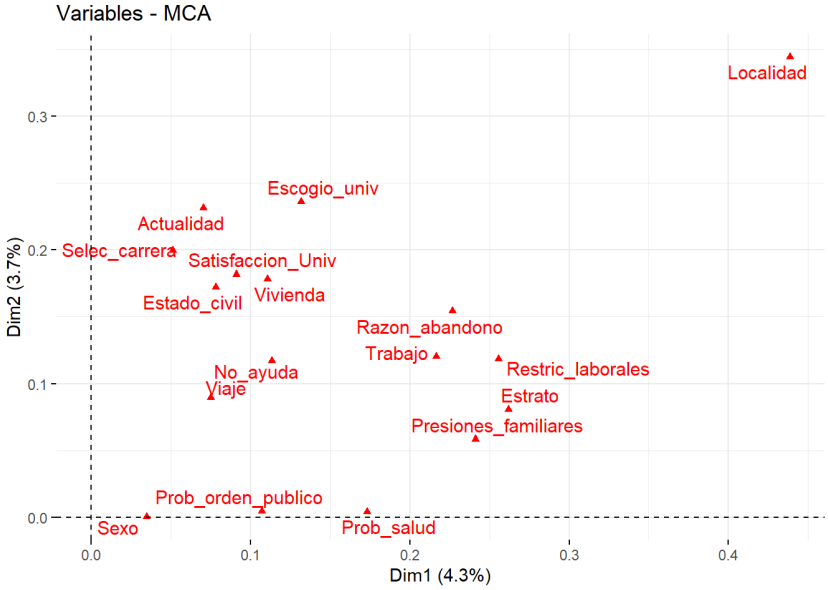
Figura 3.
Distribución de estudiantes que desertaron por programa académico



Fuente: elaboración propia.

En la exploración de los datos se realizó un análisis multivariado, tomando como referente la variable “razón de abandono” y observando su comportamiento con las demás variables como se observa en la figura 4.

Figura 4.
Análisis de correspondencia multivariado

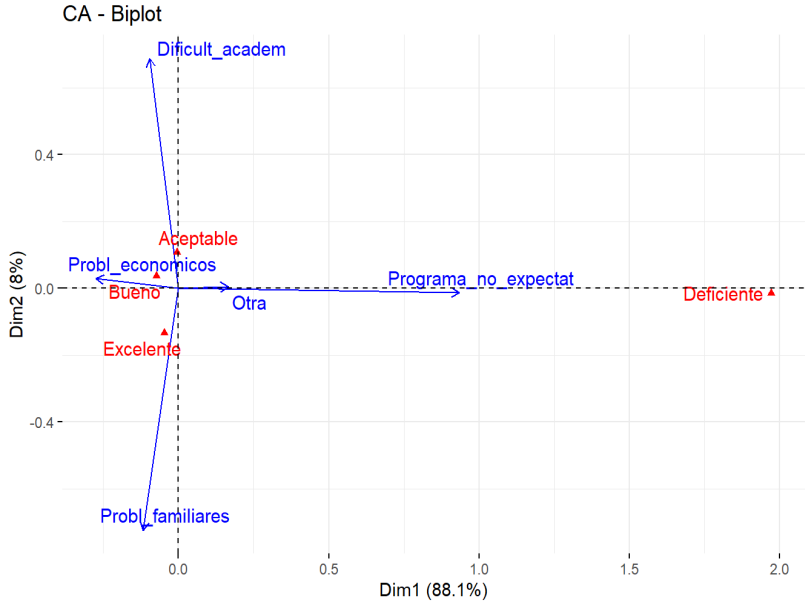


Fuente: elaboración propia.

El gráfico muestra la relación entre variables, donde la razón de abandono de los estudios se asocia con motivos laborales; el estrato, el cual tiene una alta asociación con las presiones familiares; la satisfacción con la universidad, también está asociada con el tipo de vivienda y el estado civil. Por otra parte, se puede establecer que la localidad donde vivían los estudiantes es independiente de la razón de abandono.

