UNIVERSIDAD DE EL SALVADOR

FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA DE POSGRADO

OPTIMIZACIÓN DE LA PREDICCIÓN DE RETENCIÓN ESTUDIANTIL EN LA EDUCACIÓN SUPERIOR: UN ANÁLISIS COMPARATIVO DE MODELOS DE APRENDIZAJE AUTOMÁTICO SUPERVISADOS Y NO SUPERVISADOS

PRESENTADO POR:

FRANKLIN IVAN ARGUETA BERMUDEZ

WILBER ALEXANDER ORTIZ CORTEZ

PARA OPTAR AL TITULO DE:

MAESTRO EN INGENIERIA PARA LA INDUSTRIA CON ESPECIALIZACION EN CIENCIAS DE LA COMPUTACION

CIUDAD UNIVERSITARIA, FEBRERO 2026

UNIVERSIDAD DE EL SALVADOR

RECTOR:

MSc. JUAN ROSA QUINTANILLA

SECRETARIO GENERAL:

LCDO. PEDRO ROSALÍO ESCOBAR CASTANEDA

FACULTAD DE INGENIERÍA Y ARQUITECTURA

DECANO:

MSc. LUIS SALVADOR BARRERA MANCÍA

SECRETARIO:

ARQ. RAÚL ALEXANDER FABIÁN ORELLANA

ESCUELA DE POSGRADO

DIRECTOR:

MSc. ELMER ARTURO CARBALLO RUÍZ

UNIVERSIDAD DE EL SALVADOR

FACULTAD DE INGENIERIA Y ARQUITECTURA

ESCUELA DE POSGRADO

Trabajo de Graduación previo a la opción al Grado de:

MAESTRO EN INGENIERIA PARA LA INDUSTRIA CON ESPECIALIZACION EN CIENCIAS DE LA COMPUTACION

Título:

OPTIMIZACIÓN DE LA PREDICCIÓN DE RETENCIÓN ESTUDIANTIL EN LA EDUCACIÓN SUPERIOR: UN ANÁLISIS COMPARATIVO DE MODELOS DE APRENDIZAJE AUTOMÁTICO SUPERVISADOS Y NO SUPERVISADOS

Presentado por:

FRANKLIN IVAN ARGUETA BERMUDEZ

WILBER ALEXANDER ORTIZ CORTEZ

Trabajo de Graduación Aprobado por:

Docente Asesor:

PhD. MIGUEL GARCIA TORRES

SAN SALVADOR, FEBRERO 2026

Trabajo de Graduación Aprobado por:

Docente Asesor:

Fdo digitalmente por
Miguel Garcia Torres
(18/Febrero/2026)

PhD. MIGUEL GARCIA TORRES

AGRADECIMIENTOS

Agradezco a Dios por brindarme la fortaleza, la sabiduría y la perseverancia necesaria para culminar esta etapa académica. También, quiero agradecerle a mi familia, especialmente a mi madre, por su amor incondicional, apoyo constante y confianza en cada paso de mi formación. A mi esposa, gracias por tu paciencia, comprensión y por ser mi compañera incondicional en este proceso.

Franklin Iván Argueta Bermúdez

Quiero expresar mi más profundo agradecimiento a todas las personas e instituciones que, de una u otra manera, hicieron posible la realización de esta tesis. En primer lugar, agradezco a Dios por brindarme salud, sabiduría y fortaleza durante este proceso académico. A mi asesor de tesis, Dr, Miguel Garcia Torres, por su guía constante, por compartir sus conocimientos y por sus valiosas observaciones que enriquecieron este trabajo. Su compromiso y paciencia fueron fundamentales para alcanzar este logro. A mi familia, por su amor incondicional, comprensión y apoyo moral durante todo el desarrollo de esta etapa. Especialmente a mi Madre, quien creyó en mí incluso en los momentos más difíciles. Finalmente, a mi compañero y amigo de maestría Franklin Argueta, por los espacios de colaboración, aprendizaje compartido y amistad, que hicieron este camino más enriquecedor y llevadero.

Wilber Alexander Ortiz Cortez

Contenido

Lista de Figuras

RESUMEN

La deserción estudiantil en la educación superior representa un desafío estructural de gran relevancia para el sistema educativo salvadoreño, especialmente en las universidades privadas, donde se observa un incremento sostenido en las tasas de abandono. Esta problemática no solo tiene implicaciones individuales —como el rezago académico y económico del estudiante— sino también repercusiones institucionales y sociales, afectando la eficiencia del sistema educativo, la sostenibilidad financiera de las instituciones y el desarrollo del capital humano en el país. En este contexto, la presente investigación se orientó a optimizar la predicción de la retención estudiantil mediante un análisis comparativo de modelos de aprendizaje automático supervisados y no supervisados, utilizando un enfoque integral que incorpora métodos descriptivos, exploratorios y predictivos.

El estudio parte de la recopilación, depuración y estructuración de un conjunto de datos históricos provenientes de una universidad privada de El Salvador, correspondientes al periodo 2019–2024. Este conjunto incluyó más de 83,000 registros y 31 variables demográficas, socioeconómicas, académicas e institucionales. A partir de esta base de datos, se desarrolló una metodología que permitió avanzar en tres niveles de análisis: descriptivo, exploratorio y predictivo, con el objetivo de identificar patrones relevantes, segmentar perfiles de estudiantes y construir modelos capaces de predecir con alta precisión la continuidad o abandono de los estudiantes.

En la primera fase, el análisis descriptivo permitió identificar factores clave que inciden en la retención estudiantil. Se evidenció que el género femenino presenta una mayor continuidad académica en comparación con el masculino, y que los grupos etarios entre 18 y 34 años constituyen el núcleo principal de la matrícula

universitaria. Geográficamente, los departamentos de San Salvador y La Libertad concentran el mayor número de estudiantes, aunque muestran diferencias en las tasas de deserción. En términos socioeconómicos, se observó que los estudiantes con financiamiento familiar presentan mejores tasas de retención, mientras que aquellos que dependen de recursos propios tienden a abandonar con mayor frecuencia. A nivel académico, la modalidad de estudio también desempeña un papel importante: la modalidad presencial ha disminuido desde 2019, mientras que la modalidad no presencial ha crecido, aunque esta última presenta menores tasas de deserción. Asimismo, el rendimiento académico medido por el CUM y las asignaturas aprobadas o reprobadas resultó ser un predictor significativo de la continuidad. Los estudiantes con bajo rendimiento y en niveles académicos iniciales fueron los más propensos a desertar.

En la segunda fase, se llevó a cabo un análisis exploratorio no supervisado mediante técnicas de reducción de dimensionalidad (PCA) y algoritmos de agrupamiento (K-means). El objetivo fue segmentar a la población estudiantil en clústeres homogéneos sin utilizar la variable objetivo de continuidad. El análisis reveló que las variables con mayor peso en la formación de los grupos fueron el rendimiento académico, la edad, el tipo de financiamiento, el estado civil y la modalidad de estudio. Inicialmente, se exploró una agrupación binaria (k=2), que permitió distinguir entre estudiantes con perfiles de mayor y menor riesgo de abandono. Posteriormente, el uso del método del codo determinó que k=5 era el valor óptimo para una segmentación más detallada. Esta clusterización reveló perfiles específicos asociados a ciertas facultades y modalidades. Por ejemplo, el 71.8 % de los estudiantes de la Facultad de Arte se concentraron en un único clúster, lo que sugiere comportamientos institucionales homogéneos. Estas agrupaciones permitieron comprender mejor las dinámicas internas del estudiantado y fundamentar intervenciones focalizadas.

La tercera fase implicó el desarrollo y evaluación de modelos predictivos supervisados con el propósito de anticipar la continuidad o abandono de los estudiantes. Se compararon diversos algoritmos de clasificación, incluyendo Máquinas de Soporte Vectorial (SVM), Bosques Aleatorios (Random Forest), K-Vecinos Más Cercanos (KNN), XGBoost y AdaBoost. La preparación del conjunto de datos incluyó la codificación de variables categóricas, normalización y partición estratificada en subconjuntos de entrenamiento (75 %) y prueba (25 %). El modelo Random Forest, con 50 estimadores y un valor de "*min_samples_split*" de 32, fue el que obtuvo el mejor rendimiento general, alcanzando una precisión global de 78.99 % y una precisión positiva del 93.55 %, lo que lo convierte en una herramienta especialmente eficaz para detectar estudiantes que probablemente continúen sus estudios. Por su parte, el modelo SVM con kernel RBF y parámetro C = 46 alcanzó una precisión global de 77.31 %, con un mejor equilibrio entre las tasas de clasificación de ambas clases.

Los resultados muestran que el rendimiento académico, el nivel de avance en la carrera, el tipo de financiamiento, la modalidad de estudio, el estado civil y la edad son variables críticas para la predicción de la retención. Asimismo, la comparación de modelos evidencia que no existe un único modelo ideal, sino que la elección depende de las prioridades institucionales: mientras algunos modelos maximizan la detección de estudiantes en riesgo de deserción, otros priorizan la correcta identificación de quienes permanecerán. Esta diferenciación es esencial para la formulación de estrategias institucionales efectivas.

Finalmente, la investigación demuestra que la integración de enfoques descriptivos, exploratorios y predictivos permite alcanzar una comprensión profunda y completa del fenómeno de la retención estudiantil. Al aplicar técnicas de aprendizaje

automático, se generan herramientas concretas que pueden ser utilizadas para fortalecer la toma de decisiones institucionales, diseñar políticas de intervención temprana, personalizar el acompañamiento estudiantil y, en última instancia, mejorar la eficiencia y equidad del sistema de educación superior. Esta tesis constituye un aporte relevante tanto para la literatura académica como para la gestión educativa, y sienta las bases para futuras investigaciones que busquen ampliar, adaptar o perfeccionar los modelos presentados.

Capítulo 1: El problema de Investigación

1.1. Introducción

La Educación es el pilar fundamental para que una sociedad sea crítica, pensante de su realidad, en este sentido, se debe dar prioridad a que la sociedad tenga acceso a la educación y básicamente a los jóvenes quienes serán responsables de transformar la sociedad. Sin embargo, la deserción estudiantil es un desafío clave para las instituciones de educación superior de El Salvador debido a su impacto negativo tanto en los estudiantes como en la sociedad. Las instituciones de educación superior no han tomado medidas para mejorar la retención estudiantil y es casi nula las investigaciones sobre este tema y el aprovechamiento de las tecnologías digitales. Sin embargo, este tema sigue siendo sumamente importante ya que ha sido difícil lograr mejoras sustanciales después del fenómeno del COVID-19 que impacto a nivel mundial.

En los últimos años, la disponibilidad de datos digitales de los estudiantes ha abierto nuevas oportunidades para identificar a los estudiantes en riesgo y analizar los factores que conducen a su deserción. Los expedientes académicos institucionales y los datos demográficos de los estudiantes se han incorporado a algoritmos como pueden ser la inteligencia artificial o el machine learning para predecir los estudiantes en riesgo de los programas de pregrado, para este trabajo una Universidad Privada del país proporciono una base de datos del periodo 2019-2024, para analizar qué características son las más importantes para la retención de los estudiantes y cuáles son sus posibles causas de deserción o porque no culminan una carrera universitaria.

Se analiza la importancia de los factores en la predicción del abandono del programa de pregrado de las facultades de la Universidad, en comparación con los datos demográficos y de transcripción en la educación superior. Por lo tanto, el presente estudio proporciona una comprensión profunda del poder predictivo y el papel de las características de datos más importantes de diferentes técnicas de aprendizaje automático para la predicción del abandono del programa de pregrado en la Universidad. En el presente estudio, también nos propusimos ofrecer contribuciones prácticas ayudando a los representantes de las instituciones de educación superior a comprender cómo y cuándo aplicar el aprendizaje automático en la predicción del abandono estudiantil y apoyándolos para tomar decisiones sobre estrategias de prevención para los factores importantes que contribuyen a dichas predicciones en función del tiempo.

Este fenómeno en nuestro país es reconocido por todas las instituciones educativas, hasta por el ente rector de la educación, pero aún no hay implementación de acciones concretas en las instituciones antes mencionadas, por lo que el trabajo se hizo con la finalidad de orientar las políticas de las instituciones de educación superior del país y así proveer antecedentes relevantes para la implementación de programas exitosos.

1.2. Objetivos

1.2.1. Objetivo general

- Evaluar modelos de aprendizaje automático supervisados y no supervisados para optimizar la predicción de la retención estudiantil en una Universidad Privada de El Salador, utilizando para ello conjuntos de datos históricos de estudiantes.

1.2.2. Objetivo Especifico

- Compilar un conjunto de datos exhaustivo de estudiantes universitarios, abarcando variables académicas, demográficas, financiera y de comportamiento, y preparar estos datos para su análisis mediante procesos de limpieza, transformación y normalización.

- Ejecutar análisis descriptivos en el conjunto de datos, explorando detalladamente variables académicas, demográficas, socioeconómicas y de comportamiento para identificar patrones, tendencias y relaciones entre las variables clave que pueden influir en la retención o deserción estudiantil.

- Desarrollar modelos predictivos utilizando técnicas de aprendizaje automático tanto supervisadas como no supervisadas, y aplicar técnicas de optimización para mejorar su rendimiento predictivo.

- Analizar de manera comparativa los modelos desarrollados utilizando métricas de rendimiento estándar para identificar los enfoques más efectivos en la predicción de retención estudiantil.

- Realizar métodos y factores ágiles y efectivos para prevenir la deserción estudiantil y mejorar las tasas de retención en la Universidad Privada de El Salador.

1.3. Justificación

En El Salvador la formación universitaria se convierte en una oportunidad vital para el proceso de construcción del ciudadano, en los últimos años la deserción estudiantil ha aumentado a nivel universitario en todo el territorio nacional, de manera particular en las universidades privadas; es por ello, que se vuelve necesario identificar factores que ayuden a construir estrategias para la retención estudiantil.

En este sentido la investigación busca desarrollar un análisis de los factores que están relacionados con la deserción y así poder afrontar desde una perspectiva estratégica la retención del estudiante, esto en función de las tareas esenciales de la universidad que consisten en el ingreso, la permanencia, el egreso y la graduación de quienes aspiran a la vida universitaria y profesional.

En el ámbito de la educación, a nivel mundial el uso de Machine Learning se está posicionando como un recurso valioso para adaptar y mejorar los métodos de enseñanza-aprendizaje. Las investigaciones destacan su capacidad para adaptar contenidos basándose en las necesidades específicas de cada estudiante, proporcionar análisis detallados a partir de grandes volúmenes de datos educativos, e innovar en la creación de recursos didácticos personalizados, tales como ejercicios prácticos y material audiovisual. El Salvador siendo un país en desarrollo no debe de quedar atrás en las implementaciones de estas nuevas tecnologías ya que estos avances sugieren unos grandes panoramas prometedor para las estrategias pedagógicas

1.4. Cronograma (Desarrollo de Tesis y Articulo Científico)

ACTIVIDADES	2024										2025							Recursos
	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	
Acercamiento con la Empresa y recolección de datos																			Computadora, acceso a internet
Reuniones semanales con el asesor																			Computadora, acceso a internet
Revisión de literatura																			Computadora, acceso a internet, libros, tesis, artículos científicos
Elaboración de perfil de tema de tesis																			Computadora, acceso a internet, libros, tesis, artículos científicos
Entrega de perfil de tema de tesis para "Aprobación"																			Computadora, acceso a internet
Gestión de Oficialización de tema de tesis y asesor																			Computadora, acceso a internet
Desarrollo del Anteproyecto																			Computadora, acceso a internet, libros, tesis, artículos científicos
Presentación de Anteproyecto al Asesor																			Computadora, acceso a internet
Desarrollo de Tesis																			Computadora, acceso a internet, libros, tesis, artículos científicos
Transformación de Datos																			Computadora, software estadístico
Elaboración Capitulo 1																			Computadora, acceso a internet, libros, tesis, artículos científicos

ACTIVIDADES	2024												2025								Recursos
	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto			
Mejoras al Capítulo 1 siguiendo las recomendaciones del Asesor																					Computadora, acceso a internet, libros, tesis, artículos científicos
Elaboración Capitulo 2																					Computadora, acceso a internet, software estadístico
Mejoras al Capítulo 2 siguiendo las recomendaciones del Asesor																					Computadora, acceso a internet, software estadístico
Elaboración Capitulo 3																					Computadora, acceso a internet, software estadístico
Mejoras al Capítulo 3 siguiendo las recomendaciones del Asesor																					Computadora, acceso a internet, software estadístico
Elaboración Capitulo 4																					Computadora, acceso a internet, software estadístico
Mejoras al Capítulo 4 siguiendo las recomendaciones del Asesor																					Computadora, acceso a internet, software estadístico
Elaboración Capitulo 5																					Computadora, acceso a internet, libros, tesis, artículos científicos
Mejoras al Capítulo 5 siguiendo las recomendaciones del Asesor																					Computadora, acceso a internet, libros, tesis, artículos científicos
Elaboración Conclusiones Generales																					Computadora, acceso a internet, libros, tesis, artículos científicos
Mejoras a las Conclusiones Generales siguiendo las recomendaciones del Asesor																					Computadora, acceso a internet, libros, tesis, artículos científicos

ACTIVIDADES	2024												2025								Recursos
	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto			
Elaboración de apartados: Portada, Recomendaciones, Agradecimientos, Anexos																					Computadora, acceso a internet, libros, tesis, artículos científicos
Mejoras de la documentación																					Computadora, acceso a internet, libros, tesis, artículos científicos
Elaboración de documentación final a presentar para la Empresa y Asesor																					Computadora, acceso a internet, libros, tesis, artículos científicos
Mejoras al documento siguiendo las recomendaciones del Asesor																					Computadora, acceso a internet, libros, tesis, artículos científicos
Elaboración de Articulo Científico																					Computadora, acceso a internet, libros, tesis, artículos científicos
Mejoras al Articulo siguiendo las recomendaciones del Asesor																					Computadora, acceso a internet, libros, tesis, artículos científicos
Elaboración de la Presentación para la Defensa de Tesis																					Computadora, acceso a internet, libros, tesis, artículos científicos
Mejoras a la Presentación siguiendo las recomendaciones del Asesor																					Computadora, acceso a internet, libros, tesis, artículos científicos
Defensa de Tesis																					Computadora, acceso a internet, libros, tesis, artículos científicos

1.5. Antecedentes

La retención estudiantil en la educación superior es un tema de mucha importancia tanto para las instituciones educativas como para los investigadores del ámbito académico. La capacidad de una universidad para mantener a sus estudiantes matriculados hasta la finalización de sus programas no solo es un indicador clave de éxito institucional, sino que también tiene implicaciones económicas y sociales de gran alcance (Tinto, Through the eyes of students, 2017). La retención de estudiantes afecta directamente el valor de un estudiante como cliente, ya que el abandono temprano de los programas puede reducir el retorno de la inversión (ROI) tanto para los estudiantes como para la universidad (Hagedorn, 2021). La comprensión de los factores que influyen en la deserción es esencial para la creación de estrategias que optimicen tanto la experiencia académica como los recursos institucionales.

El presente estudio busca comparar el rendimiento de modelos de aprendizaje automático supervisados y no supervisados para optimizar la predicción de la retención estudiantil en una universidad privada de El Salvador. Esto permitirá no solo identificar al perfil del estudiante más propenso a desertar, sino también determinar en qué facultades y carreras se presentan mayor deserción, y cuáles son los momentos críticos en los que los estudiantes tienden a abandonar. Así, este análisis contribuirá a la elaboración de estrategias preventivas para mejorar las tasas de retención y la personalización del aprendizaje, con el uso de herramientas de inteligencia artificial que permiten intervenciones más personalizadas y efectivas.

El valor del estudiante puede ser conceptualizado en términos económicos y de capital humano, similar al valor de un cliente en una empresa (Kotler & Armstrong, 2018). La probabilidad de abandono se ve influenciada por factores como el rendimiento académico, la integración social, las variables demográficas y el apoyo

financiero (Simões & Soares, 2019). Al igual que en el ámbito empresarial, donde se busca maximizar el valor de un cliente a lo largo de su ciclo de vida (Gupta y otros, 2006), las universidades deben medir y optimizar el valor del estudiante para asegurar la sostenibilidad y el crecimiento institucional. Para ello, es fundamental el uso de modelos predictivos que permitan identificar a los estudiantes con mayor riesgo de abandono, lo que a su vez facilita la intervención temprana (Herzog, 2006).