Al relacionar la calificación sobre la satisfacción con la universidad como razón de abandono, se encontró que existe una asociación entre los estudiantes que manifestaron inconformismo con la universidad (deficiente) y la razón de que el programa donde se encontraban no cumplió con las expectativas, mientras que aquellos que calificaron con “bueno”, son aquellos que tuvieron problemas económicos.

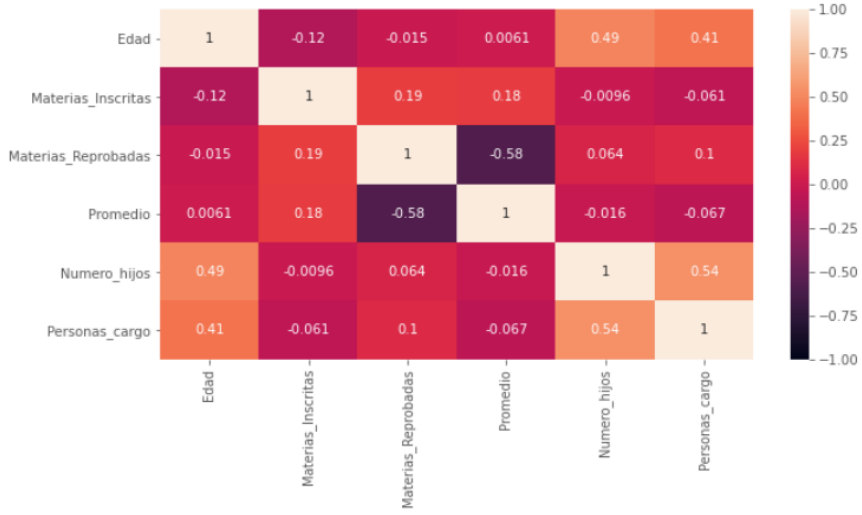
Figura 5.
Análisis de correspondencia de la satisfacción de la universidad



Fuente: elaboración propia.

En cuanto al análisis de correlación entre las variables numéricas, se evidencia que la edad y el número de hijos tiene una mediana correlación (49 %), mientras que el promedio y el número de materias reprobadas tiene una correlación inversa y medianamente alta (58.4 %)

Figura 6.
Matriz de correlación de las variables numéricas

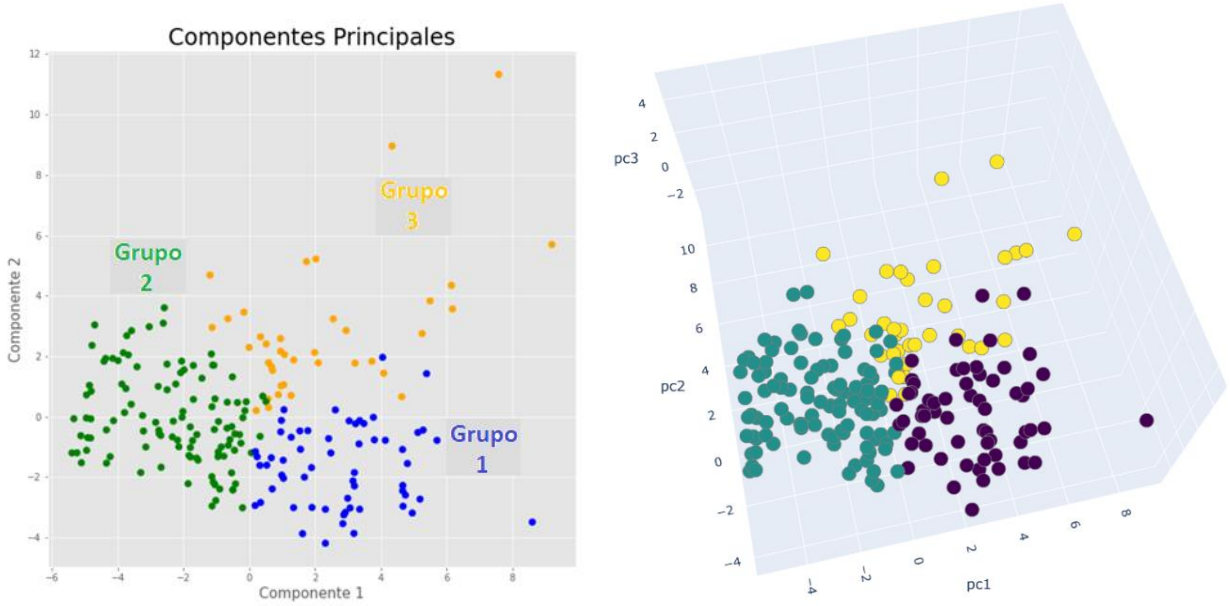


Fuente: elaboración propia.

Para el caso del presente estudio, la técnica del aprendizaje no supervisado, permitió reducir la dimensionalidad de una manera práctica y sencilla, fue el análisis de componentes principales y el algoritmo que permitió agrupar las variables maximizando su similitud en un grupo, y

minimizando la similitud entre grupos fue k-Means, obteniendo 3 grupos como se evidencia en la siguiente en la figura 7.

Figura 7.
Representación gráfica de los grupos en dos y tres dimensiones



Fuente: elaboración propia.

El gráfico de la izquierda muestra la proyección de los 3 grupos en las dos primeras componentes principales, y el gráfico de la derecha muestra los mismos grupos, pero con una proyección en las tres primeros componentes.

De acuerdo con la clasificación en 3 grupos dada por K-means sobre los componentes principales, el grupo 1 (0 = azul) tiene 60 estudiantes, el grupo 2 (1 = verde) tiene 109 estudiantes y el grupo 3 (2 = amarillo) tiene 38 estudiantes, este último presenta un comportamiento atípico.

A partir del análisis gráfico de cada uno de los grupos se establecieron diferentes causas de deserción, lo que permitió encontrar similitudes y diferencias en cada grupo y la siguiente clasificación: el grupo 1 se denominó, estudiantes inconformes con hijos, el grupo 2, estudiantes mayores satisfechos y el grupo 3, estudiantes con dificultades económicas con hijos.

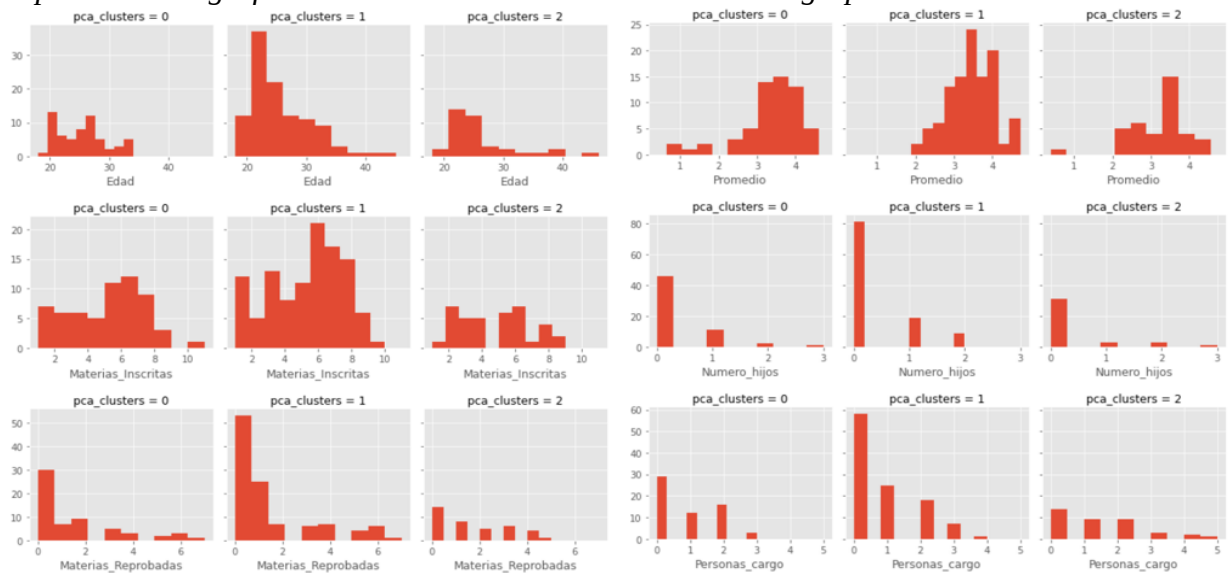
La clasificación de los grupos permite caracterizar diversos aspectos con los cuales el algoritmo los agrupó, por ejemplo, el grupo de estudiantes inconformes con hijos, establece que son estudiantes de estrato 2 en su mayoría que han tenido máximo 3 hijos, principalmente vivían fuera de Bogotá, provenientes de colegios públicos, con padres casados o en unión libre. Este grupo presenta algunas coincidencias con el grupo de estudiantes con dificultades económicas y que también presentaron un máximo de 3 hijos (Grupo 3). La diferencia radica en la percepción negativa de la universidad frente al personal administrativo, la planta física, los servicios universitarios como bienestar universitario entre otros.