El uso de datos históricos en la predicción de la retención estudiantil ha ganado popularidad debido a su capacidad para revelar patrones y tendencias que pueden ser difíciles de detectar a simple vista (Cabrera y otros, The role of finances in the persistence process: A structural model, 2018). A través del análisis de grandes volúmenes de datos, como el rendimiento académico, la asistencia, el apoyo financiero y las actividades extracurriculares, las instituciones educativas pueden tomar decisiones más informadas y personalizadas. Esta evolución ha estado marcada por el creciente interés en las técnicas de aprendizaje automático (machine learning), las cuales han demostrado ser herramientas eficaces para la predicción de deserción estudiantil (Romero y otros, 2013).

Dentro del aprendizaje automático, se distinguen dos grandes enfoques: el aprendizaje supervisado y el no supervisado. En el aprendizaje supervisado, los modelos se entrenan con datos etiquetados, lo que permite clasificar a los estudiantes según su riesgo de abandono (Cortez & Silva, 2008). Los modelos no supervisados, como los algoritmos de clustering, permiten identificar grupos de estudiantes con características similares, lo que facilita la segmentación de grupos de interés. Ambos enfoques tienen aplicaciones significativas en la predicción de la retención estudiantil, y el presente estudio busca comparar su efectividad en un contexto específico (una universidad privada en El Salvador).

Otra investigación relevante es la de (Asif y otros, 2017), quienes evaluaron diversos algoritmos de clasificación, incluyendo Support Vector Machines (SVM) y Random Forest, para predecir la deserción estudiantil en una universidad del Reino Unido. Sus hallazgos indicaron que los factores socioeconómicos y el rendimiento académico inicial eran predictores clave de la retención. Adicionalmente, un estudio realizado por (Abdullah y otros, 2019) en universidades de Malasia exploró la efectividad de algoritmos no supervisados como K-means para identificar patrones de abandono en los primeros años de estudio.

Finalmente, la importancia de las variables demográficas, como la edad, el género, y el origen socioeconómico, en la predicción de la retención ha sido bien documentada (Thomas, 2018). Estas variables juegan un papel crucial en la adaptación de los modelos predictivos al contexto regional, lo cual es particularmente relevante para el caso de El Salvador, donde las brechas económicas y sociales son significativas en la educación superior (Meneses y otros, 2020).

Capítulo 2: Marco Teórico

El concepto de valor del cliente en el ámbito comercial se refiere a la cantidad de ingresos que un cliente genera para una empresa durante el período en el que interactúa con ella. Esta idea es clave para la planificación estratégica, ya que permite a las organizaciones concentrar sus esfuerzos en retener a los clientes más valiosos (Kotler & Keller, 2016). Por su parte, el valor del estudiante en el contexto educativo se relaciona con el impacto que tiene la permanencia de un estudiante en la institución sobre sus ingresos y su reputación (Fahd y otros, 2022). En ambos casos, la retención es un factor crítico, ya que influye directamente en la estabilidad financiera y el crecimiento de la organización, sea esta una empresa o una institución educativa.

Una de las principales diferencias entre ambos contextos es la naturaleza de los factores que influyen en la retención. En el ámbito comercial, la retención de clientes depende principalmente de la calidad del producto o servicio, la experiencia del cliente y las políticas de fidelización (Kotler & Armstrong, 2018). En cambio, la retención estudiantil en la educación superior está influenciada por una gama más amplia de factores, incluidos los académicos, socioeconómicos, demográficos y psicosociales (Tinto, Completing college: Rethinking institutional action, 2012).

El aprendizaje automático (machine learning) ha emergido como una herramienta valiosa para predecir la retención estudiantil mediante la identificación de patrones ocultos en grandes conjuntos de datos. Los algoritmos supervisados, como la regresión logística, los árboles de decisión y las máquinas de soporte vectorial (SVM), requieren de datos etiquetados para entrenar el modelo. Estos algoritmos se utilizan comúnmente para predecir eventos binarios como el abandono o la

permanencia de los estudiantes (Vandamme y otros, 2017). En contraste, los algoritmos no supervisados, como el clustering o las redes neuronales auto-organizadas, no requieren datos etiquetados y son útiles para identificar subconjuntos de estudiantes con características similares, permitiendo una intervención personalizada (Breiman, 2001).

La regresión logística es ampliamente utilizada en la predicción de la retención, ya que permite evaluar la probabilidad de abandono en función de diversas variables independientes (Cortez & Silva, 2008). Por su parte, los modelos de clustering, como el K-means, se utilizan para agrupar a los estudiantes en función de características comunes, lo que facilita la detección de patrones de comportamiento que no son evidentes a simple vista (Kurniawan & Taufiq, 2019).

Diversos estudios han identificado múltiples factores que inciden en la retención estudiantil. Estos factores pueden dividirse en demográficos, académicos y psicosociales. Los factores demográficos incluyen edad, género, nivel socioeconómico y situación laboral (Tinto, Completing college: Rethinking institutional action, 2012). Los factores académicos, por otro lado, se relacionan con el rendimiento académico previo, el nivel de satisfacción con el programa de estudios y la percepción de la carga de trabajo. Los factores psicosociales abarcan aspectos como la motivación personal, la integración social en la universidad y el apoyo familiar (Astin, 1999).

La probabilidad de abandono aumenta considerablemente cuando los estudiantes enfrentan múltiples factores de riesgo, como bajos ingresos familiares y dificultades académicas (Cabrera y otros, College persistence: Structural equations modeling test of an integrated model of student retention, 2006). La detección temprana de estos factores mediante modelos predictivos permite la implementación de estrategias de

intervención, como tutorías personalizadas o programas de apoyo financiero (Martínez y otros, 2021).

La evaluación de los modelos de predicción de retención estudiantil se basa en varias métricas de desempeño. Entre las más utilizadas se encuentran la curva ROC (Receiver Operating Characteristic) y el AUC (Área Bajo la Curva), que permiten medir la capacidad del modelo para distinguir entre estudiantes que abandonarán y los que no lo harán (Fawcett, 2006). Además, métricas como la precisión, la sensibilidad y la especificidad se emplean para determinar la efectividad del modelo en la clasificación correcta de los casos (Chicco & Jurman, 2020).

La precisión (accuracy) indica el porcentaje de predicciones correctas en relación con el total de predicciones realizadas, mientras que la sensibilidad (recall) mide la capacidad del modelo para identificar correctamente a los estudiantes que están en riesgo de abandonar (Powers, 2011). Estas métricas permiten una evaluación integral del desempeño del modelo y facilitan la comparación entre diferentes algoritmos.

Diversos estudios han implementado modelos de aprendizaje automático para abordar el problema de la retención estudiantil. (Vandamme y otros, 2017) compararon modelos supervisados como la regresión logística y el árbol de decisión, encontrando que este último ofrecía una mejor precisión en la predicción de la deserción. Por otro lado, (Kurniawan & Taufiq, 2019) utilizaron el algoritmo de clustering K-means para agrupar estudiantes en función de variables demográficas y académicas, lo que permitió identificar grupos con alto riesgo de deserción.

En un estudio reciente, (Arqawi y otros, 2022) se centran en predecir la retención de estudiantes universitarios utilizando aprendizaje automático y aprendizaje

profundo, demostrando una alta precisión de predicción con modelos optimizados. Este enfoque híbrido resultó ser efectivo para mejorar la precisión y reducir los falsos positivos en la predicción.

Capítulo 3: Análisis Descriptivo

3.1. Introducción

El análisis descriptivo permite comprender las características fundamentales de los datos antes de aplicar técnicas de modelado, en este capítulo se busca ofrecer una visión clara de las tendencias, patrones y distribuciones de las variables clave que pueden influir en la continuidad de los estudiantes e identificar relaciones preliminares entre las variables.

Al examinar variables como la edad, el género, el tipo de financiamiento, la modalidad de estudio y los promedios académicos, es posible identificar grupos de estudiantes que presentan mayor riesgo de abandono. Esta comprensión inicial facilita la construcción de modelos más precisos, ya que se puede seleccionar un conjunto de variables relevantes y excluir aquellas que no aportan valor al modelo.

Objetivo del análisis descriptivo: explorar patrones que puedan influir en la retención estudiantil.

3.2. Explicación de la estructura de la base de datos

La base de datos utilizada para este estudio está compuesta por un conjunto de datos históricos de estudiantes de una Universidad Privada. Contiene un total de **83,175 observaciones** (filas) y **31 variables** (columnas) que abarcan aspectos demográficos, socioeconómicos, académicos y de progreso estudiantil. Las variables presentes en la base de datos son tanto categóricas como numéricas, permitiendo un análisis multidimensional de los factores que pueden influir en la retención estudiantil.

A continuación, se detallan las principales variables de la base de datos:

- **ID_Alumno:** Identificación única del estudiante.
- **Género:** Clasificación por sexo (M/F).

- **Edad:** Edad del estudiante al momento de ingresar a la universidad.
- **TipoFinanciamiento:** Fuente de financiamiento del estudiante, como "Ayuda Familiar" o "Fondos Propios".
- **EstadoFamiliar:** Estado civil del estudiante, como "Soltero", "Casado", etc.
- **Facultad y Carrera:** Facultad y carrera a la que pertenece el estudiante.
- **Modalidad:** Modalidad de estudio, que puede ser presencial o no presencial.
- **AIngreso y CIngreso:** Año y ciclo de ingreso del estudiante.
- **TipoIngreso:** Tipo de ingreso del estudiante, como "Nuevo", "Antiguo", entre otros.
- **Departamento y Municipio:** Ubicación geográfica de origen del estudiante.
- **Cum:** Promedio académico acumulado (CUM) del estudiante.
- **Nivel, Aprobadas, Reprobadas:** Información sobre el nivel académico del estudiante y el número de materias aprobadas o reprobadas.
- **Continuidad:** Estado de continuidad del estudiante, ya sea que haya finalizado o esté aún en el sistema educativo.
- **Cohorte y Periodo:** Año y semestre de la cohorte de ingreso del estudiante.

En términos de **temporalidad**, la base de datos abarca seis años académicos (2019-2024) y dos ciclos (semestres) por año, lo que permite observar tendencias a lo largo del tiempo, como la matrícula, la deserción o la continuidad de los estudiantes en diferentes periodos.

3.3. Análisis Demográfico

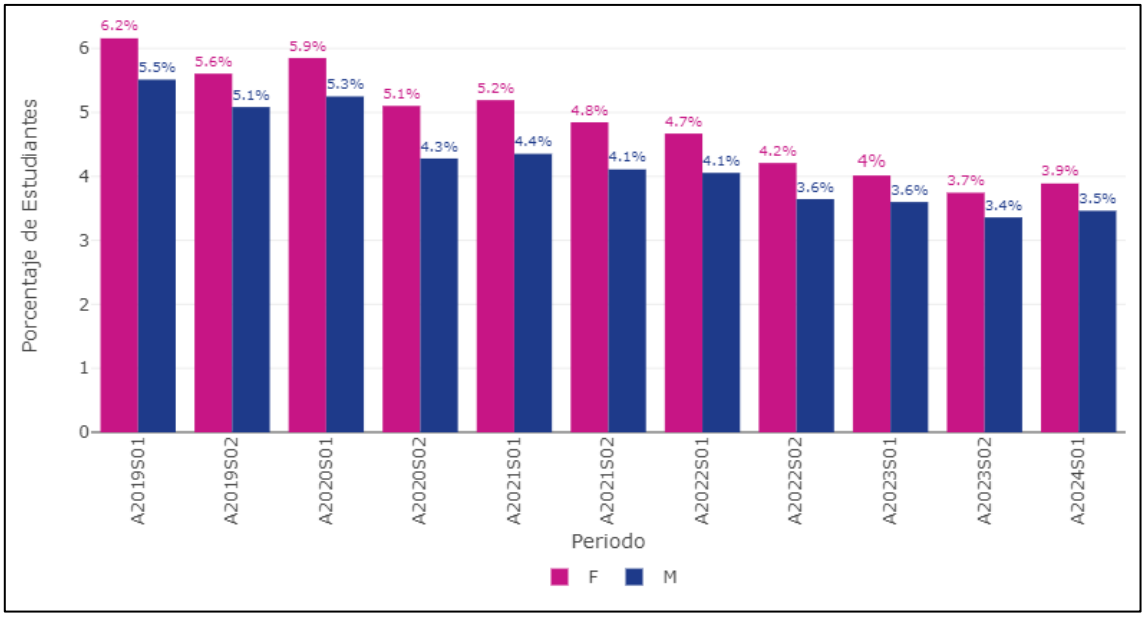
El análisis demográfico busca entender cómo las características propias de los estudiantes, como el género y la edad, están relacionadas con la retención estudiantil. Estas variables pueden influir significativamente en la continuidad académica de los estudiantes, proporcionando información clave para desarrollar estrategias de intervención y apoyo.

A través de este análisis, se busca identificar posibles diferencias en los patrones de continuidad académica entre grupos demográficos desglosando por el nivel de continuidad (Finalizado, Continuando, No Continuando).

3.3.1. Distribución de Estudiantes por Género

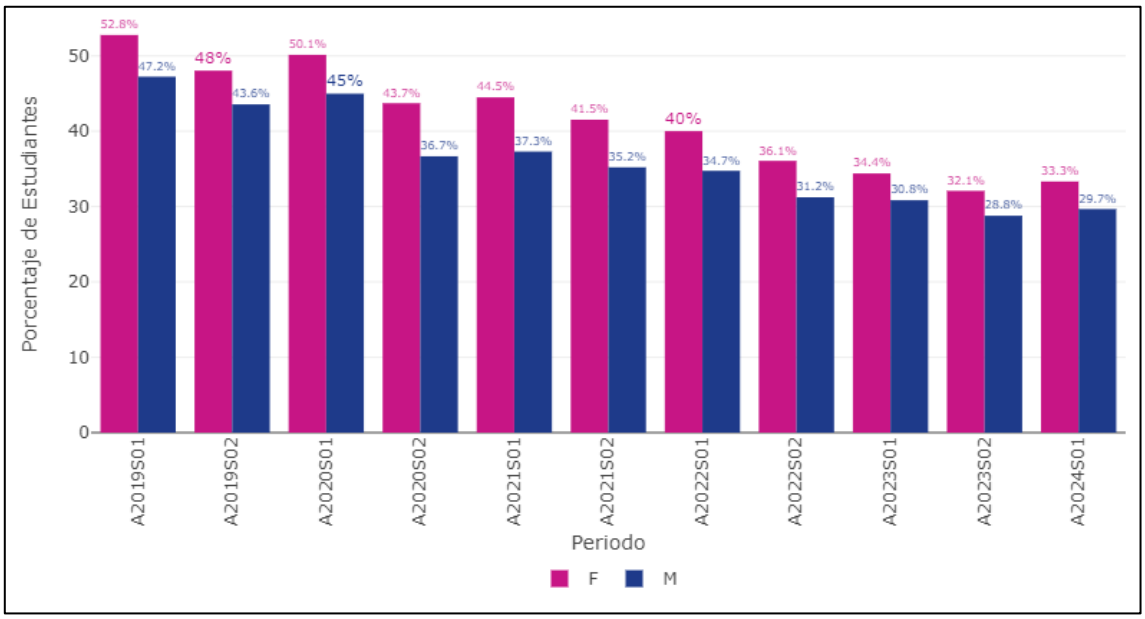
En la Figura 1, se puede observar que la mayoría de los estudiantes en cada semestre son de género femenino. También se puede observar una tendencia negativa en cuanto al total de estudiantes por periodo.

Figura 1: Distribución porcentual de estudiantes por género y periodo



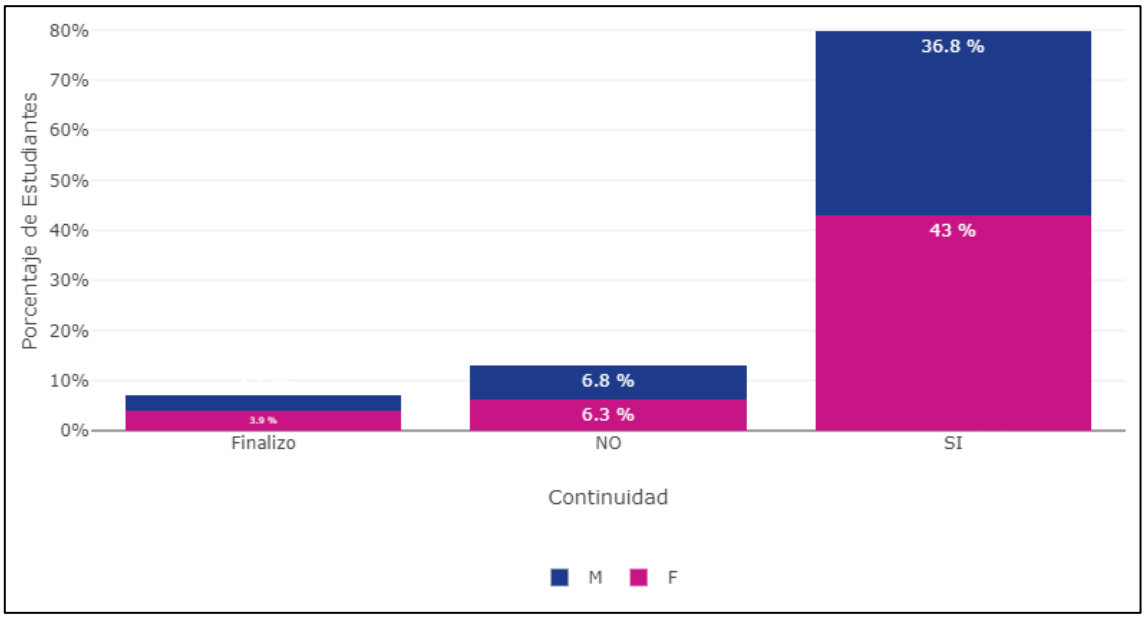
Ahora tomando como base el total de estudiantes del semestre uno del año 2019 tenemos que en este ciclo el 52.8% eran estudiantes del género femenino y el 47.2% eran del género masculino. Además, en la Figura 2, se puede observar que comparando el total de alumnos se ve una tendencia negativa la cual ha generado que en el semestre uno del año 2024 se tiene un 63% de estudiantes en cuanto a cantidad con respecto al semestre uno del 2019.

Figura 2: Distribución porcentual de estudiantes por género y periodo relativo al semestre 01 del año 2019



Finalmente, analizando la proporción de estudiantes por género y continuidad nos damos cuenta de que del total de estudiantes que han estudiado en la universidad desde el 2019 semestre 01 hasta el 2024 semestre 01 se tiene que han finalizado su carrera un 7.04%, no han continuado en la carrera de un ciclo al siguiente un 13.1% y si han continuado en la carrera de un ciclo a otro un 79.86%. Además, se observa en la Figura 3, que el género femenino es el que es más propenso a continuar en la carrera de un ciclo a otro.