El grupo de estudiantes mayores satisfechos (Grupo 2), de las localidades de Bosa, Suba y Engativá principalmente, se caracterizaron porque son estudiantes con mayor edad, de estrato 3 en su mayoría, que recibieron apoyo institucional, con padres casados, divorciados y/o solteros, con un promedio académico más alto con respecto a los demás grupos, no les agradó el currículo de la carrera y calificaron los aspectos institucionales con una percepción de buena y excelente.

En cuanto al grupo de estudiantes que presentaron dificultades económicas con hijos (grupo 3), se pueden establecer que son estudiantes de estrato 2 de las localidades de Engativá y Kennedy, con un máximo de 3 hijos, que también recibió apoyo institucional, con más personas a cargo y donde la percepción de la universidad en la mayoría de los casos es aceptable y buena. A diferencia de los demás grupos, este se caracteriza porque dentro de los motivos de deserción, manifestaron que desertaron por mala relación con los compañeros, número elevado de estudiantes, trabajar, tener presiones familiares y dificultades económicas.

Figura 8.

Representación gráfica de las variables numéricas en los tres grupos



Fuente: elaboración propia.

Los gráficos muestran el comportamiento de los grupos para las variables “Edad”, “Materias inscritas”, “Materias reprobadas”, “Promedio académico”, “Número de hijos” y “Número de personas a cargo”. Para el grupo 1 (pca_clusters = 0), se evidencia que tiene un intervalo de edad menor con respecto a los demás grupos, como también presenta promedios académicos por debajo de 2.0. En cuanto al grupo 2 (pca_clusters = 1), se tienen promedios un poco más altos por encima de 4.0 y con un máximo de dos hijos. El grupo 3 (pca_clusters = 2), presenta más personas a cargo y menos número de materias reprobadas.

En cuanto a la variable “razón de abandono, clasificada por grupo, se puede establecer que efectivamente los problemas económicos (65%) son la principal razón por la cual los estudiantes desertan de la educación superior, pero a diferencia de otros estudios, las dificultades académicas es la causa menos probable de deserción como se puede observar en el promedio general.

También se encontraron otras razones de deserción como lo fueron el inconformismo con las instalaciones de la universidad o que los docentes no cumplieron con las expectativas, entre otros.

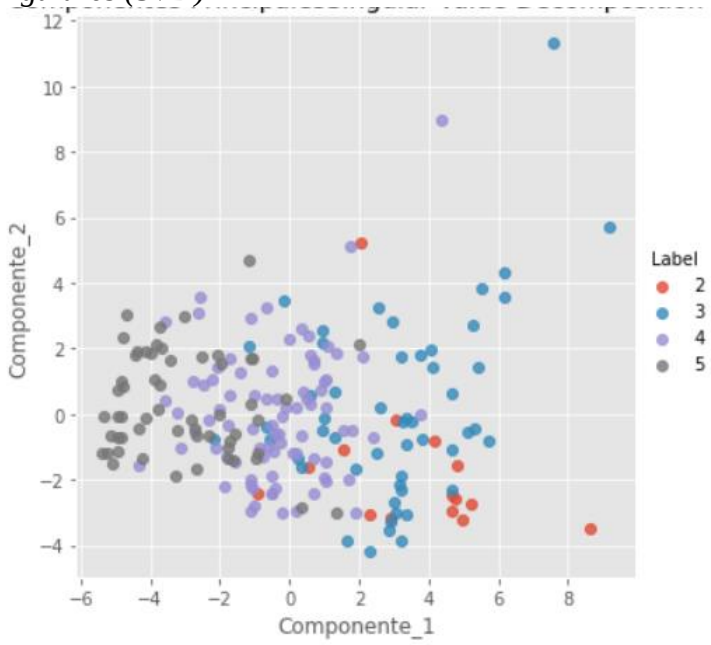
Tabla 3.
Frecuencias de la razón de abandono por grupo