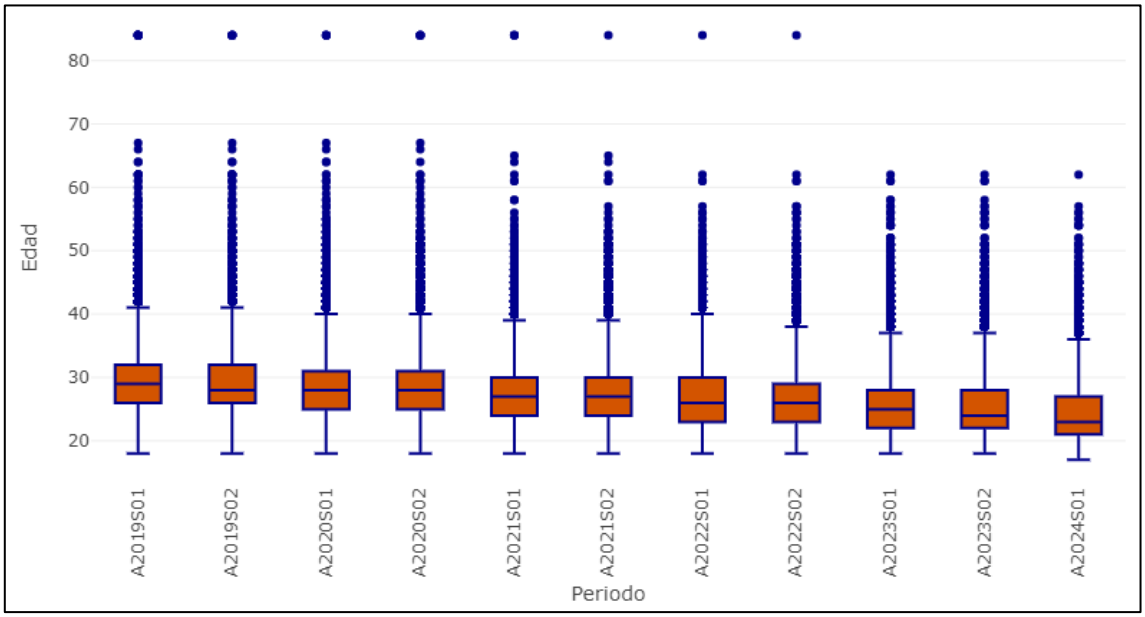
Figura 3: Proporción de estudiantes por género y continuidad



3.3.2. Distribución por Edad

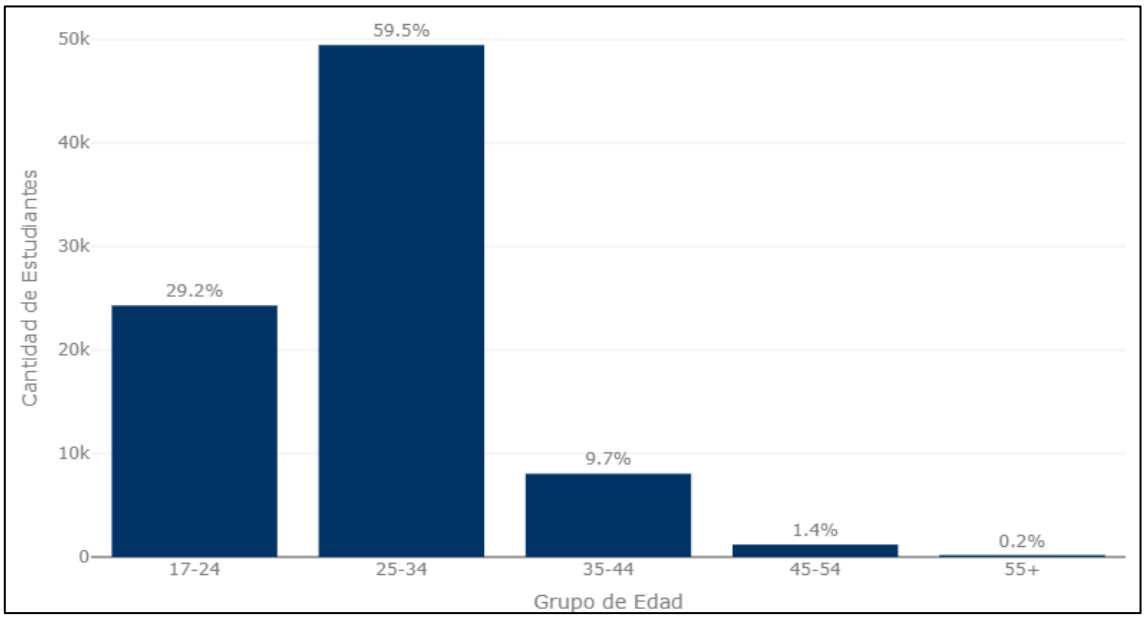
La mayoría de los estudiantes se encuentran entre los 18 y 38 años para todos los semestres, tal como se observa en la Figura 4.

Figura 4: Distribución de estudiantes por edad y periodo



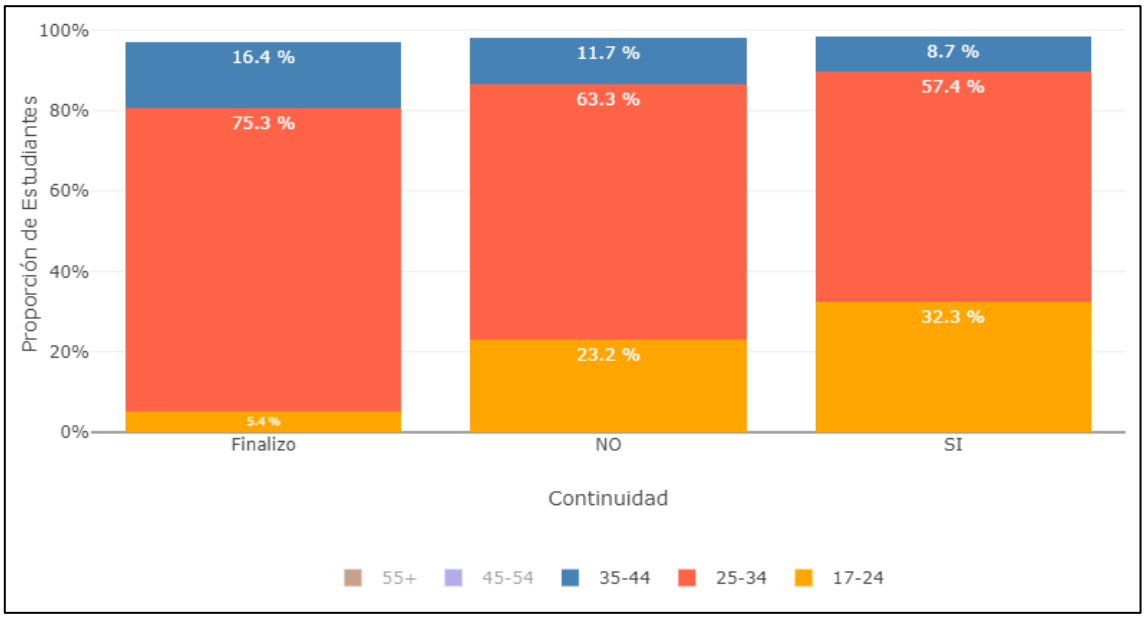
En la Figura 5 se puede observar que la mayoría de los estudiantes se encuentran en el grupo etario entre 25 y 34 años, con un 59.5%, seguido del grupo etario entre 18 y 24 años y los estudiantes mayores de 54 tienen un 0.2%

Figura 5: Distribución de estudiantes por intervalo de edad



Analizando cómo varía la distribución de las edades en función de la retención de los estudiantes sin considerar los periodos tenemos que el grupo etario de 17-24 y 25-34 son los grupos de interés para comprender de mejor manera la continuidad académica de un ciclo a otro. A continuación, se presenta de forma gráfica la proporción de retención por intervalo de edad.

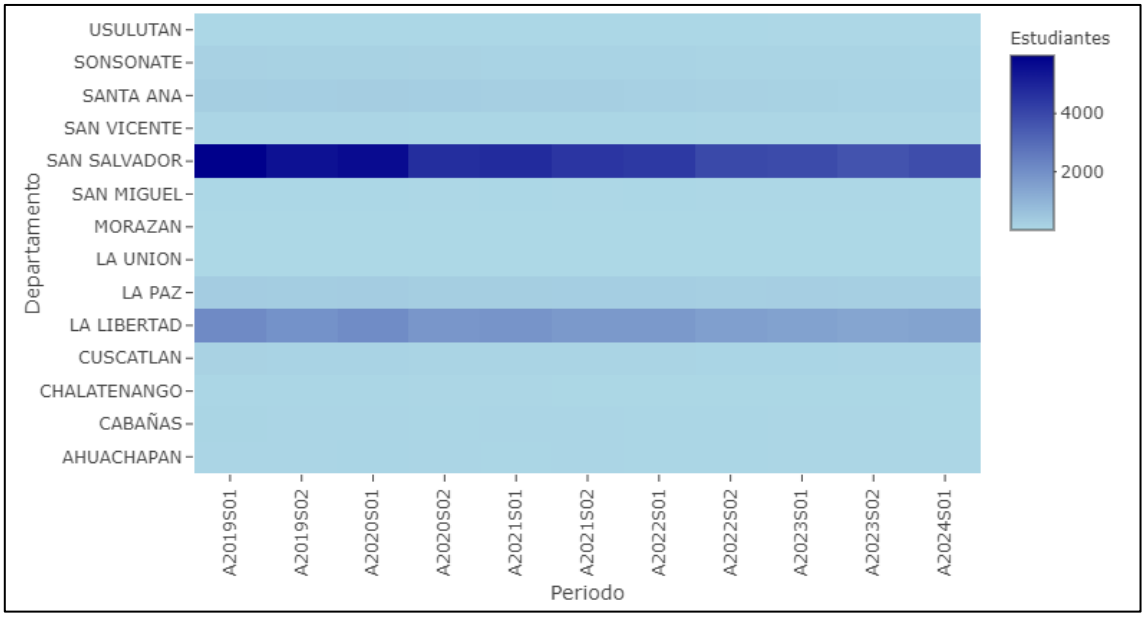
Figura 6: Proporción de retención por intervalo de edad y nivel de continuidad



3.3.3 Distribución Geográfica (Departamentos y Municipios)

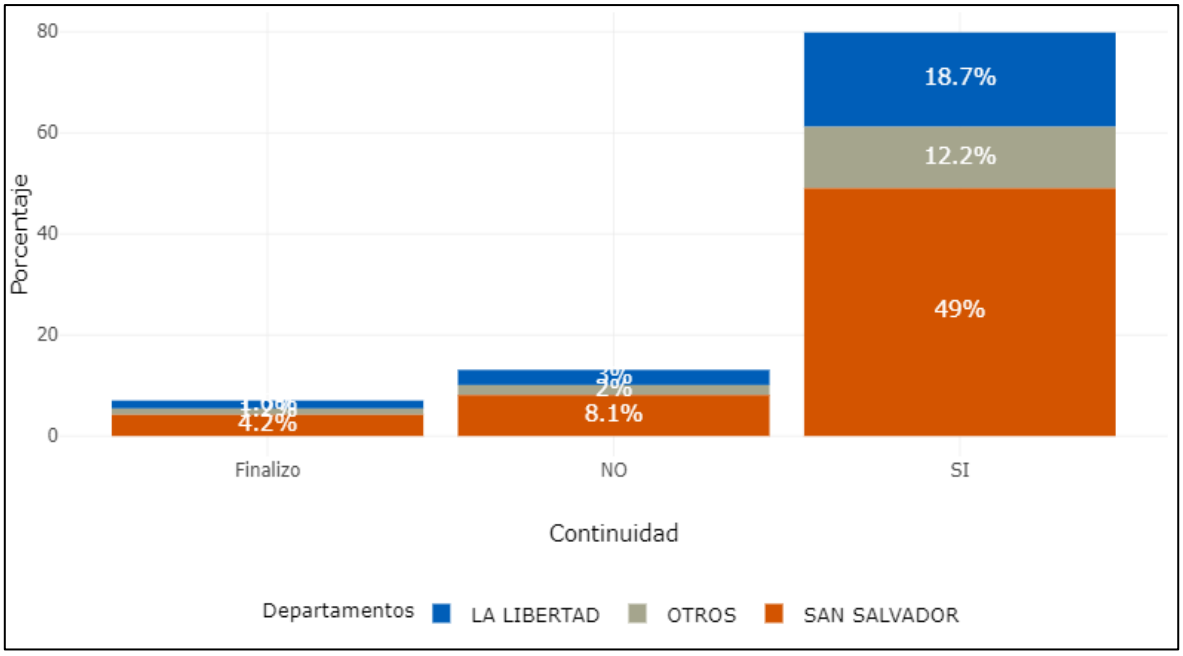
Los mayores grupos poblacionales se encuentran en los departamentos de San Salvador y La Libertad con el 61.3% y 22.3% respectivamente. En la Figura 7 se observa que esta relación se mantiene en todos los periodos.

Figura 7: Mapa de distribución de procedencia de los estudiantes



En la Figura 8 se analiza los departamentos con mayores tasas de deserción o retención.

Figura 8: Porcentaje de retención por continuidad y departamento



3.4. Análisis Socioeconómico

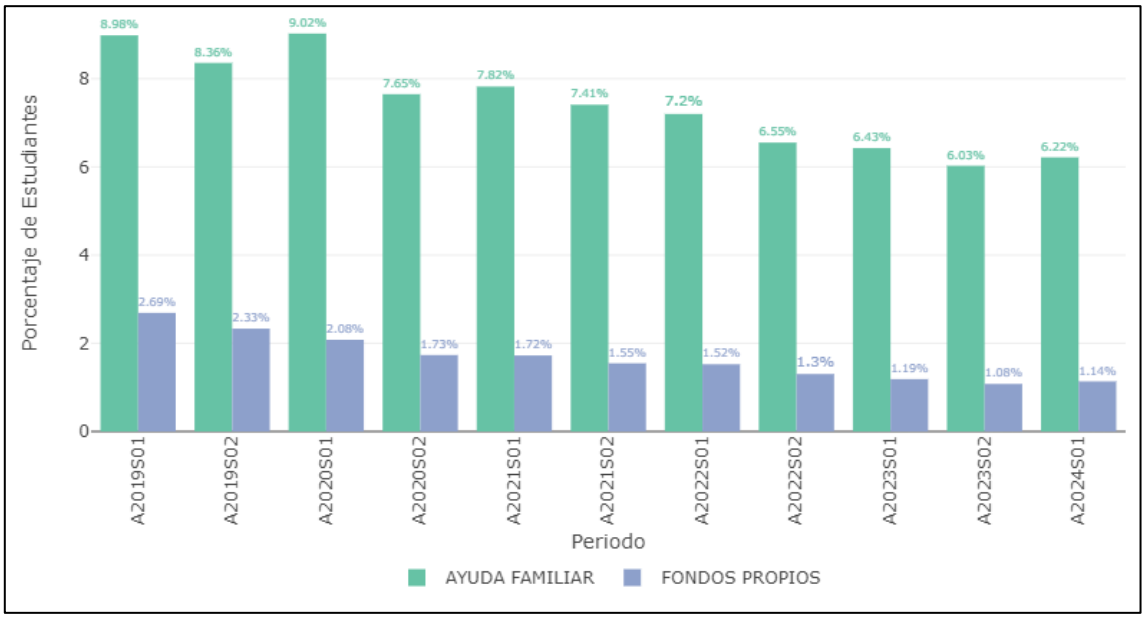
El análisis de las variables socioeconómicas es fundamental para entender cómo factores externos influyen en la retención estudiantil. Aspectos como el tipo de financiamiento y el estado familiar pueden tener un impacto significativo en la continuidad académica de los estudiantes, ya que estos factores suelen estar vinculados con las posibilidades de acceso a recursos, estabilidad emocional y otros determinantes que afectan el rendimiento académico y la permanencia en la educación superior.

En esta sección se examinarán dos dimensiones clave: la distribución por tipo de financiamiento y la distribución por estado familiar, teniendo como meta identificar patrones de riesgo que pueden contribuir a la deserción o, por el contrario, facilitar la permanencia y finalización de los estudios.

3.4.1. Distribución por Tipo de Financiamiento

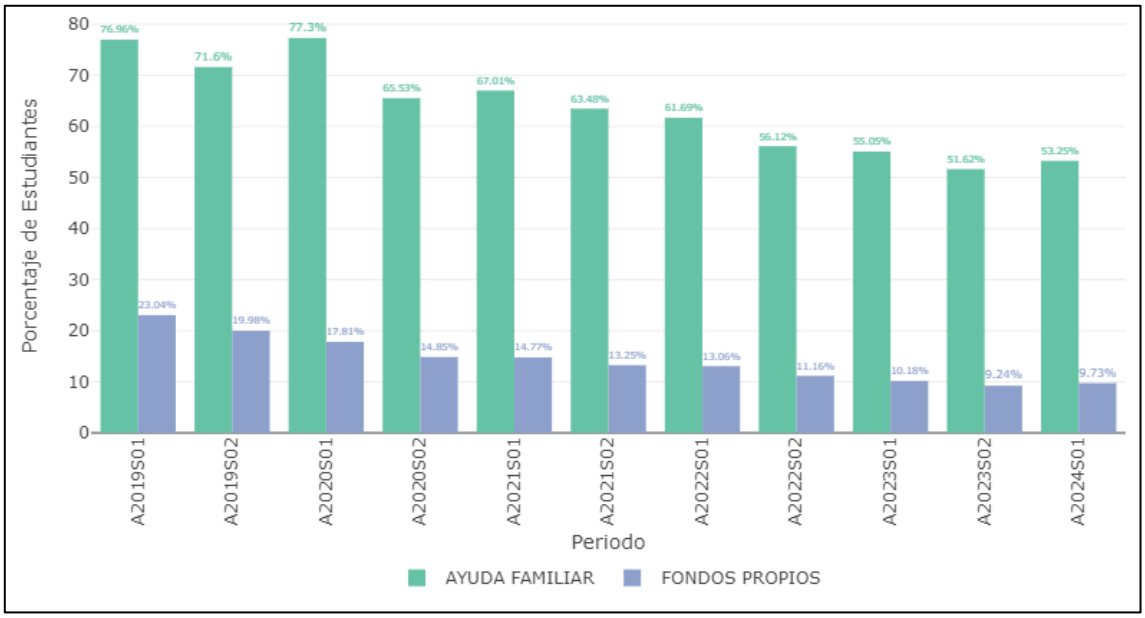
En la Figura 9, se puede observar que la mayoría gracias a la ayuda familiar que les brindan en cada semestre. También se puede observar una tendencia negativa en cuanto al total de estudiantes por periodo.

Figura 9: Distribución porcentual de estudiantes por tipo de financiamiento y periodo



Tomando como base el total de estudiantes del semestre uno del año 2019 tenemos que en este ciclo el 76.96% eran estudiantes que recibían ayuda familiar y el 23.04% estudian con fondos propios (trabajo a medio tiempo, tiempo completo o pasantías remuneradas, becas). Además, en la Figura 10, se puede observar que comparando el total de estudiantes se ve una tendencia negativa la cual ha generado que en el semestre uno del año 2024 se tiene un 62.98% de estudiantes, es decir un 37.02% menos en cuanto a cantidad con respecto al semestre uno del 2019.

Figura 10: Distribución porcentual de estudiantes por tipo de financiamiento y periodo relativo al semestre 01 del año 2019



Finalmente, comparando la Figura 11 y 12, podemos afirmar que, los que estudian con fondos propios son más propensos a dejar la carrera entre ciclos, pero también son los que más logran graduarse.

Figura 11: Proporción de estudiantes por tipo de financiamiento y continuidad

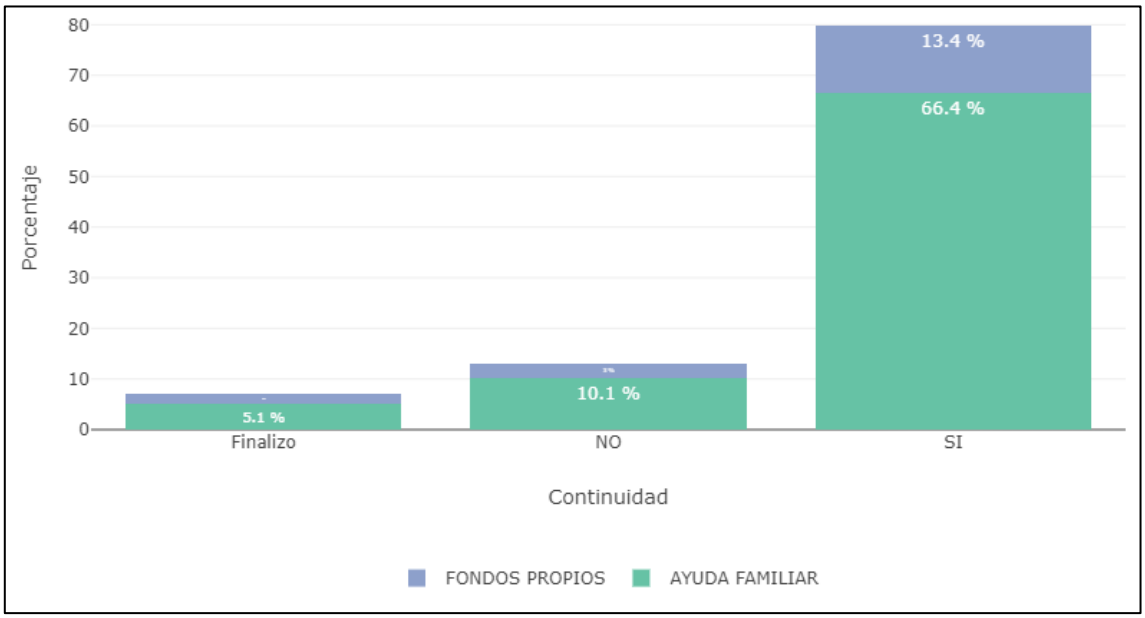
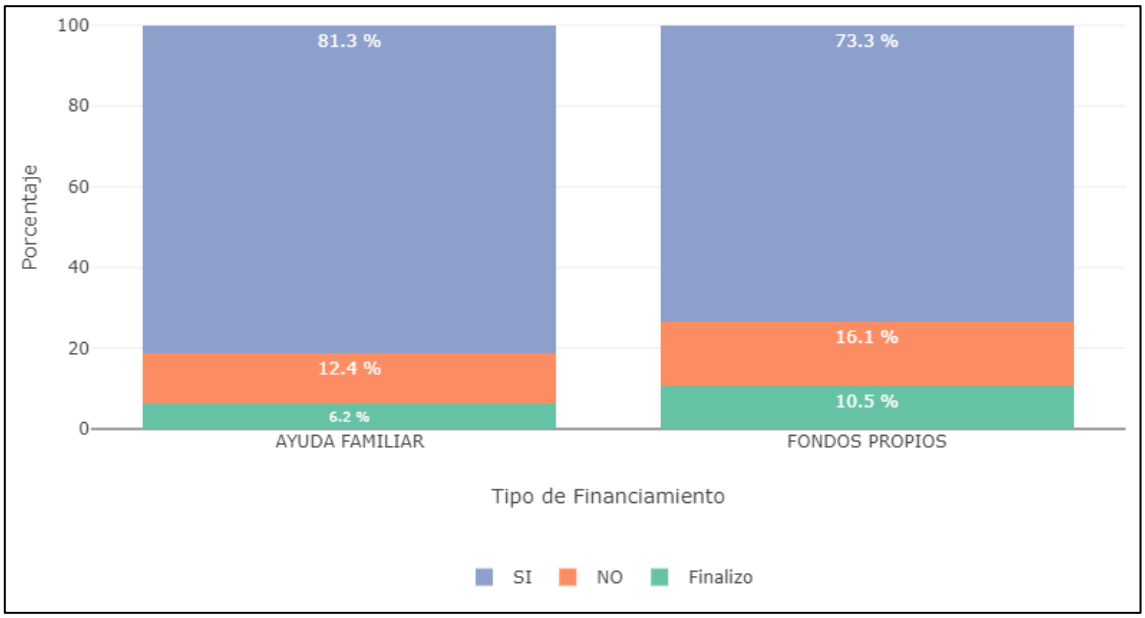


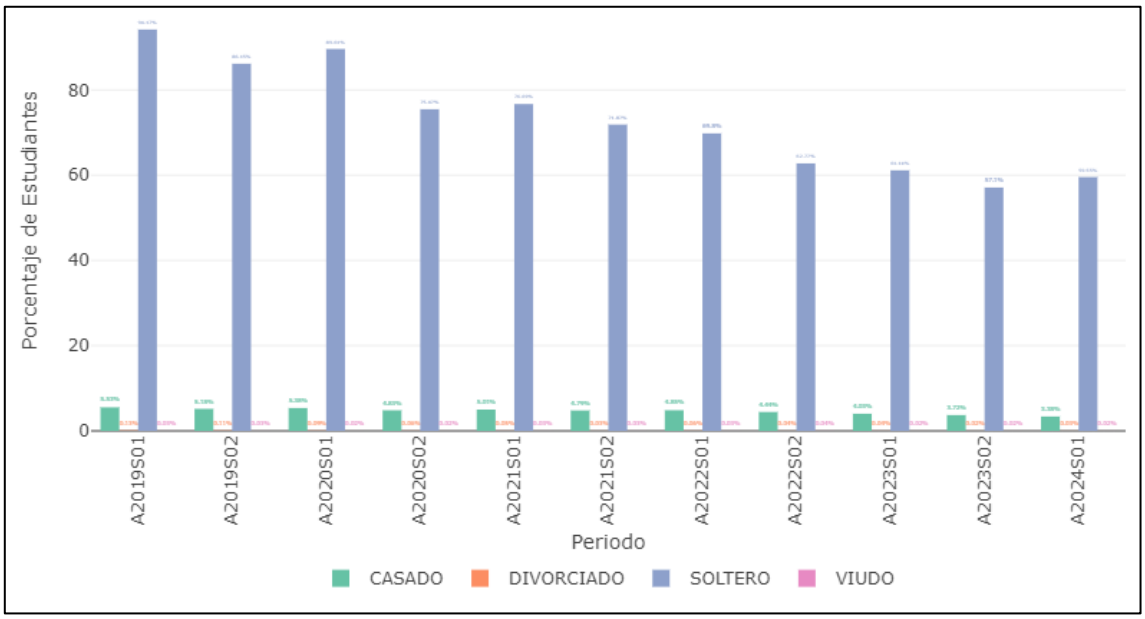
Figura 12: Comparativo de continuidad por tipo de financiamiento



3.4.2. Distribución por Estado Familiar

En la Figura 13, se ha tomado como base el total de estudiantes del semestre uno del año 2019 y con ello puede observar que comparando el total de alumnos se ve una tendencia negativa la cual ha generado que en el semestre uno del año 2024 se tenga una reducción de estudiantes en cuanto a cantidad con respecto al semestre uno del 2019. Por ejemplo, para Solteros están 94.17% y 59.55% lo cual representa una reducción del 34.62%, y para Casados están 5.53% y 3.38% representando una reducción del 2.15%.

Figura 13: Distribución porcentual de estudiantes por estado familiar y periodo relativo al semestre 01 del año 2019



Finalmente, comparando la Figura 14 y 15, podemos observar que, la continuidad entre ciclos es diferente en proporciones por cada tipo de estado familiar.

Figura 14: Proporción de estudiantes por estado familiar y continuidad

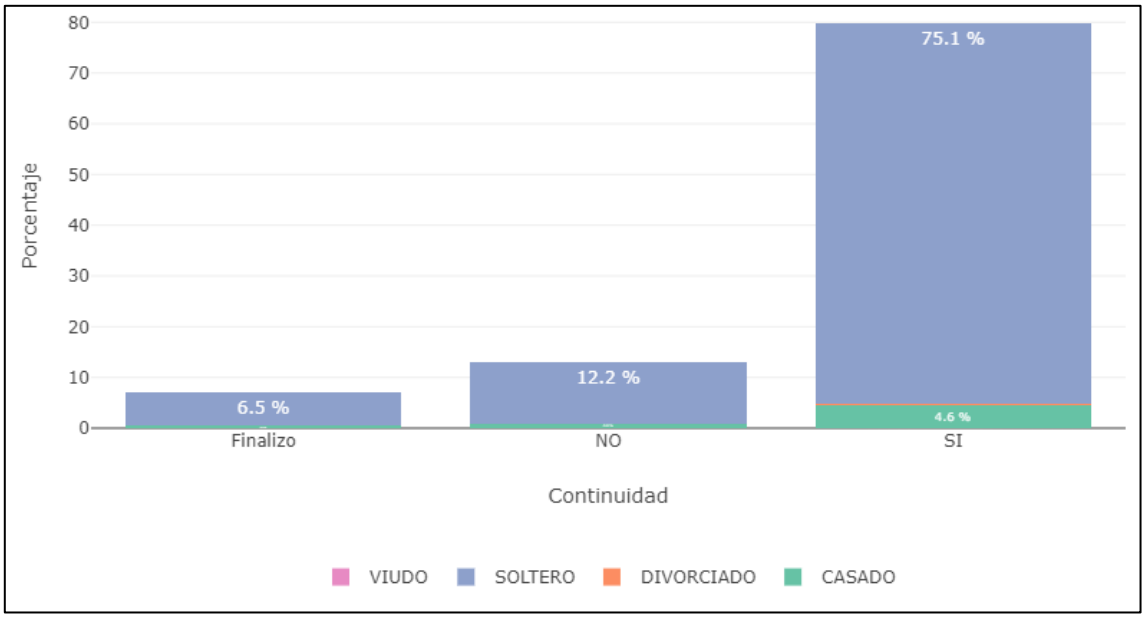
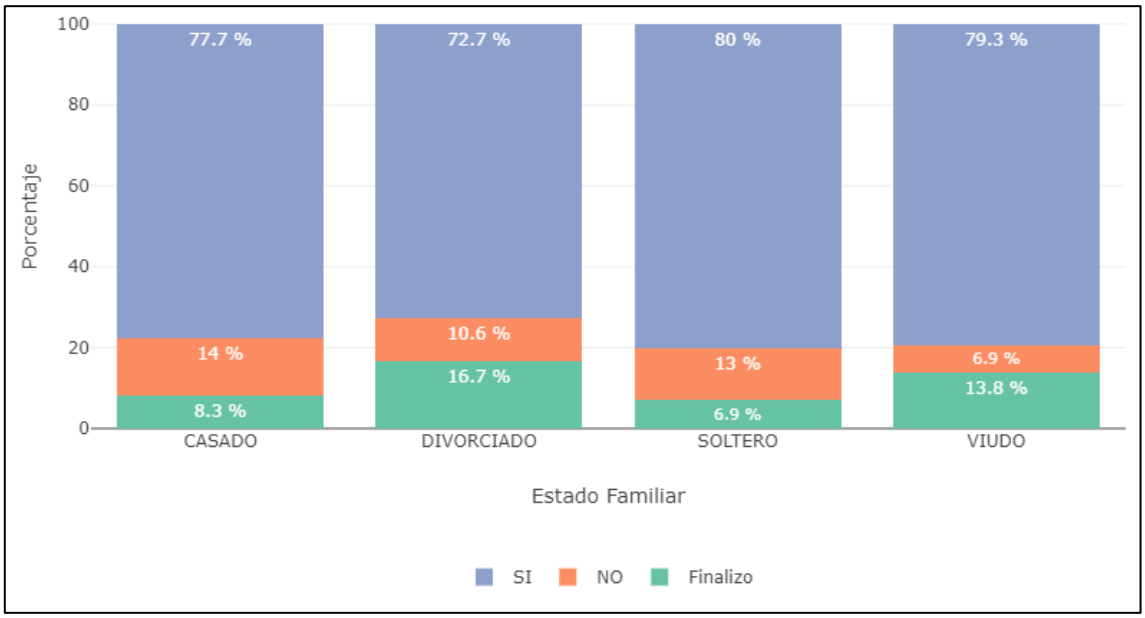


Figura 15: Comparativo de continuidad por estado familiar



3.5. Análisis Académico

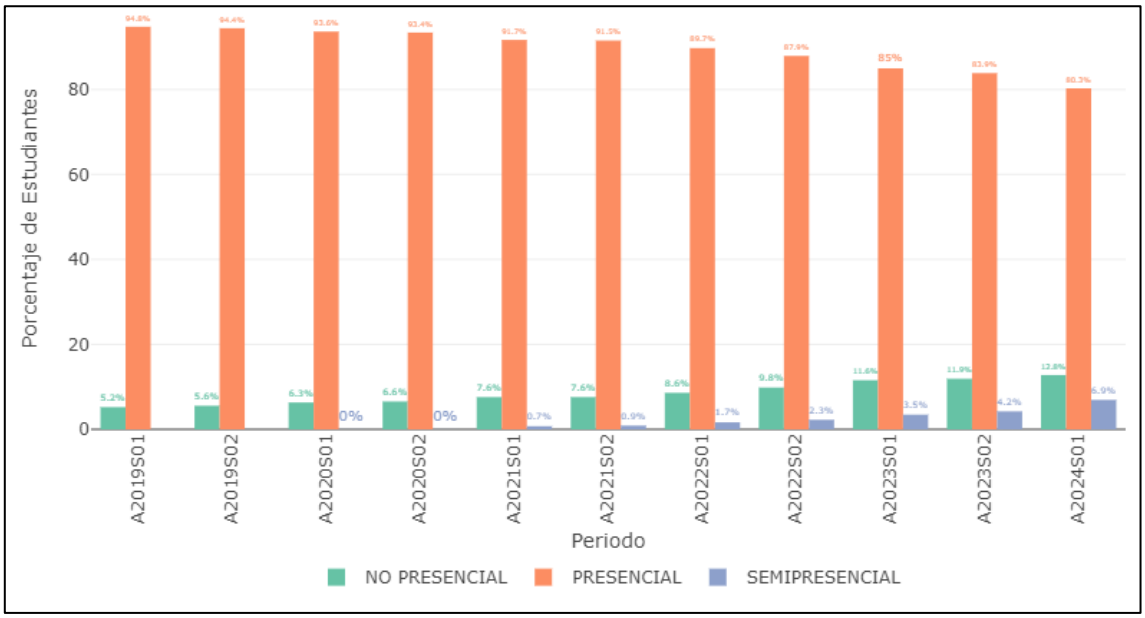
El análisis de las variables académicas busca proporcionar una comprensión más profunda de cómo factores institucionales, como la modalidad de estudio, la facultad y la carrera elegida, influyen en la trayectoria y la retención de los estudiantes. En esta sección se examina cómo estos factores se han distribuido y evolucionado a lo largo del tiempo, y se explora su relación con la continuidad académica.

Se abordarán tres dimensiones clave: distribución por modalidad, distribución por facultad y distribución por carrera. Mediante gráficos de distribución, proporción y tendencia, se analizará el comportamiento de los estudiantes en función de estos factores, destacando patrones de matrícula y deserción por periodo. Adicionalmente, se compararán los niveles de continuidad (Finalizado, Sí, No) en función de estas variables para identificar qué modalidades, facultades y carreras presentan mayores tasas de retención o riesgo de abandono.

3.5.1. Distribución de Estudiantes por Modalidad

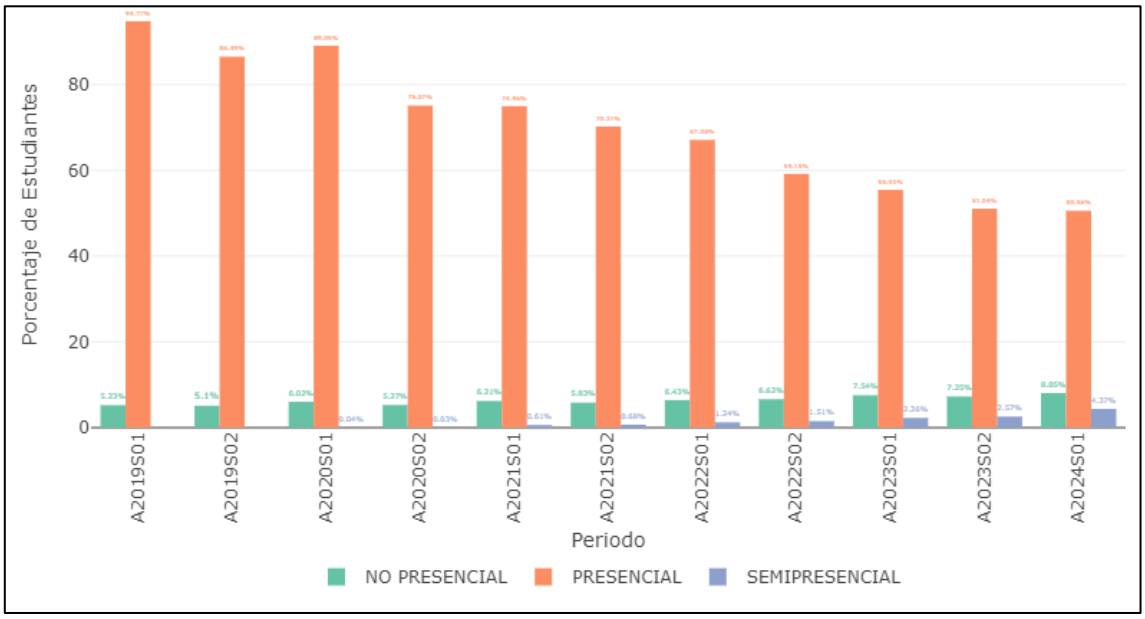
Para analizar cómo se distribuyen los estudiantes según la modalidad de estudio (presencial, semi presencial y no presencial) a lo largo de los diferentes periodos académicos, se presenta la Figura 16. Este nos permite visualizar cómo ha variado la inscripción en cada modalidad durante el periodo de estudio (2019-2024). Siendo la modalidad presencial la que predomina teniendo su valor más alto de 94.8% en el semestre uno del año 2019 y bajando hasta llegar al 80.3% el semestre uno del año 2024

Figura 16:Distribución de estudiantes por modalidad y periodo



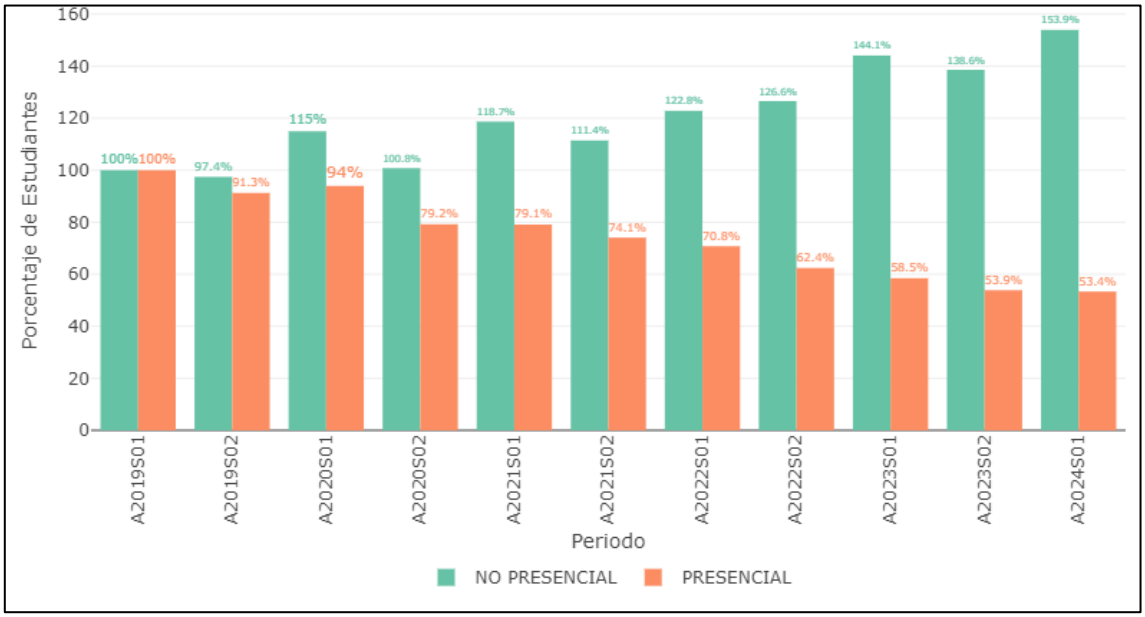
A continuación, se muestra un gráfico comparativo tomando como base el total de estudiantes del ciclo uno del año 2019, lo que permite observar las variaciones porcentuales en la inscripción por modalidad en los periodos posteriores. Este análisis nos ayuda a entender cómo ha decrecido la participación en cada modalidad a lo largo del tiempo, de modo que para el ciclo uno del año 2024 se tiene un 62.98% de estudiantes es decir un 37.02% menos.