Razón de Abandono					
Clusters	**1**	**2**	**3**	**Total**	**%**
Otra	19	23	6	48	23%
Dificultades académicas	2	4	1	7	3%
El programa no llenó expectativas	3	3	2	8	4%
Problemas económicos	36	72	27	135	65%
Problemas familiares	0	7	2	9	4%
Total	**60**	**109**	**38**	**207**	**100%**

Fuente: elaboración propia.

Al observar la proyección de cada una de las variables en las dos primeras componentes, de acuerdo con la distribución de los grupos dada por el algoritmo K-Means, la técnica que mejor las representó gráficamente fue la descomposición en valores singulares (SVD) basado en el análisis de componentes principales (ACP), permitiendo una visualización del comportamiento de los grupos.

Figura 9.
*Proyección de la **c**alificación del personal administrativo con la descomposición en valores singulares (SVD)*



Fuente: elaboración propia.

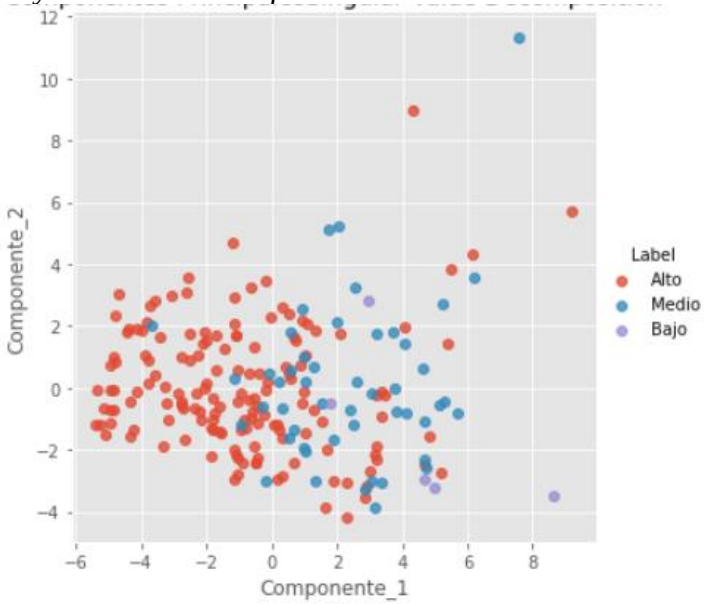
El gráfico de descomposición en valores singulares basado en el ACP, muestra la misma estructura de agrupación que se realizó con k-means como se observa en la figura 7. En este gráfico se puede analizar la proyección de la variable "Calificación del personal administrativo" de la universidad en las dos primeras componentes que explican el 20% de la variabilidad, donde se puede evidenciar el comportamiento de los diferentes grupos con respecto a la calificación de

la variable. El grupo de estudiantes inconformes con hijos (Grupo 1) registró una calificación de deficiente y aceptable (2 y 3), lo que corresponde a una percepción negativa para el personal administrativo.

El grupo de estudiantes mayores satisfechos (Grupo 2), presentó una calificación de 5 indicando una percepción buena y excelente y el grupo de estudiantes con dificultades económicas con hijos (Grupo 3) presentó una calificación de aceptable y buena (3 y 4).

Para la percepción que tenían los estudiantes acerca de qué tan satisfechos se sentían con la universidad, se clasificó la calificación en "deficiente" con un nivel bajo, "aceptable" nivel medio y la calificación de "bueno" y "excelente" con un nivel alto, encontrando que el grupo de estudiantes inconformes con hijos presentaron una calificación de bajo y algunos como medio. El grupo de estudiantes mayores satisfechos presentaron una calificación alta como se evidencia en la figura 10.