Figura 17: Distribución de la proporción de estudiantes por modalidad y periodo, tomando como base el total de estudiantes del ciclo 01 año 2019



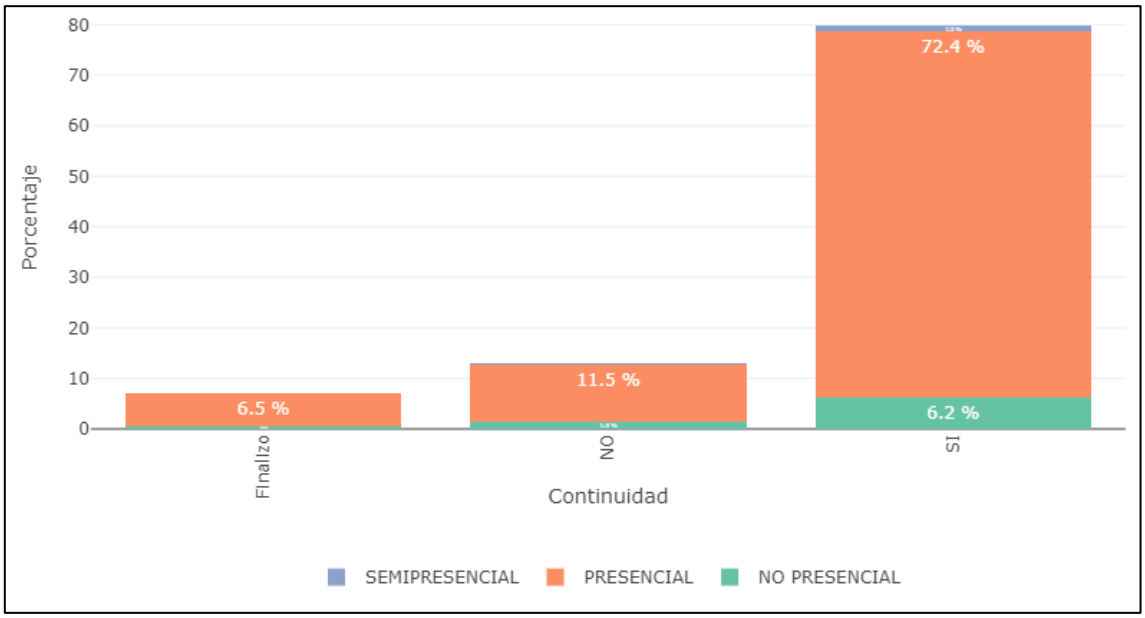
El siguiente gráfico muestra la distribución porcentual de estudiantes por modalidad, comparando cada periodo con el ciclo uno del año 2019, pero segmentado por modalidad. Con este gráfico, se puede identificar claramente qué la modalidad "No Presencial" ha experimentado un mayor crecimiento llegando hasta un 153.9%, en cambio la modalidad "Presencial" una disminución del 46.6% siendo esta ultima la que marca la tendencia por serla modalidad predominante.

Figura 18: Distribución porcentual de estudiantes por modalidad y periodo, tomando como base el total de estudiantes del ciclo 01 año 2019



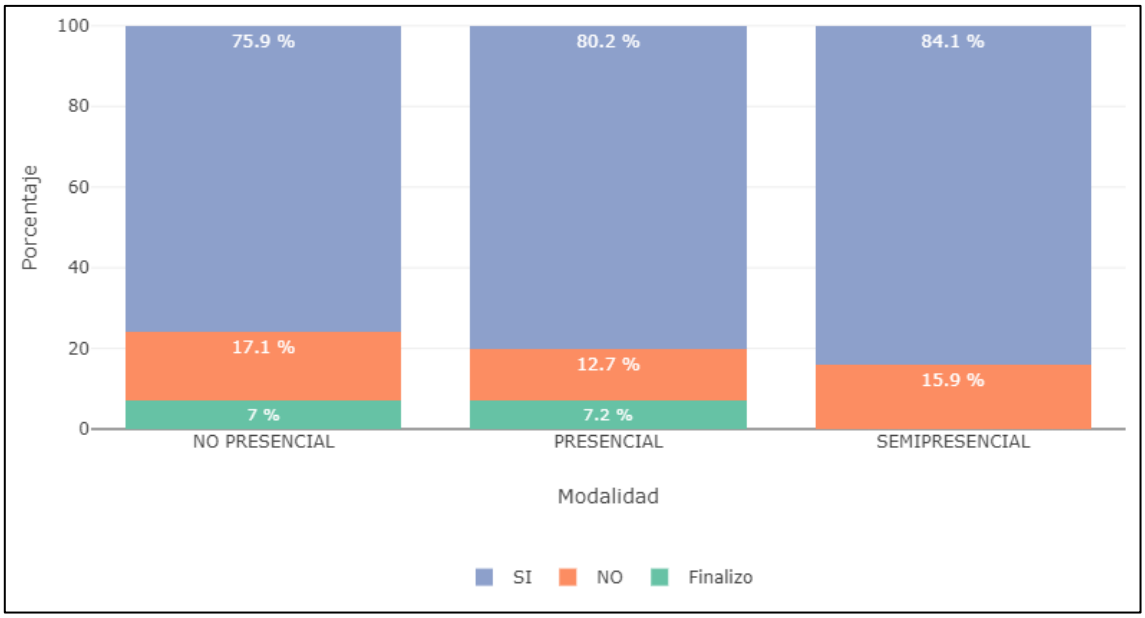
Además, se presenta un análisis sobre la proporción de retención de estudiantes según la modalidad de estudio y el nivel de continuidad (Finalizado, Sí, No). Este gráfico ilustra cómo la modalidad de estudio "Presencial" influye en las tasas de retención, por lo que, presenta mayor riesgo de deserción.

Figura 19: Proporción de retención de estudiantes por modalidad y nivel de continuidad



Finalmente, se compara la retención entre modalidades utilizando la modalidad en el eje X para visualizar claramente las diferencias en los porcentajes de continuidad entre las modalidades.

Figura 20: Comparación de los porcentajes de continuidad entre modalidades



3.5.2. Distribución por Facultad

A continuación, se presenta la tendencia de matrícula por facultad y periodo académico. La Figura 21, nos permite comparar cómo ha evolucionado la matrícula en cada facultad a lo largo del tiempo, y la Figura 22 nos permite identificar qué facultades han experimentado un decrecimiento significativo en relación con el ciclo uno del año 2019.

Figura 21: Tendencia de matrícula por periodo y facultad

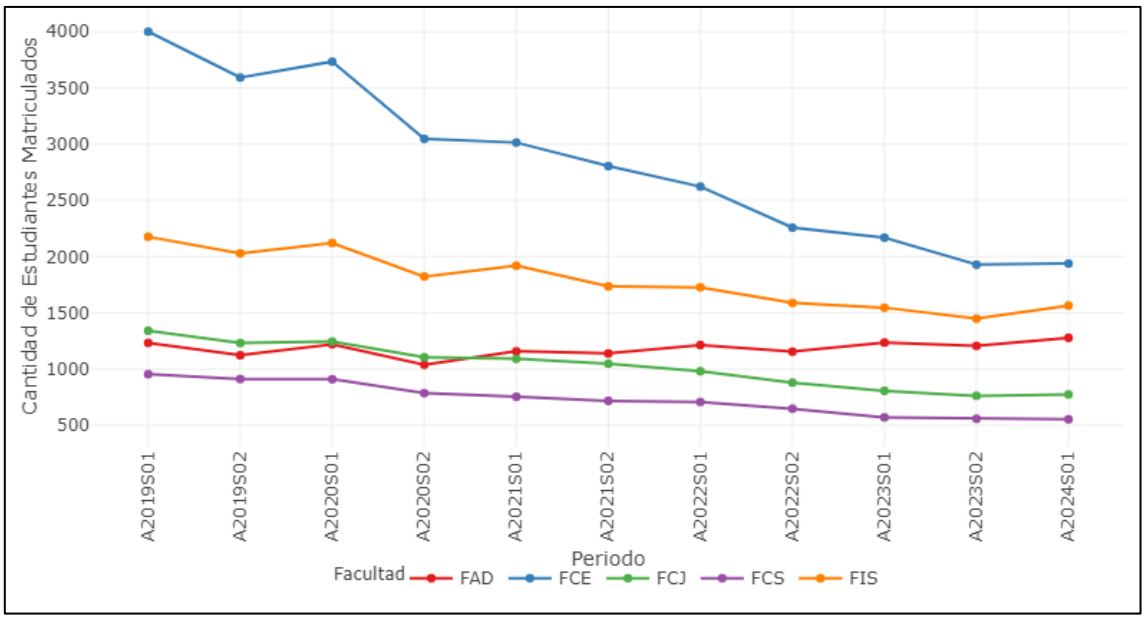
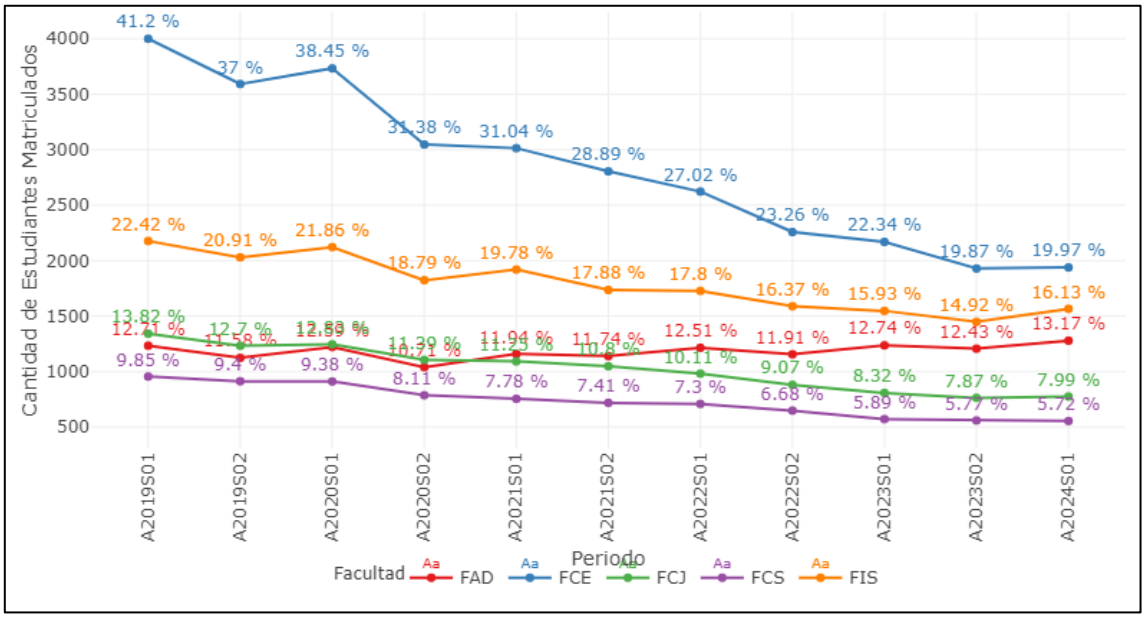
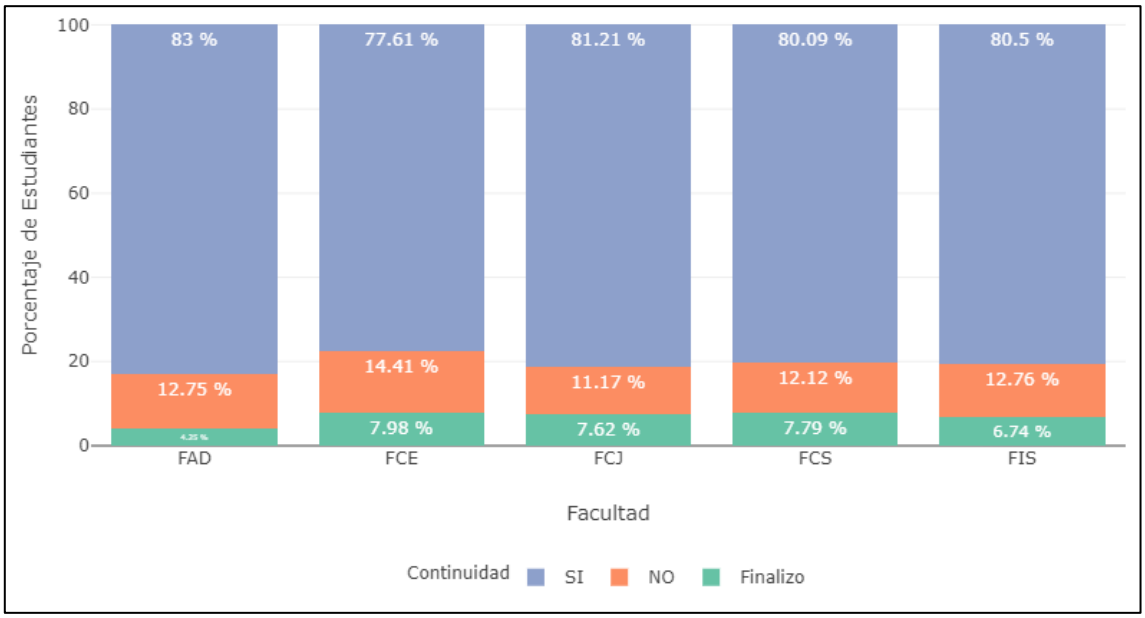


Figura 22: Tendencia de matrícula por periodo y facultad tomando como base el ciclo 01 del año 2019



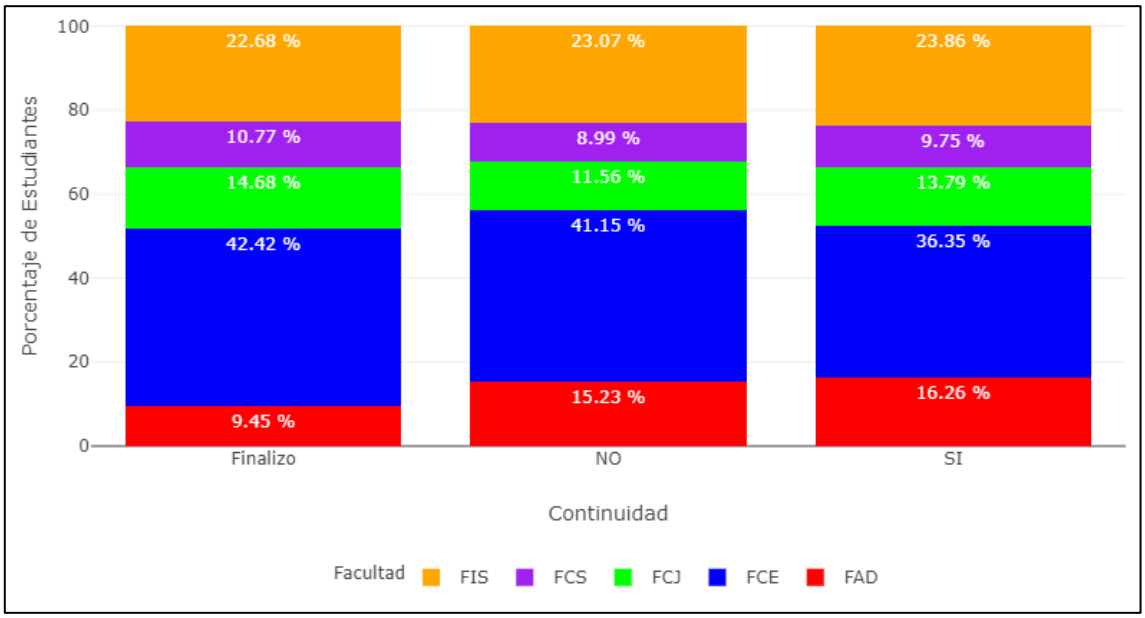
Se analiza también la continuidad de los estudiantes por facultad, comparando los niveles de retención y deserción. La Figura 23, permite identificar que la facultad FCE presentan mayores problemas de retención con un 14.41% y las que mantienen una continuidad más estable entre los estudiantes son FAD y FCJ con un 83% y 81.21% respectivamente.

Figura 23: Distribución de la continuidad por facultad



Finalmente, se compara la retención entre facultades utilizando el nivel de continuidad como eje para visualizar cómo varía la retención entre las diferentes facultades en función del porcentaje de participación, destacando la facultad FCE por manejar los mayores porcentajes en cada nivel de continuidad.

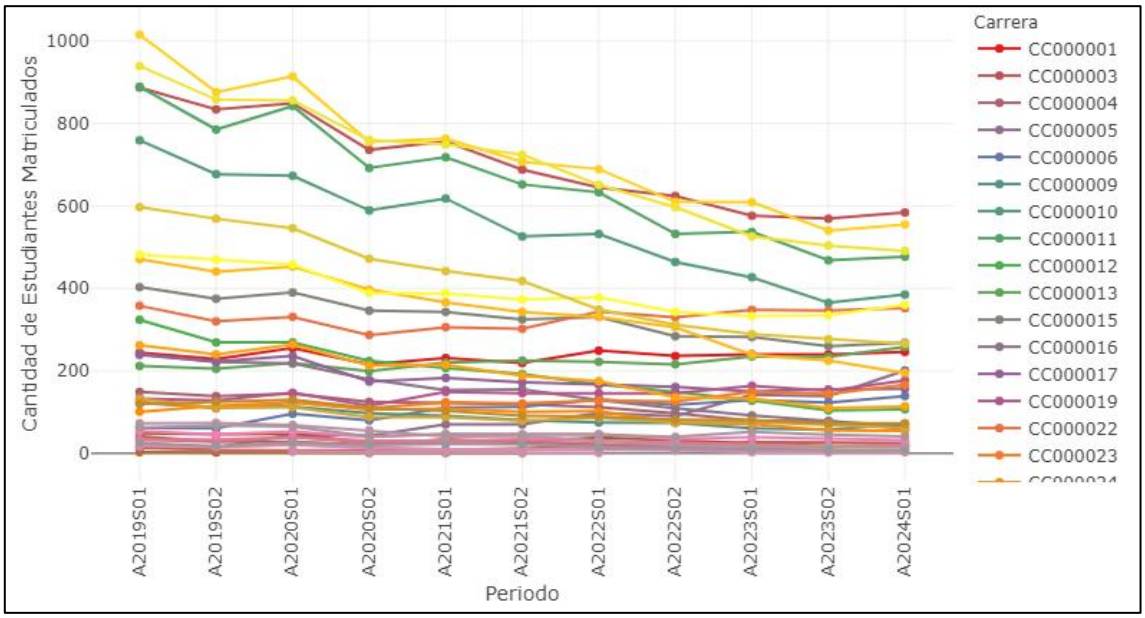
Figura 24: Comparación de facultades por continuidad



3.5.3. Distribución por Carrera

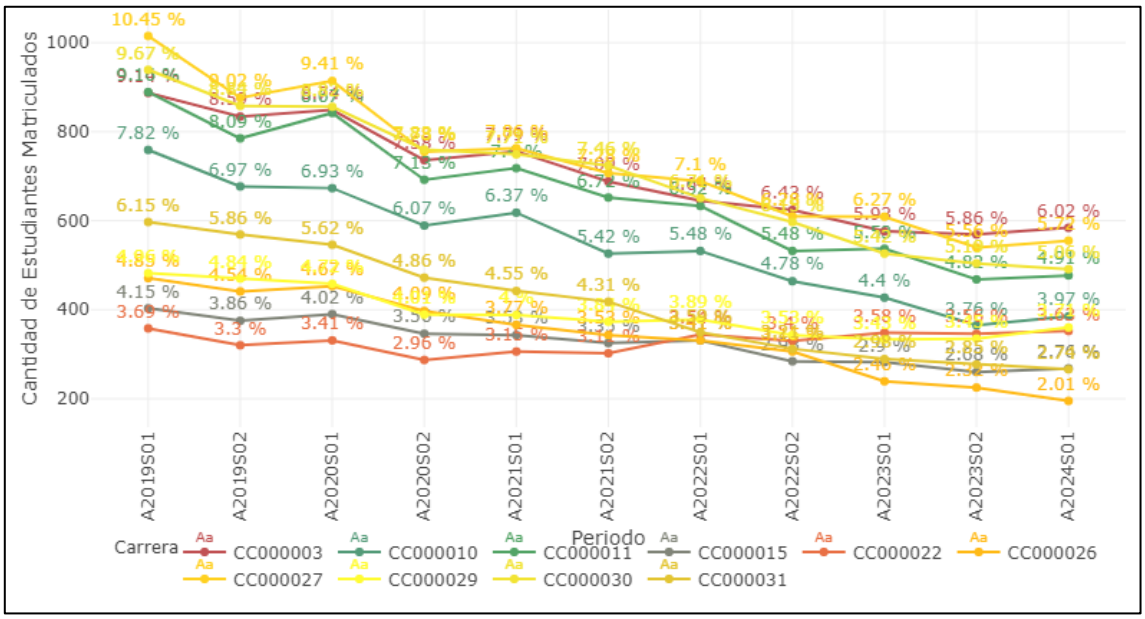
La Figura 25 muestra la tendencia de matrícula por carrera y periodo, permitiendo visualizar el comportamiento de la inscripción de los estudiantes en cada carrera a lo largo del tiempo. Este análisis ayuda a identificar qué carreras han mantenido una inscripción constante y cuáles han experimentado fluctuaciones importantes.

Figura 25: Tendencia de matrícula por periodo y carrera



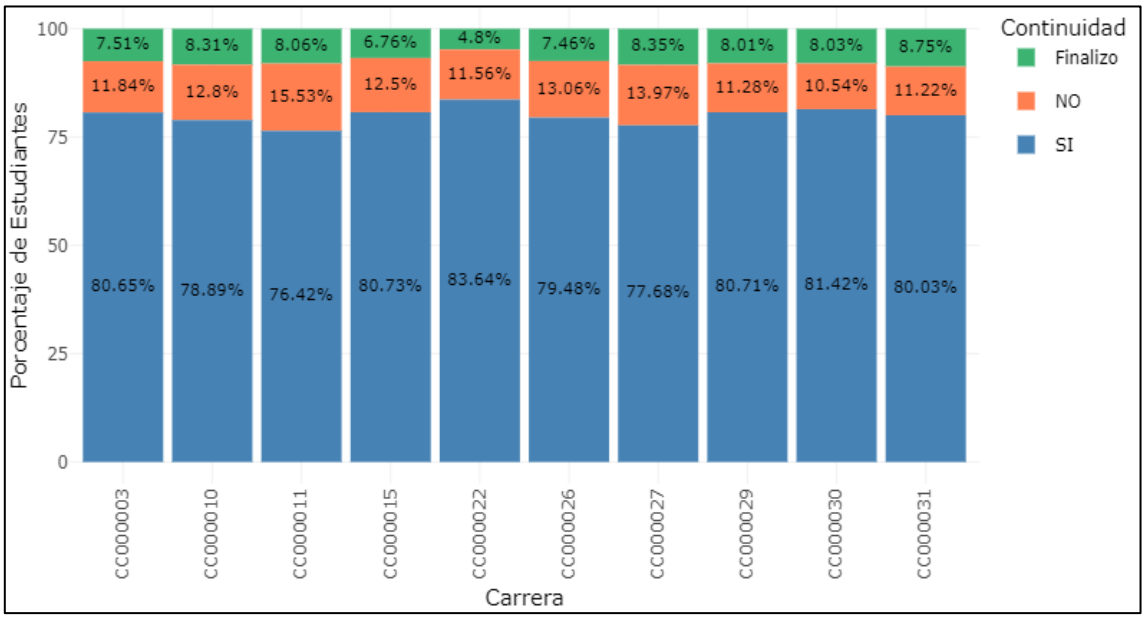
Además, se presenta la Figura 26 enfocada en el top 10 de carreras con mayor matrícula, comparando los periodos con el periodo inicial ciclo 01 del año 2019. Este análisis permite identificar qué carreras han mostrado una disminución o aumento significativo.

Figura 26: Tendencia de matrícula por periodo y top 10 de carreras, tomando como base ciclo 01 año 2019



A continuación, se analiza la continuidad de los estudiantes en las principales 10 carreras. La Figura 27 permite observar qué carreras presentan mayor retención y cuáles tienen mayor riesgo de deserción.

Figura 27: Distribución de la continuidad por top 10 de carreras



3.6. Análisis del Progreso Académico

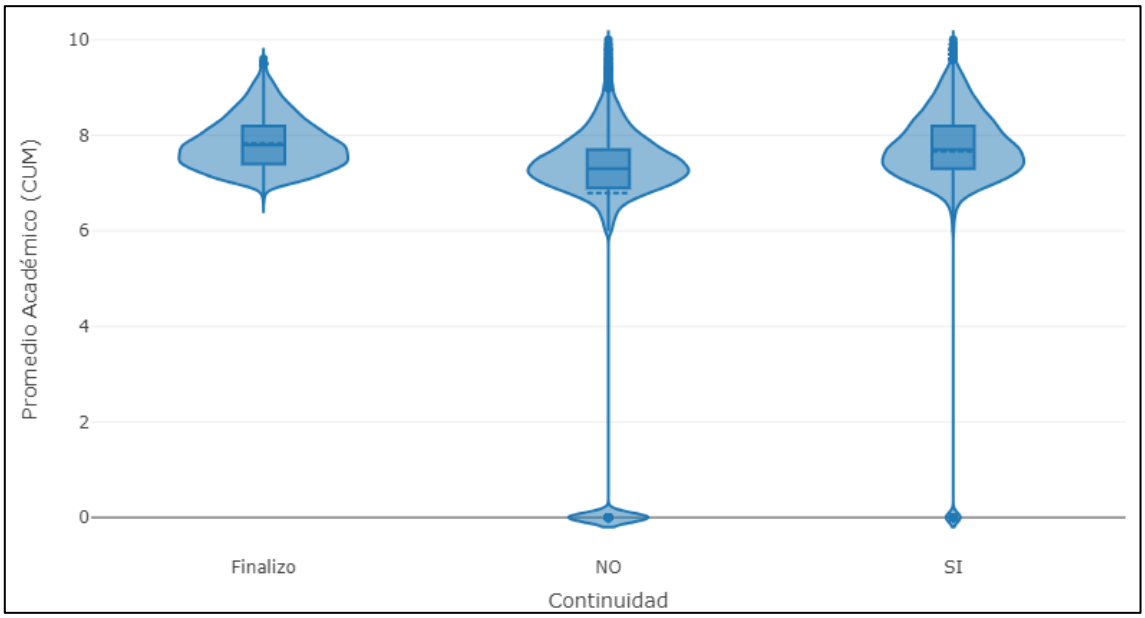
El progreso académico es un factor por tomar en cuenta para determinar la probabilidad de retención o deserción estudiantil. Variables como el promedio acumulado (CUM), el número de materias aprobadas o reprobadas, y el nivel académico alcanzado, proporcionan una visión detallada del desempeño de los estudiantes y su relación con la continuidad en sus estudios.

En esta sección se analizarán tres dimensiones clave del progreso académico: **Promedio Académico (CUM) y Nivel**, **Materias Aprobadas y Reprobadas**, y **Continuidad**. A través de gráficos y análisis bivariados, se explorará cómo el rendimiento académico y la acumulación de materias aprobadas o reprobadas impactan en la retención estudiantil, identificando patrones de riesgo y posibles puntos de intervención para mejorar la continuidad de los estudiantes en la universidad.

3.6.1. Promedio Académico (CUM) y Nivel

El rendimiento académico, medido a través del promedio acumulado (CUM), es un factor determinante en la continuidad estudiantil. En la Figura 28 se muestra la distribución del CUM entre los estudiantes, clasificada según su estado de continuidad (Finalizado, Sí, No). Este análisis permite observar cómo los promedios bajos se correlacionan con la retención y deserción.

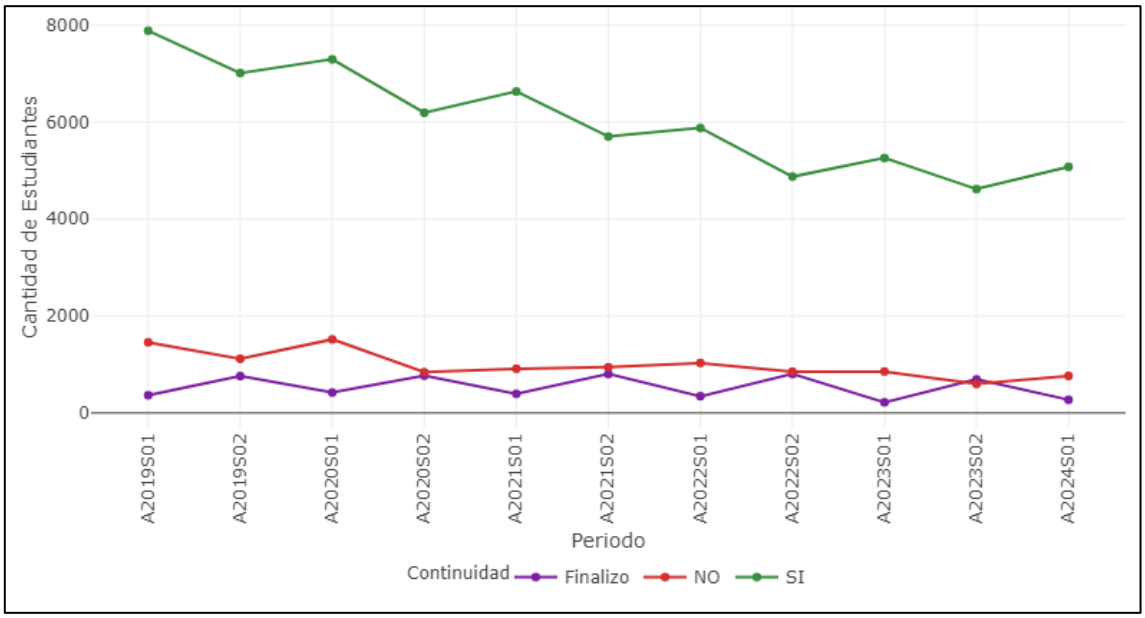
Figura 28: Distribución del promedio académico (CUM) y continuidad



3.6.2. Continuidad

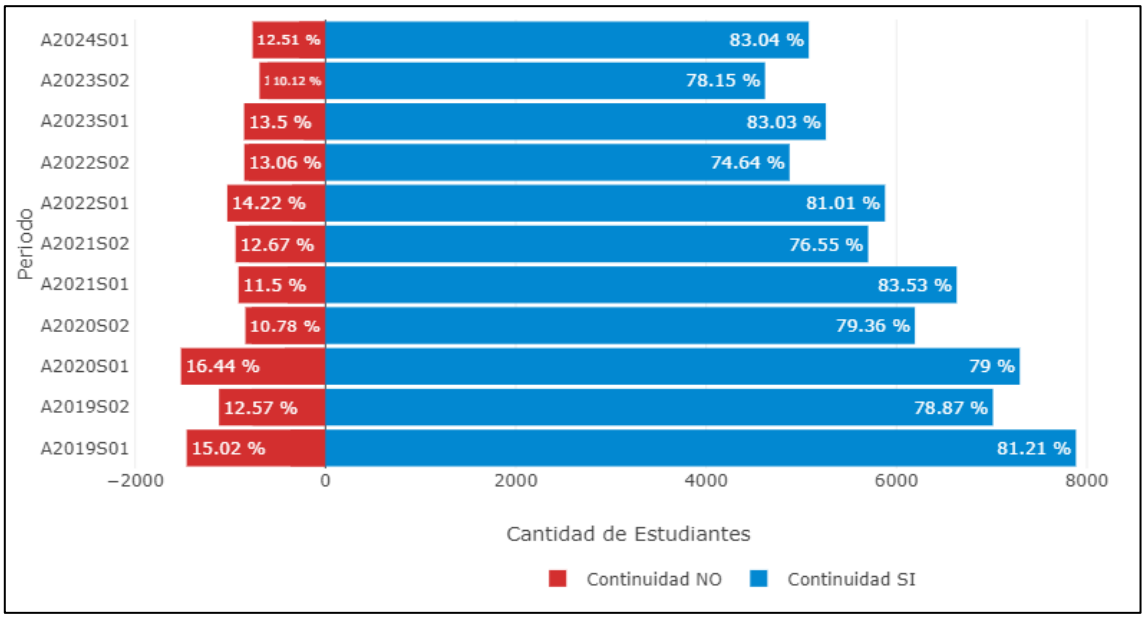
La Figura 29 presenta la evolución de la continuidad académica de los estudiantes a lo largo de los distintos periodos, desglosando la población estudiantil en tres categorías: estudiantes que han continuado, estudiantes que no han continuado, y aquellos que han finalizado sus estudios. El análisis visual permite identificar patrones temporales que pueden ser clave para comprender los factores que influyen en la retención y la finalización, así como para detectar posibles variaciones a lo largo de los periodos analizados.

Figura 29: Tendencia de tipo de continuidad por periodo



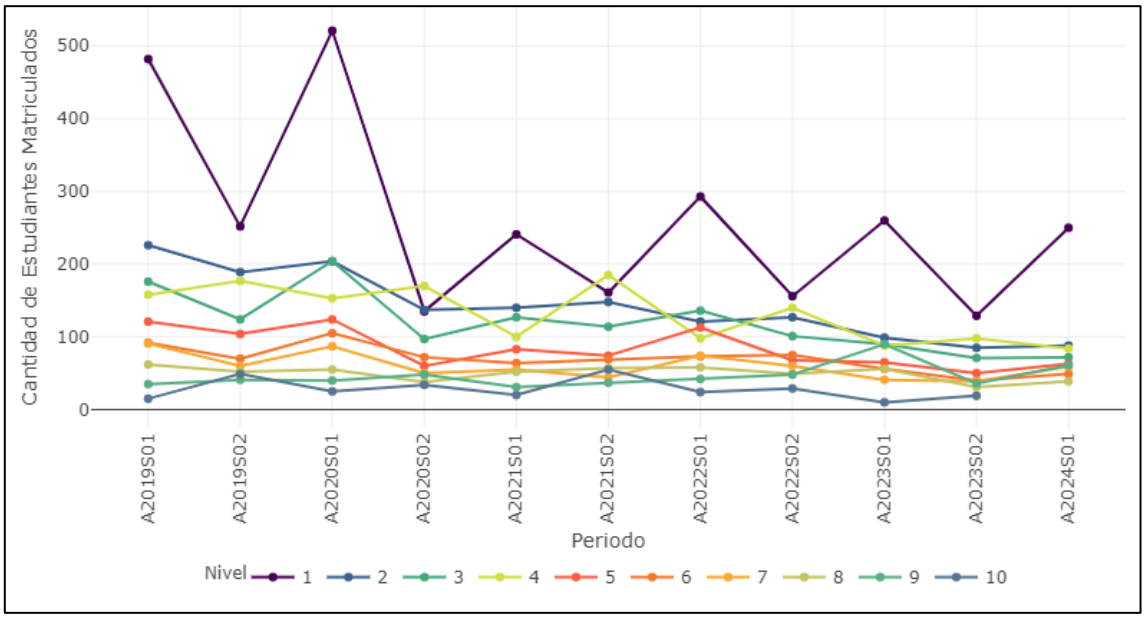
La Figura 30 presenta un gráfico piramidal el cual compara la distribución porcentual de los estudiantes que continúan y los que no continúan en cada periodo. Este enfoque facilita la visualización de la proporción relativa de ambos grupos, proporcionando una perspectiva clara sobre la estabilidad o fluctuación de la continuidad académica a lo largo del tiempo. Los porcentajes presentados dentro de las barras permiten una comprensión más precisa de la magnitud de estos fenómenos en cada periodo.

Figura 30: Distribución porcentual piramidal por periodo y continuidad



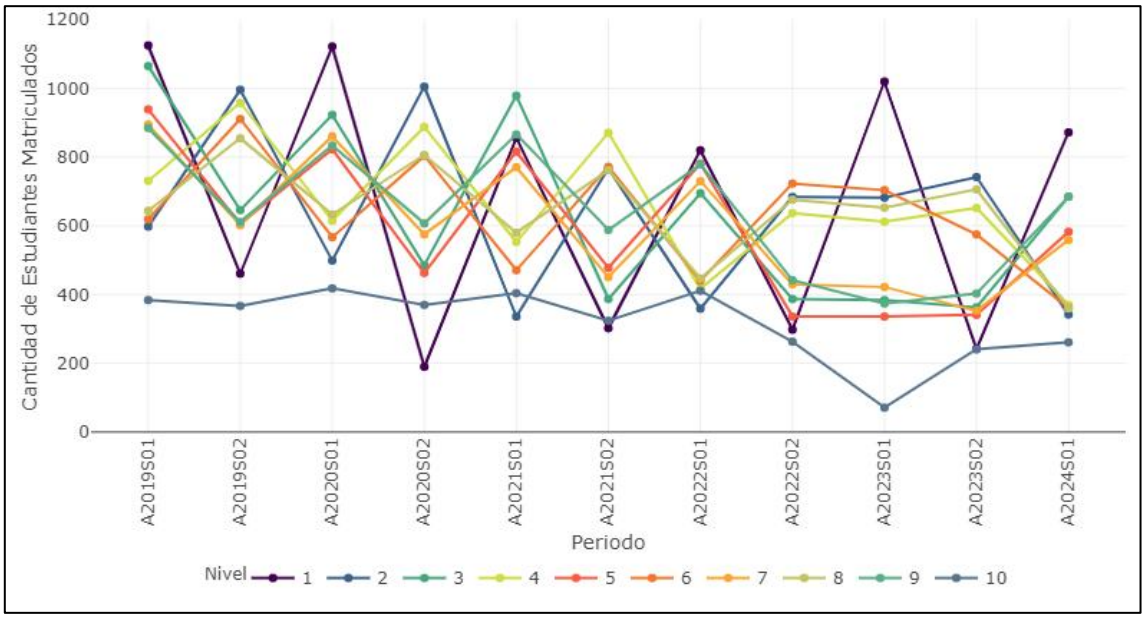
La Figura 30 muestra la evolución de los estudiantes que no han continuado sus estudios de un ciclo a otro, segmentados por su nivel académico y el periodo correspondiente. El objetivo es identificar en qué niveles académicos se concentran las mayores tasas de deserción y cómo estas tasas varían a lo largo de los periodos, lo que puede ser un indicio de los desafíos específicos que enfrentan los estudiantes en distintas etapas de su formación, siendo el nivel uno el que mas destaca entre todos ellos.

Figura 31: Tendencia de estudiantes que "No Continúan" por periodo y nivel



La Figura 32 muestra la tendencia de los estudiantes que han decidido continuar sus estudios de un ciclo a otro, categorizados por nivel académico y periodo. El análisis permite evaluar cómo la retención varía entre los diferentes niveles y en qué periodos se ha observado un mayor compromiso de los estudiantes con su continuidad académica, ofreciendo una visión completa del progreso académico de los estudiantes que permanecen en la institución.

Figura 32: Tendencia de estudiantes que "Sí Continúan" por periodo y nivel



3.7. Discusión del Análisis Descriptivo

En el análisis demográfico se evidencia que la población estudiantil está dominada por el género femenino en todos los periodos, con una tendencia decreciente en el total de estudiantes matriculados. A pesar de esta disminución, las mujeres presentan una mayor probabilidad de continuidad en la universidad en comparación con los hombres. Además, los grupos etarios de 18-24 años y 25-34 años son los más representativos, siendo también los más importantes para analizar la retención académica. Por otra parte, los departamentos de San Salvador y La Libertad concentran la mayor parte de la matrícula, manteniendo esta tendencia en todos los periodos, aunque presentan diferencias en las tasas de deserción que merecen mayor atención. Estos patrones destacan la relevancia de factores demográficos en la retención estudiantil, lo que puede guiar intervenciones focalizadas.

En el análisis socioeconómico, se observa una tendencia decreciente en el total de estudiantes, destacándose que la mayoría recibe ayuda familiar para financiar sus estudios. Aunque el grupo que depende de fondos propios es menor, son estos estudiantes los que presentan mayor probabilidad a abandonar la carrera entre ciclos, pero también son quienes tienen una mayor tasa de graduación. Además, la reducción en la matrícula es evidente tanto en estudiantes solteros como casados, siendo los solteros el grupo más afectado. La continuidad entre ciclos varía según el estado familiar, lo que resalta la influencia de las circunstancias socioeconómicas en la retención y finalización académica. Este análisis sugiere la importancia de considerar factores económicos y familiares al diseñar estrategias de retención estudiantil.

En el análisis académico, se observa que la modalidad presencial, aunque sigue siendo la predominante, ha experimentado una notable disminución en la matrícula desde 2019, mientras que la modalidad no presencial ha crecido significativamente. Este cambio sugiere una adaptación de los estudiantes a modalidades más flexibles, posiblemente impulsada por factores externos como la digitalización educativa. A su vez, la modalidad presencial muestra una mayor tasa de deserción, lo que indica la necesidad de intervenciones más focalizadas en este grupo. En cuanto a la distribución por facultad, la Facultad FCE enfrenta los mayores desafíos de retención, mientras que facultades como FAD y FCJ muestran una estabilidad en la continuidad estudiantil. Este patrón resalta la importancia de adaptar estrategias de retención según las características y necesidades específicas de cada facultad. En el análisis por carrera, se identifican fluctuaciones en la matrícula a lo largo de los años, destacando las 10 carreras con mayor inscripción. Algunas carreras muestran una tendencia decreciente, mientras que otras mantienen su atractivo. Además, la continuidad en estas carreras varía considerablemente, lo que sugiere que ciertas áreas académicas requieren más atención para mejorar las tasas de retención y disminuir la deserción.

En el análisis del progreso académico, se destaca la fuerte relación entre el promedio académico acumulado (CUM) y la continuidad estudiantil. Los estudiantes con promedios más bajos son más propensos a abandonar, mientras que aquellos con promedios más altos tienen mayores tasas de finalización. Este hallazgo sugiere que el rendimiento académico sería un predictor clave de la retención. La evolución temporal de la continuidad revela patrones significativos de fluctuación en la permanencia de los estudiantes, donde ciertos periodos parecen tener tasas de deserción más altas. Este análisis permite identificar los momentos en que la retención se ve más comprometida, ofreciendo una base para intervenciones más oportunas.