Figura 10.
Proyección de la satisfacción de la universidad con SVD



Fuente: elaboración propia.

Finalmente, se aplicó la prueba no paramétrica de Kruskal-Wallis que verifica la existencia o no existencia de una diferencia estadísticamente significativa entre las medianas de tres o más grupos independientes. Ver tabla 4.

Tabla 4.
Prueba estadística

Prueba	Variables	Estadístico	P_value
Independencia entre los grupos	Edad	5.977e-05	0.999
	Promedio	3.206	0.2012
	Materias Reprobadas	1.218	0.543

Fuente: elaboración propia.

Con una significancia del 5% y un $p > 0.05$ se puede afirmar que: no existe suficiente evidencia estadística para determinar que existe independencia entre las variables numéricas de los diferentes grupos. Aunque existen causas comunes de deserción como la dificultad económica, esta categorización por grupos permitió obtener adicionalmente las razones por las cuales los estudiantes desertan, donde la calificación que los estudiantes asignaron a las diferentes dependencias institucionales son un indicador para la organización de los grupos y determinar algunos aspectos instituciones a mejorar en la universidad.

Discusión

A partir de los resultados obtenidos se pueden establecer acercamientos o diferencias con otros estudios, resaltando que estos resultados pueden estar ligados a la calidad de los datos, las técnicas empleadas, la ubicación temporal, las características socio-demográficas y económicas de la población, entre otras.

En efecto, la principal causa de deserción en la IES fueron los problemas económicos (65%) y otros factores (35%) de menor impacto (Castillo y Sánchez, 2021), pero al igual que el estudio de educación inicial en Paraguay realizado por Chaves & Torres (2019) donde se aplicaron técnicas del aprendizaje no supervisado, se coincidió que el factor económico es una de las principales causas de deserción. En ese mismo contexto, existen otros estudios que difieren en cuanto a que la principal causa de deserción ha sido el factor académico. Es así como el rendimiento académico se corrobora como la variable determinante en la decisión de permanencia o abandono, asociado al número de créditos aprobados y reprobados por el estudiante, planteado por Casanova et al. (2018), donde el ser mujer tiene mayor relevancia al desertar, mientras que, en nuestro caso, la mayor deserción se presentó en los hombres (67.6%).

En cuanto a las técnicas del aprendizaje no supervisado para el análisis de datos mixtos, al igual que Giraldo (2021), la técnica de Análisis de Componentes Principales (ACP) y la descomposición en valores singulares (SVD), proporcionan una base fundamental para el procesamiento de la cuantificación optima de datos cualitativos encontrando mejores relaciones entre las variables.

En consecuencia, las técnicas empleadas para lograr la agrupación y caracterización de los estudiantes que desertaron, permitieron encontrar resultados consistentes con los encontrados con otros estudios, asimismo, como en el estudio de las distintas causas para abandonar los estudios universitarios planteado por Vries et al. (2011), donde se realiza una investigación cualitativa a un grupo de estudiantes desertores, se pueden encontrar diversas coincidencias frente a que la deserción se debe en gran medida a la incompatibilidad de estudios y trabajo, como se observó en los resultados de los programas de ingenierías y ciencias que suelen ser tradicionalmente más exigentes, mostrando un mayor índice de deserción Vries et al. (2011). Este estudio también plantea que es necesario reconocer que existe una diversidad de estudiantes que pueden tener una perspectiva distinta y que las razones de abandono están ligadas al contexto institucional y nacional, por lo que se debe pensar en la flexibilidad curricular y en la adecuación de la política de retención estudiantil.

Finalmente, las condiciones socio-económicas juegan un papel más importante que el clima organizacional o el acoplamiento entre estudiante y universidad Vries et al. (2011) y aunque en el presente estudio se encontró una asociación entre la razón de abandono con la percepción que tiene los estudiantes en el aspecto institucional, la principal razón de abandono (problemas económicos) pudo estar ligada a la crisis de salud pública ocasionada por la pandemia, aumentando de índice de deserción en esta institución de educación superior.