Por último, el nivel académico es un factor influyente, ya que los estudiantes en los primeros niveles presentan mayores tasas de abandono, el análisis de la retención a lo largo de los niveles académicos muestra que los estudiantes de niveles más avanzados tienden a continuar con mayor frecuencia, lo que resalta el compromiso académico a medida que los estudiantes progresan en sus estudios. Estos resultados proporcionan una visión integral del progreso académico, ofreciendo información clave para estrategias de retención.

3.8. Conclusión del Análisis Descriptivo

El análisis descriptivo realizado ha permitido obtener una visión detallada de los factores demográficos, socioeconómicos, académicos y de progreso estudiantil que influyen en la retención y deserción de los estudiantes en la universidad. Los resultados muestran que, si bien existe una tendencia general a la disminución de la matrícula en los últimos periodos, esta no afecta de manera uniforme a todos los grupos estudiantiles.

Desde la perspectiva demográfica, se identificó que las mujeres no solo representan la mayoría del estudiantado, sino que también presentan mayores niveles de continuidad. Asimismo, los grupos etarios jóvenes (18-34 años) concentran la mayor parte de la matrícula, constituyéndose como segmentos clave para la retención. Geográficamente, los departamentos de San Salvador y La Libertad lideran la concentración estudiantil, pero presentan diferencias en los niveles de deserción que requieren atención focalizada.

En cuanto a las condiciones socioeconómicas, se evidencia que la mayoría de los estudiantes cuenta con apoyo familiar para financiar sus estudios, mientras que aquellos que dependen de sus propios recursos enfrentan mayores riesgos de deserción, aunque también presentan una mayor proporción de finalización. Estas dinámicas muestran cómo las variables económicas y familiares afectan significativamente la permanencia académica.

A nivel académico, se observa un cambio estructural en la modalidad de estudio, con un crecimiento sostenido de la modalidad no presencial, la cual ha superado en crecimiento a la modalidad presencial desde 2019. Sin embargo, es la modalidad presencial la que presenta mayores niveles de abandono, lo que sugiere la necesidad de estrategias específicas para mejorar la experiencia y apoyo en este grupo. Asimismo, la continuidad varía considerablemente entre facultades y carreras, destacando algunas como la FCE por sus retos en retención, y otras como FAD y FCJ por su estabilidad.

Finalmente, el análisis del progreso académico confirma que el rendimiento académico, medido por el CUM, el número de materias aprobadas y reprobadas, y el nivel alcanzado, son factores clave para predecir la continuidad. Los estudiantes

con promedios más bajos y que se encuentran en niveles iniciales son más propensos a desertar, mientras que aquellos con buen desempeño y que han avanzado en su carrera muestran mayor compromiso y persistencia.

Capítulo 4: Análisis Exploratorio y Segmentación Estudiantil

4.1. Introducción

La retención estudiantil es un fenómeno complejo influenciado por múltiples factores académicos, socioeconómicos y personales. Comprender las razones detrás de la permanencia o deserción de los estudiantes requiere un análisis detallado que permita identificar patrones ocultos en los datos. En este sentido, el análisis exploratorio y la segmentación mediante técnicas de aprendizaje no supervisado sirven para descubrir patrones y tendencias en la población estudiantil y facilitar el diseño de estrategias de intervención más efectivas.

El objetivo de este capítulo es explorar la base de datos de estudiantes para segmentar la población en grupos homogéneos, utilizando técnicas de reducción de dimensionalidad y algoritmos de agrupamiento. Para ello, se implementa el Análisis de Componentes Principales (PCA) con el propósito de transformar las variables originales en un espacio de menor dimensión, facilitando la visualización de los datos y la identificación de las características más relevantes. Posteriormente, se aplica el algoritmo K-medias (K-means) que requiere definir el número de grupos óptimos, y la clusterización jerárquica que permite observar la estructura de los datos desde una perspectiva jerárquica sin necesidad de predefinir el número de clústeres, lo que permite una comprensión más profunda de los perfiles existentes y de sus posibles riesgos asociados.

Además, se emplea el método del codo para determinar el número óptimo de clústeres y lograr un equilibrio entre granularidad e interpretabilidad. Los

resultados obtenidos a partir de esta segmentación ofrecen una primera aproximación a la retención estudiantil desde un enfoque exploratorio, permitiendo diferenciar grupos de estudiantes con alto y bajo riesgo de deserción en función de sus características académicas, económicas y demográficas.

4.2. Metodología de Análisis No Supervisado

En esta sección se describen las técnicas estadísticas y de aprendizaje no supervisado utilizadas para segmentar la población estudiantil con base en características comunes, sin hacer uso de etiquetas previamente definidas. La finalidad es identificar patrones y grupos homogéneos que puedan explicar diferencias relevantes asociadas a la retención o deserción estudiantil. Se emplearon métodos de reducción de dimensionalidad y técnicas de agrupamiento, específicamente el Análisis de Componentes Principales (PCA), el algoritmo de K-medias y la clusterización jerárquica. Asimismo, se aplicaron criterios cuantitativos para la selección del número óptimo de clústeres.

4.2.1. Análisis de Componentes Principales (PCA)

El Análisis de Componentes Principales (PCA, por sus siglas en inglés) se utilizó como técnica preliminar para reducir la dimensionalidad del conjunto de datos, conservando la mayor cantidad de información posible. Dado que la base de datos contiene múltiples variables numéricas y categóricas transformadas, el PCA permite transformar dichas variables en un nuevo conjunto de componentes no correlacionados entre sí, los cuales explican la mayor proporción de la varianza observada.

Esta técnica facilita tanto la visualización de los datos como el proceso de clusterización posterior, al eliminar redundancias y reducir el ruido presente en las variables originales. En particular, se emplearon los primeros componentes

principales para proyectar los datos en espacios bidimensionales y tridimensionales, lo que permitió identificar visualmente la formación de agrupaciones naturales en la población estudiantil.

4.2.2. Algoritmo K-medias (K-means)

El algoritmo de K-medias fue seleccionado como técnica principal de agrupamiento no supervisado debido a su eficiencia y simplicidad para segmentar grandes conjuntos de datos. Este algoritmo parte de una asignación aleatoria de centros de clúster y realiza iteraciones sucesivas para minimizar la varianza intra-clúster, es decir, la distancia promedio entre cada observación y el centro del clúster al que pertenece.

Su aplicación en este estudio permitió identificar grupos de estudiantes con patrones similares en variables académicas, demográficas y socioeconómicas. La facilidad de implementación y su compatibilidad con datos numéricos estandarizados justifican su elección para una primera aproximación a la segmentación de la población.

4.2.3. Clusterización Jerárquica

Además del enfoque particional de K-medias, se empleó la **clusterización jerárquica** como método alternativo para analizar la estructura de los datos desde una perspectiva más flexible. Esta técnica permite generar un dendrograma que representa la similitud entre las observaciones, agrupándolas de manera ascendente (aglomerativa) a partir de las distancias entre pares de individuos o grupos.

El uso de la clusterización jerárquica permitió observar cómo se van formando las agrupaciones a diferentes niveles de similitud, sin necesidad de predefinir un número específico de clústeres. Esta característica fue especialmente útil para validar la consistencia de los grupos obtenidos con K-medias y para explorar posibles relaciones jerárquicas entre los perfiles estudiantiles.

4.2.4. Selección del Número Óptimo de Clústeres

La elección del número de clústeres (k) en el algoritmo de K-medias se realizó mediante el **método del codo**, el cual consiste en analizar la variación explicada por los clústeres en función de diferentes valores de k. Se identificó el punto de inflexión en la curva de la suma de cuadrados intra-clúster, donde las ganancias marginales en la reducción de la varianza comienzan a disminuir. Este punto representa un equilibrio entre simplicidad y capacidad explicativa del modelo.

En un primer análisis se estableció **k = 2**, lo que permitió realizar una segmentación básica de la población estudiantil en dos grandes grupos, diferenciando perfiles de mayor y menor riesgo de deserción. Sin embargo, con el objetivo de obtener una segmentación más detallada, se aplicó nuevamente el método del codo, y se determinó que **k = 5** ofrecía una mejor representación de la diversidad de perfiles estudiantiles, capturando patrones más específicos y significativos.

4.3. Resultados de la Segmentación

El análisis de datos educativos es una herramienta poderosa para identificar patrones y tendencias que impactan la trayectoria académica de los estudiantes. Dentro de este contexto, la clusterización se presenta como una técnica clave para segmentar grandes volúmenes de información en grupos homogéneos, ayudando a descubrir similitudes y diferencias en los perfiles estudiantiles. En este estudio, se decidió aplicar la clusterización por nivel, utilizando un conjunto de datos históricos de una universidad privada, con el propósito de mejorar la comprensión de los factores que influyen en la retención y el abandono.

La elección de realizar la clusterización en el Nivel 1 responde a la importancia crítica de esta etapa en la vida universitaria. El Nivel 1 incluye a los estudiantes que están en su primer año o en los primeros créditos de su carrera, una fase que suele determinar su continuidad en la educación superior. Durante este periodo inicial, los estudiantes enfrentan desafíos significativos, como adaptarse a la dinámica universitaria, manejar el aumento en la carga académica y equilibrar responsabilidades personales, sociales y económicas. Esta etapa también es el momento en el que los estudiantes construyen su sentido de pertenencia con la institución, un factor crucial para su permanencia. Las tasas de abandono en este nivel son históricamente más altas, ya que algunos estudiantes descubren que la carrera o la modalidad no cumple con sus expectativas, mientras que otros enfrentan barreras externas como problemas financieros o falta de apoyo familiar. Por estas razones, analizar los perfiles de los estudiantes en el Nivel 1 se vuelve esencial para identificar los factores que contribuyen al abandono temprano y diseñar estrategias de intervención personalizadas que aseguren un inicio exitoso en su trayectoria académica.

Para realizar el análisis de clusterización en el Nivel 1, se seleccionaron un conjunto de variables

1. **Demográficas:**
    - **Género:** Representado como masculino (M) y femenino (F), esta variable permite analizar posibles diferencias en patrones de retención entre hombres y mujere
    - **Estado Civil:** Incluye categorías como soltero, casado o divorciado, proporcionando información sobre el contexto familiar de los es
    - **Edad:** Permite identificar si existen tendencias relacionadas con estudiantes de mayor o menor edad en este nivel.
2. **Socioeconómicas:**
    - **Tipo de Financiamiento:** Clasificado en opciones como fondos propios, ayuda familiar, becas o préstamos, esta variable refleja el nivel de apoyo económico disponible para el estudiante.

- **Estado Familiar:** Se refiere a si el estudiante vive con su familia, es independiente o tiene dependientes económicos, lo que puede influir en su tiempo disponible y nivel de compromiso.

3. **Académicas:**
    - **Promedio Acumulado (CUM):** Una métrica clave para identificar el rendimiento académico de los estudiantes.
    - **Asignaturas Aprobadas y Reprobadas:** Estas variables ofrecen un panorama del progreso académico en términos de desempeño en los cursos matriculados.
    - **Modalidad de Estudio:** Clasificada en presencial, semipresencial y no presencial, refleja las preferencias de los estudiantes y su disponibilidad para participar en actividades académicas.
4. **Progreso Estudiantil:**
    - **Continuidad en la Matrícula:** Indica si el estudiante ha mantenido la inscripción regular durante su trayectoria académica.
    - **Participación Extracurricular:** Refleja el grado de integración del estudiante en actividades fuera del aula, lo cual puede estar relacionado con su nivel de compromiso y satisfacción con la experiencia universitaria.

4.3.1. Visualización de los Grupos con PCA

Como primer paso del análisis no supervisado, se aplicó el Análisis de Componentes Principales (PCA) con el objetivo de reducir la dimensionalidad del conjunto de datos y facilitar la visualización de posibles agrupaciones naturales entre los estudiantes. Esta técnica permitió proyectar las observaciones en un espacio bidimensional, preservando la mayor proporción posible de varianza total.

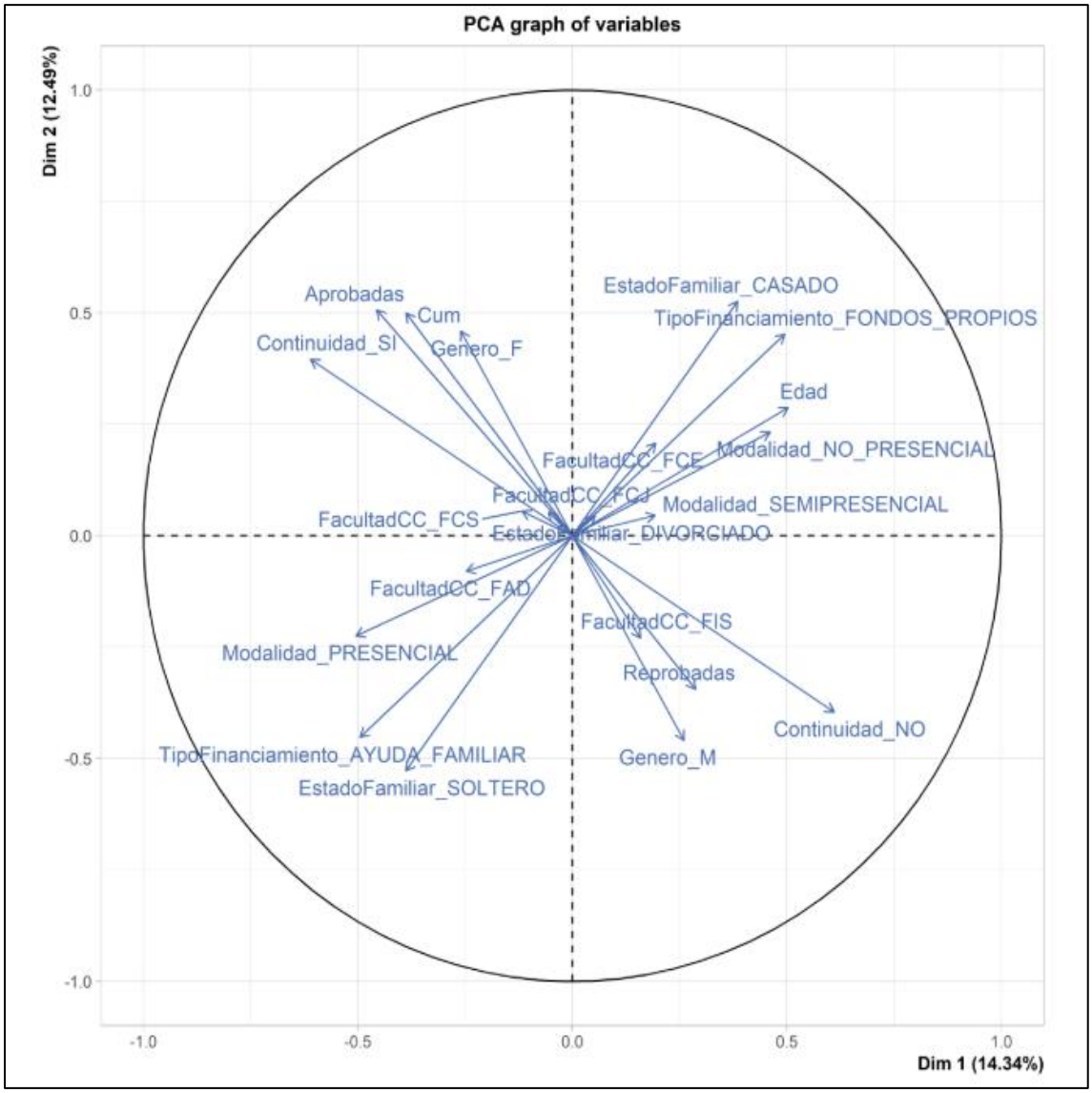
Antes de aplicar el PCA, se llevó a cabo un tratamiento exhaustivo de las variables para garantizar que fueran adecuadas para el análisis. Este proceso incluyó los siguientes pasos:

1. **Transformación a Variables Dummy:**
    - Las variables categóricas, como "modalidad de estudio" y "estado civil", se transformaron en variables dummy para convertirlas en un formato numérico compatible con el PCA. Por ejemplo, la modalidad de estudio (presencial, semipresencial, en línea) se dividió en columnas binarias donde cada columna indicaba la presencia o ausencia de cada categoría para cada estudiante.
    - Esta transformación permite que los algoritmos consideren las categorías como características separadas, evitando cualquier interpretación errónea de valores ordinales que no existen en estas variables.
2. **Estandarización:**
    - Las variables continuas, como edad, promedio acumulado (CUM) y el número de asignaturas aprobadas o reprobadas, se estandarizaron utilizando la técnica de z-score. Esto implica que cada valor de la variable fue transformado restando su media y dividiendo entre su desviación estándar.
    - La estandarización fue crucial porque el PCA depende de la magnitud de los valores; las variables con escalas más grandes podrían dominar el análisis si no se ajustan. Por ejemplo, sin estandarización, el rango de la variable "edad" podría influir más que las asignaturas aprobadas debido a la diferencia en sus escalas.
3. **Selección de Variables Relevantes:**
    - Solo se incluyeron las variables con mayor relevancia para la retención estudiantil y aquellas que podían contribuir significativamente a los patrones latentes en los datos. Estas variables incluyeron: edad, promedio acumulado, número de asignaturas reprobadas y aprobadas, estado civil, modalidad de estudio y tipo de financiamiento.
    - La selección de estas variables fue guiada por la literatura previa y análisis exploratorios, buscando un balance entre preservar la información y reducir la dimensionalidad.

La representación gráfica de la Figura 33, presenta la relación entre las diferentes variables incluidas en el análisis de retención estudiantil. Aquí se descompone el

comportamiento de las variables y su asociación en el espacio definido por los componentes principales (Dim 1 y Dim 2).

Figura 33: Grafica de variables



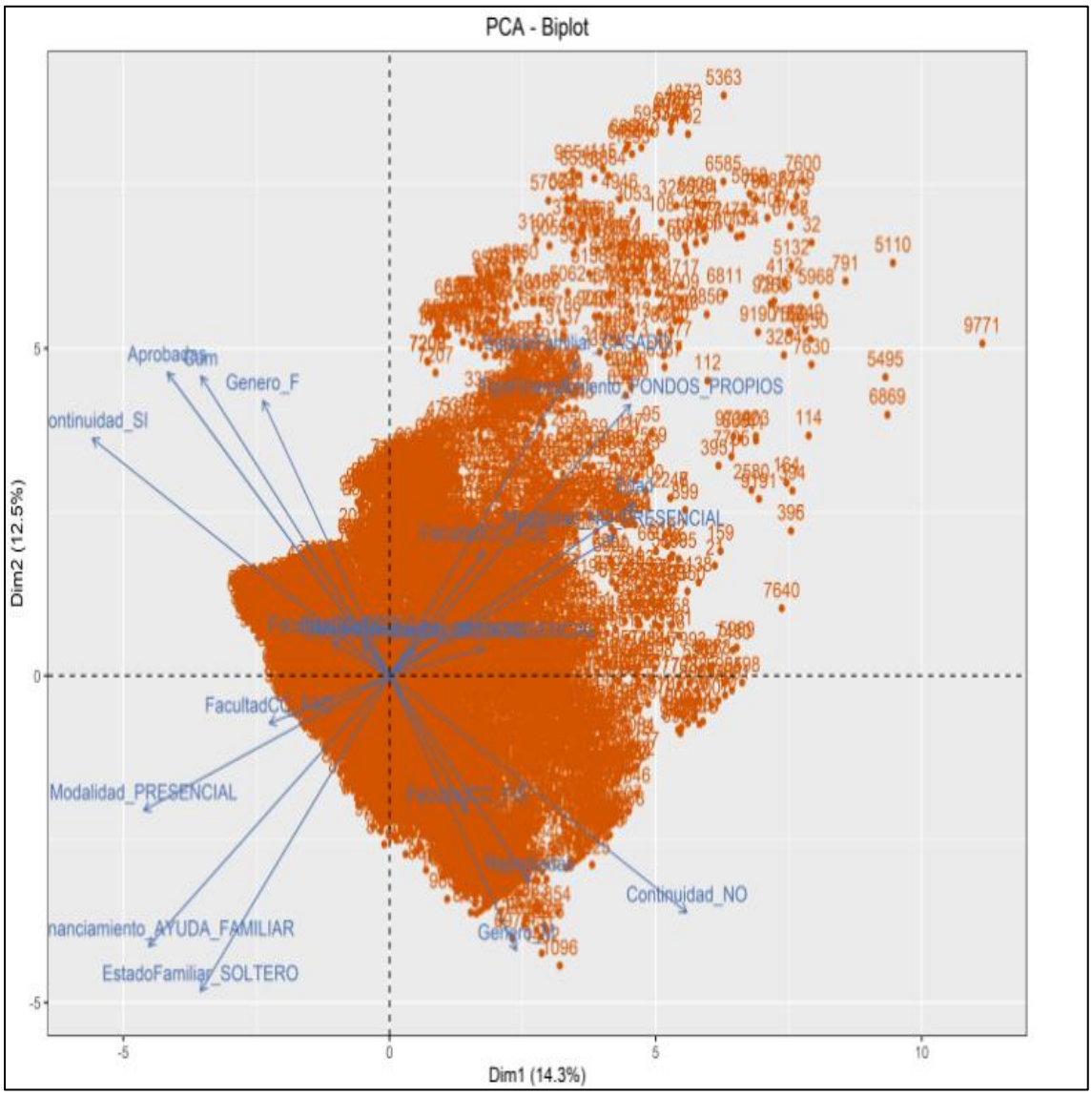
La gráfica de variables PCA ilustra.