Conclusiones

Con el desarrollo del presente trabajo se identificaron de manera práctica y efectiva las principales causas de deserción en una muestra de 207 estudiantes que desertaron de una institución de educación superior durante el 2020, a partir de técnicas del aprendizaje no supervisado que permitieron encontrar patrones de asociación entre variables y grupos de estudiantes.

La aplicación de técnicas como el análisis de componentes principales (PCA), la descomposición en valores singulares (SVD) y el algoritmo de agrupación k-Means, permitieron un análisis sencillo, rápido y útil para reducir la dimensionalidad a 24 componentes con el 71% de la variabilidad explicada y la agrupación en tres grupos de estudiantes: (1) estudiantes inconformes con hijos, (2) estudiantes mayores satisfechos y (3) estudiantes con dificultades económicas con hijos. Esta agrupación identificó un patrón asociado a la percepción de algunos aspectos institucionales como se evidenció en la proyección de las variables a partir de las dos primeras componentes dadas por SVD, lo que facilitó encontrar la existencia de causas particulares de deserción.

Es así como, el primer grupo, estudiantes con hijos principalmente de estrato 2 que vivían fuera de Bogotá, manifestaron estar inconformes con aspectos institucionales como el personal administrativo, la planta física, los servicios universitarios, el bienestar universitario, entre otros. El segundo grupo, son estudiantes con mayor edad de estrato 3 en su mayoría, con un promedio académico más alto con respecto a los demás grupos que no les agradó el currículo de la carrera y manifestaron estar satisfechos con la universidad. El tercer grupo, son estudiantes de estrato 2, que tiene más personas a cargo, recibieron apoyo institucional y presentaron mayor dificultad económica, además, tienen una percepción de la universidad en la mayoría de los casos aceptable y buena. A diferencia de los demás grupos, este grupo se caracterizó porque desertaron por mala relación con los compañeros, por el número elevado de estudiantes, por problemas en el trabajo y presiones familiares.

En comparación con otros estudios relacionados con la deserción escolar, se encontró que existen algunas coincidencias frente a las dificultades económicas, problemas familiares y deficiencias académicas pasadas (Castillo y Sánchez, 2021). Adicionalmente, se pudo establecer que en esta institución el factor académico no fue una de las principales causas de deserción como ocurrió en otros estudios (Gutiérrez et al., 2021), tampoco lo fue el sector donde habitaba el estudiante. Cabe resaltar que, en el estudio se identificó que existen causas relacionadas con aspectos institucionales, la falta de información de programas y ayudas, la dificultad de trabajar y estudiar al mismo tiempo y el ambiente universitario.

La identificación de estas causas particulares permite propiciar estrategias eficaces que contribuyan al mejoramiento del plan de permanencia estudiantil y la política de retención con que cuenta la universidad. Cada uno de los elementos encontrados como causas de la deserción debe ser analizado con el fin de establecer estrategias de mejora en la institución o de facilitación de la continuidad para los estudiantes. De esta manera la investigación que hace uso de técnicas estadísticas brinda información precisa y confiable para establecer acciones de mejora que aporten a la disminución de la tasa de deserción en las IES, y al fomento del capital humano con calidad en el país.

Referencias bibliográficas

Bacos, C. A. Machine learning and education in the human age: a review of emerging technologies. In *Science and information conference* (pp. 536-543). Springer, Cham. (2019, April)

Bravo Núñez A. et al. Reducción de dimensiones: revisión y aplicaciones en clasificación automática. 2020.

Casanova, J. R., Cervero Fernández-Castañón, A., Núñez Pérez, J. C., Almeida, L. S., & Bernardo Gutiérrez, A. B. (2018). Factors that determine the persistence and dropout of university students. Psicothema, 30.

Castiblanco W. Modelo de ausentismo y deserción retención y permanencia estudiantil en la universidad ECCI. 2020.

Castillo, G. T. U., & Sánchez, B. A. M. (2021). Factores que inciden en la deserción universitaria. *TZHOECOEN*, *13*(2), 56-65.

Chaves, V. E. J., & Torres, M. G. (2019). Análisis de la Educación Inicial en Paraguay a través de las Técnicas de Aprendizaje Automático. Revista de la Sociedad Científica del Paraguay, 24(2), 293-304.

Giraldo Otálvaro, J. D. (2021). Estudio de las técnicas de reducción de dimensión basadas en componentes principales: Análisis de componentes principales no lineales.

Gutiérrez, D., Díaz, J. F. V., & López, J. (2021). Indicadores de deserción universitaria y factores asociados. *EducaT: Educación virtual, Innovación y Tecnologías*, *2*(1), 15-26.

Hernández Sampieri, Roberto. (2018). Metodología de la investigación: las rutas: cuantitativa y cualitativa y mixta. México: Mc Graw Hill- educación.

Jolliffe, IT (2002). Representación gráfica de datos utilizando componentes principales. *Análisis de componentes principales*, 78-110.

Kuvcak, Danijel y Jurieié. (2018). Aprendizaje máquina en educación: Una encuesta de las tendencias actuales de investigación. *Annals of DAAAM Proceedings*.

Martínez J., (2022). Clustering (Agrupamiento), K-Means con ejemplos en Python. IArtificial.net. https://www.iartificial.net/clustering-agrupamiento-kmeans-ejemplos-en-python/

Mysiak k., (2015). Explicación de la agrupación en clústeres de DBSCAN. Hacia la ciencia de datos. https://towardsdatascience.com/explaining-dbscan-clustering-18eaf5c83b31

Nainggolan, R. Perangin-angin, E. Simarmata, and A. F. Tarigan. (2019). Improved the performance of the k-means cluster using the sum of squared error (sse) optimized by using the elbow method. In Journal of Physics: Conference Series, volume 1361, page 012015. IOP Publishing.

Navarro Céspedes J. M. (2008). Análisis de Componentes Principales y Análisis de Regresión para datos categóricos. Aplicación en HTA. PhD Thesis, Universidad Central "Marta Abreu" de Las Villas.

Patiño Garzón L. and A. M. Cardona Pérez. (2012). Revisión de algunos estudios sobre la deserción estudiantil universitaria en Colombia y Latinoamérica. Theoría: Ciencia, Arte y Humanidades, 21(1):9 – 20. ISSN 0717196X. URL https://search-ebscohost-com.craiustadigital.usantotomas.edu.co/login.aspx?direct=true&db=a9h&AN=112611591&lang=es&site=ehost-live.

Sánchez L. G., G. A. Osorio, and J. F. Suárez. (2008). Introducción a kernel acp y otros métodos espectrales aplicados al aprendizaje no supervisado. Revista Colombiana de Estadística, 31(1):19–40.

Suárez Rodríguez J. M. (2015). Caracterización de los hurtos a personas que afectan la localidad los mártires de la ciudad de Bogotá mediante el uso de los algoritmos de agrupamiento de

minería de datos espaciales dbscan y k-means. Tesis de Ingeniería Catastral y Geodesia. Universidad Distrital Francisco José de Caldas.

Tinto V. (1975). Dropout from higher education: A theoretical synthesis of recent research. Review of Educational Research, 45(1):89–125. ISSN 00346543, 19351046. URL http://www.jstor.org/stable/1170024.

Toledo, J. A. J., & Pereira, S. R. T. (2015). Caracterización de la deserción estudiantil en educación superior con minería de datos. *Revista Tecnológica-ESPOL*, *28*(5).

Vélez White C. et al. (2008). Deserción estudiantil en la educación superior colombiana. Elementos para su diagnóstico y Tratamiento. Ministerio de Educación Nacional. Bogotá. Colombia, pages 7–34.

Vries, W. D., León Arenas, P., Romero Muñoz, J. F., & Hernández Saldaña, I. (2011). ¿Desertores o decepcionados? Distintas causas para abandonar los estudios universitarios. Revista de la educación superior, 40(160), 29-49.

Zúñiga, J. (2021). El algoritmo k-means aplicado a clasificación y procesamiento de imágenes. Disponible en https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html