- **Dim 1 (14.34%):** Representa la mayor variabilidad en los datos. En este eje, variables relacionadas con el **estado académico** (aprobadas/reprobadas), modalidad de estudio y la continuidad académica tienen mayor peso.

- **Dim 2 (12.49%):** Representa la segunda mayor variabilidad, influenciada principalmente por **variables sociodemográficas** como estado familiar, edad y tipo de financiamiento.
- Variables como "**Aprobadas", "Cum" y "Continuidad_SI"** se agrupan en la parte superior izquierda, lo que indica que están relacionadas positivamente con el éxito académico y la continuidad.
- Variables como "**Reprobadas" y "Continuidad_NO"** aparecen en el cuadrante inferior derecho, mostrando su correlación negativa con las anteriores.
- Variables de tipo de financiamiento y modalidad de estudio se dispersan entre los cuadrantes derecho e inferior, mostrando su impacto diferenciado en la variabilidad de los datos.
- La **modalidad de estudio** (presencial, semipresencial, no presencial) tiene una presencia significativa, destacándose como una variable crítica que puede influir en los clústeres formados.
- Las **variables de género** (Género_F y Género_M) muestran una orientación opuesta, lo que sugiere una posible segmentación de patrones entre hombres y mujeres.
- Variables **sociodemográficas** como "EstadoFamiliar_CASADO" y "EstadoFamiliar_SOLTERO" también tienen una distribución clave, indicando diferentes perfiles en el espacio PCA.

Aunque el PCA no realiza segmentación por sí mismo, en este estudio se utilizó para observar la disposición general de los datos. En la Figura 34, el biplot del análisis de componentes principales se muestra las variables originales (flechas azules) y las observaciones (puntos naranjas) proyectadas en las dos primeras dimensiones principales (**Dim 1** y **Dim 2**), que explican el porcentaje máximo posible en dos dimensiones del conjunto de datos.

Figura 34: Biplot del Análisis de Componentes Principales (PCA)



La Dim1 explica el 14.3% de la variabilidad total de los datos y Dim2 el 12.5%. Aunque estas cifras no suman un porcentaje muy alto, representan las direcciones principales de variación en los datos. Las observaciones están proyectadas en este plano para visualizar patrones importantes, aquí se representan prácticamente todas las observaciones del conjunto de datos. Esto genera una distribución densa, donde se ven claramente patrones globales, pero puede dificultar identificar detalles más finos. Las flechas representan las variables originales, indicando su contribución a

las componentes principales. Por ejemplo, variables como "Continuidad_SI", "Aprobadas", y "Modalidad_PRESENCIAL" tienen una fuerte relación con Dim1 y Dim2, mostrando cómo impactan la agrupación de los datos.

Aunque la gráfica es densa, parece haber un patrón de dispersión más amplio hacia la derecha (Dim1 positiva). Esto podría estar asociado con estudiantes con características específicas, como "Fondos propios" y "Modalidad presencial". La variabilidad en esta dirección sugiere que estas variables tienen un impacto significativo en la separación de los datos, la acumulación de puntos cerca del centro indica que muchos estudiantes comparten características similares o que estas no tienen una gran variabilidad respecto a las componentes principales. Las visualizaciones mostraron una estructura de datos compatible con la existencia de múltiples grupos lo cual permite realizar una exploración con distintos valores de grupos (k).

4.3.2. Análisis descriptivo por clúster

Esta sección tiene como propósito caracterizar los grupos obtenidos a partir de los procesos de segmentación no supervisada, mediante un análisis descriptivo de variables sociodemográficas, académicas y económicas. Para ello, se presentan los resultados de dos escenarios de agrupamiento: uno con **k = 2**, orientado a identificar una posible distinción binaria relacionada con la retención estudiantil, y otro con **k = 5**, que permite una segmentación más fina y detallada del estudiantado.

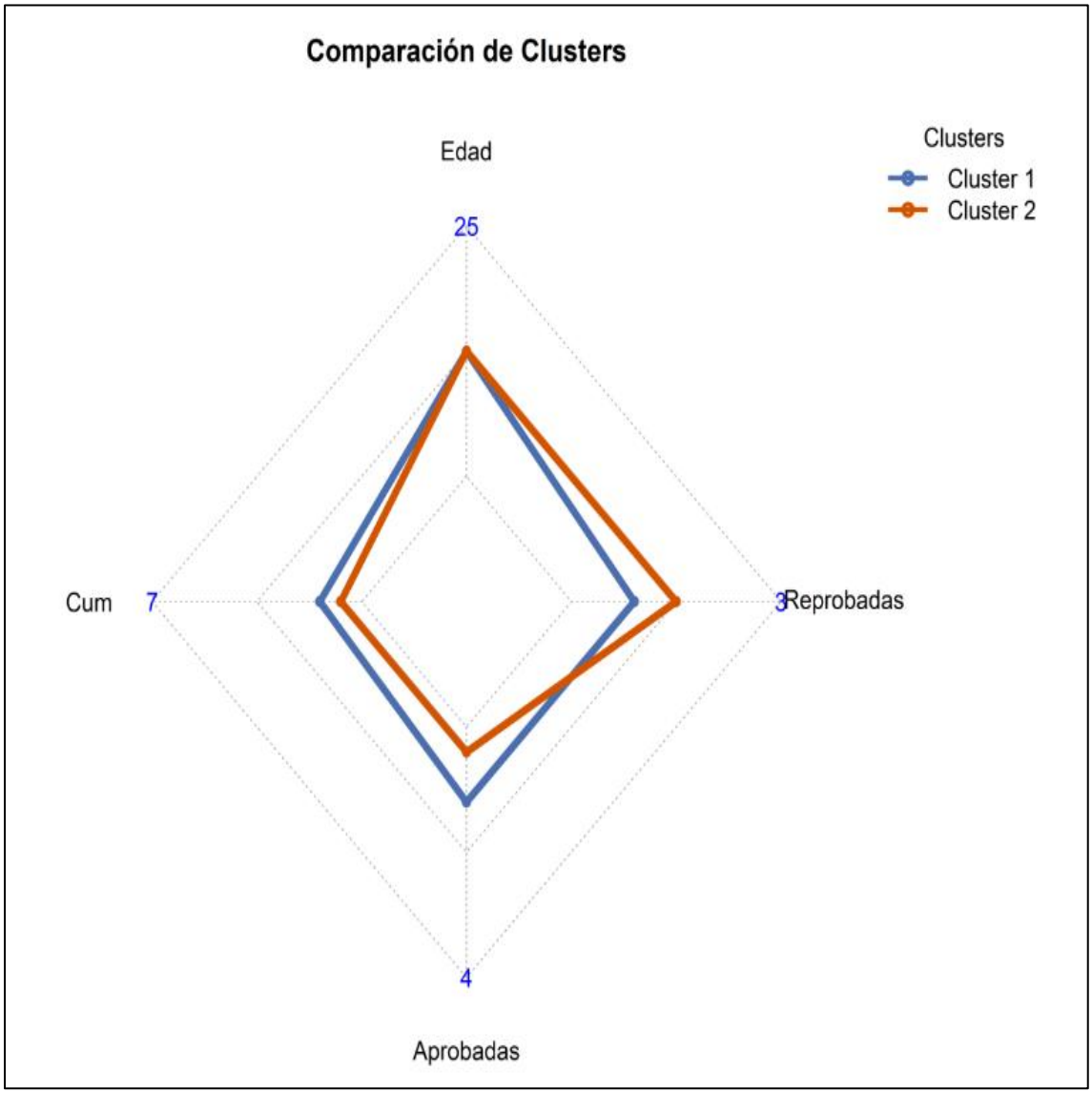
4.3.2.1. Clusterización k=2

Las variables cuantitativas por continuidad se refieren a cómo las variables numéricas continuas (como la edad, el promedio académico o los ingresos) pueden ser tratadas y analizadas en los modelos y técnicas de análisis, como los de

clusterización y predicción. Los siguientes gráficos representa que es especialmente importante para que los resultados sean más significativos y útiles.

El primer análisis de agrupamiento se realizó con k = 2, buscando identificar una división general entre estudiantes con características asociadas a la continuidad y aquellos con propensión a la deserción. La Figura 35, muestra la comparación de ambos grupos identificados mediante este enfoque.

Figura 35: Comparación de Clusters con K=2



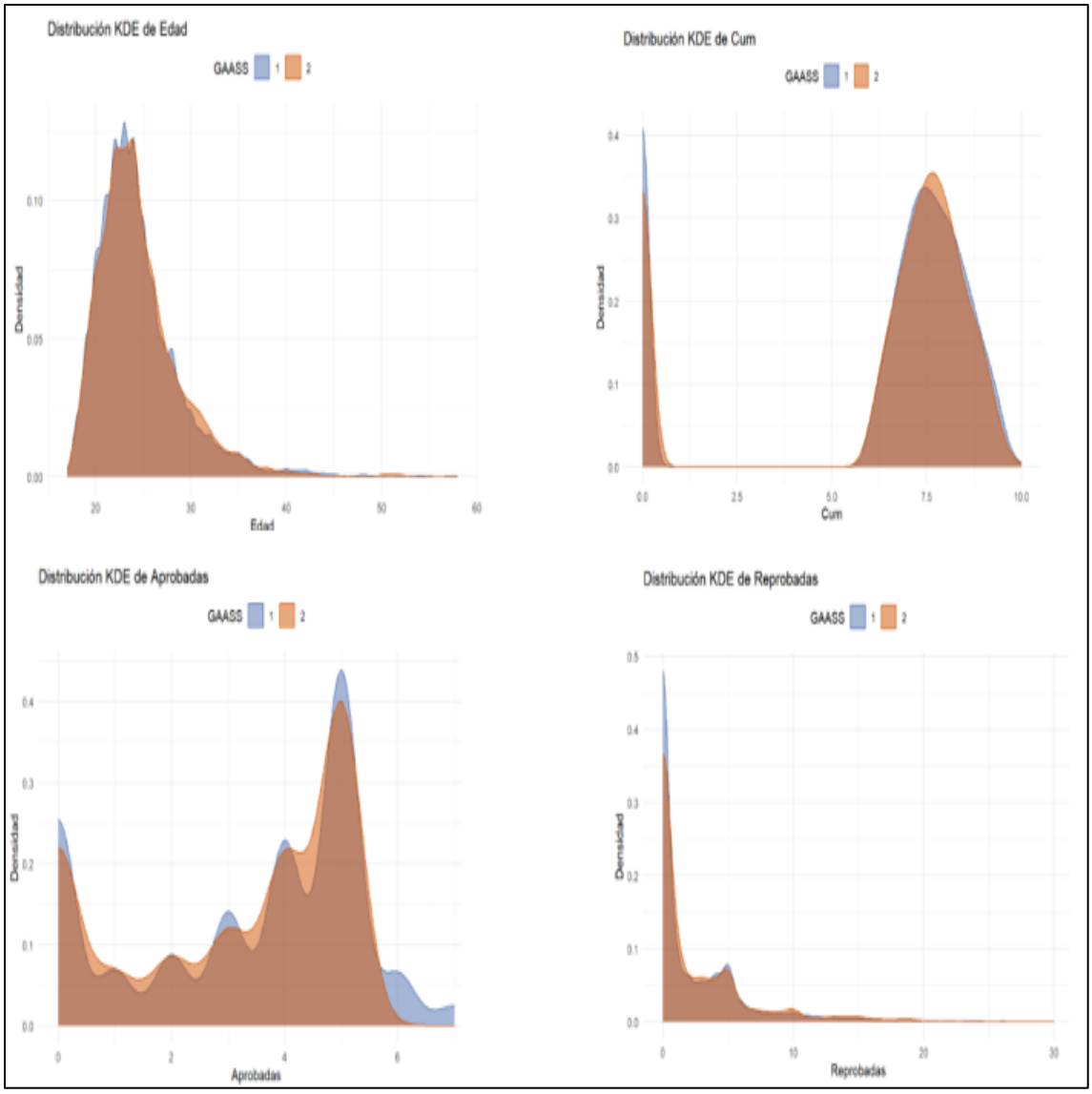
Al evaluar los resultados de los análisis de clusterización, es común encontrar casos donde los clústeres identificados presentan una alta similitud, tanto en términos de centroides como en las gráficas de distribución

- **El Cluster 1** parece agrupar a estudiantes más jóvenes con un rendimiento académico ligeramente más alto y menos materias reprobadas. Esto podría indicar un perfil más tradicional de estudiantes que ingresan directamente después de la escuela secundaria y enfrentan menos desafíos en términos de continuidad académica.
- **El Cluster 2** agrupa a estudiantes de mayor edad, con un balance entre más materias aprobadas, pero también un mayor promedio de reprobadas. Este perfil podría corresponder a estudiantes con mayores responsabilidades externas, como trabajo o familia, o aquellos que regresan a sus estudios después de una pausa.

La separación en dos clústeres permite entender las diferencias principales en el perfil estudiantil. La Figura 36, presenta la distribución de variables cuantitativas como el CUM, número de materias reprobadas y edad. Se observa que uno de los grupos tiene, en promedio, un mejor desempeño académico y menor número de asignaturas reprobadas, lo cual refuerza la hipótesis de una segmentación relacionada con la continuidad en los estudios.

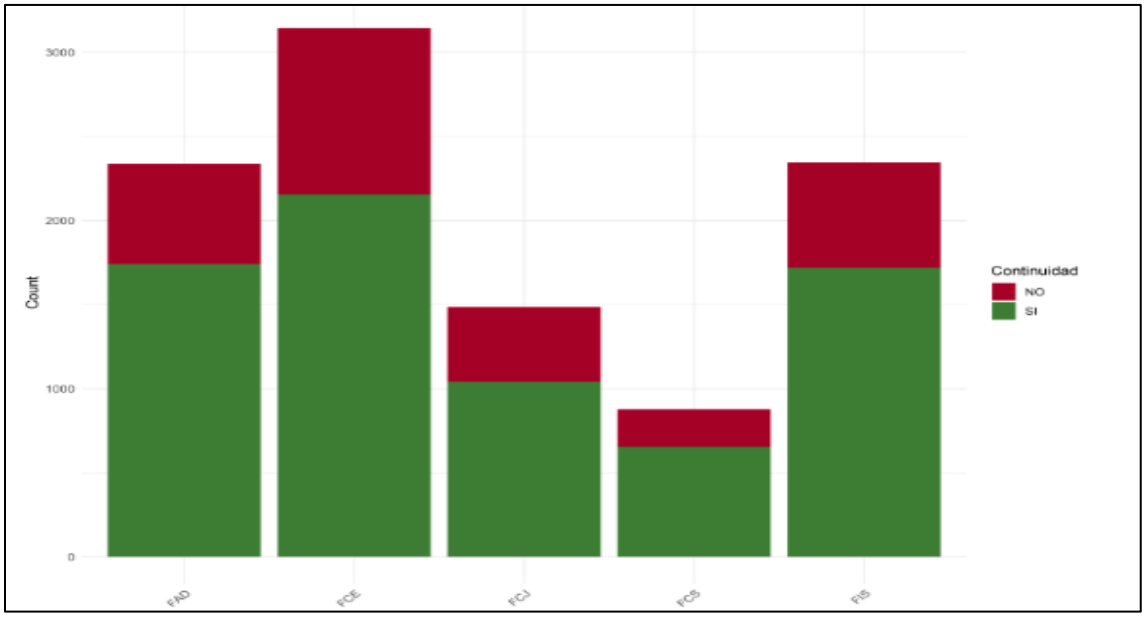
El Cluster 1 (perfil tradicional) muestra a estudiantes jóvenes con mejor rendimiento general y menor carga de reprobaciones, para este cluster se podrían ofrecer programas de mentorías para mantener su rendimiento académico.El Cluster 2 (perfil no tradicional) muestra estudiantes de mayor edad con mayor carga de reprobaciones, pero que avanzan en sus estudios, para este cluster, sería valioso implementar estrategias de apoyo académico y asesoramiento para gestionar mejor la carga académica y personal.

Figura 36: Distribución de variables cuantitativas



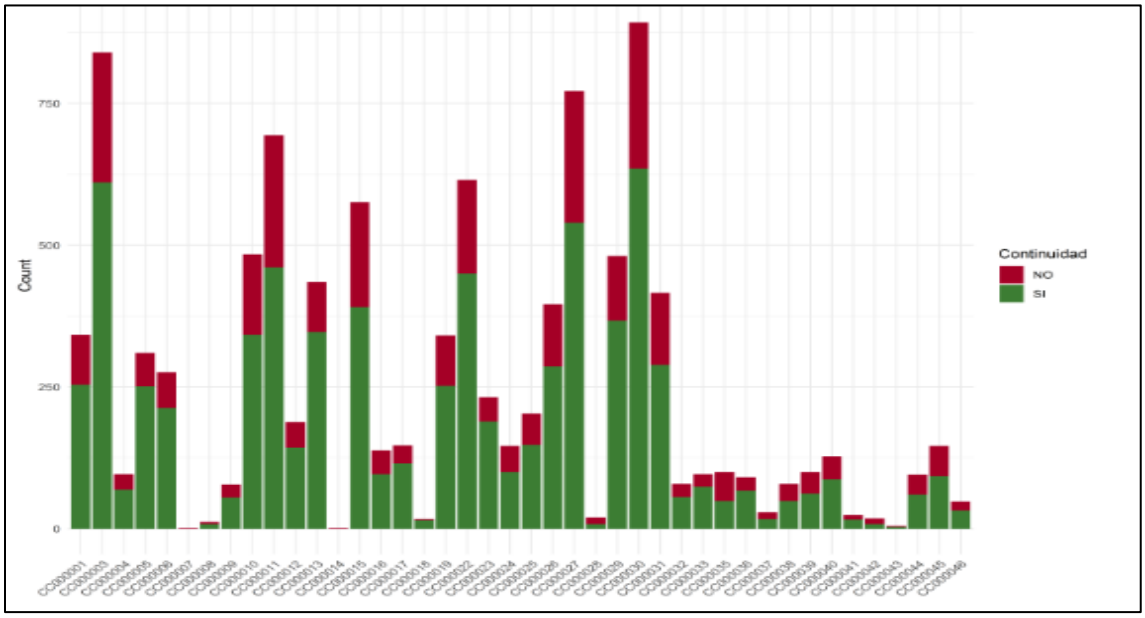
En cuanto a variables categóricas, la Figura 37, muestra la distribución de estudiantes por facultad. Se evidencia que ciertas facultades se concentran más en uno de los clústeres, lo que podría estar relacionado con particularidades propias de los programas académicos o contextos institucionales. Por otro lado, la Figura 38, evidencia la distribución por carrera, mostrando que algunas carreras específicas están sobrerrepresentadas en el grupo con mayor propensión a la deserción, lo cual merece un análisis posterior más detallado.

Figura 37: Clusters por facultad con K=2



En este análisis de clusterización, las variables cualitativas, como el género, la modalidad de estudio (presencial o en línea) o el tipo de financiamiento educativo, también pueden contribuir significativamente a la formación de clústeres.

Figura 38: Clusters de carreras con k=2

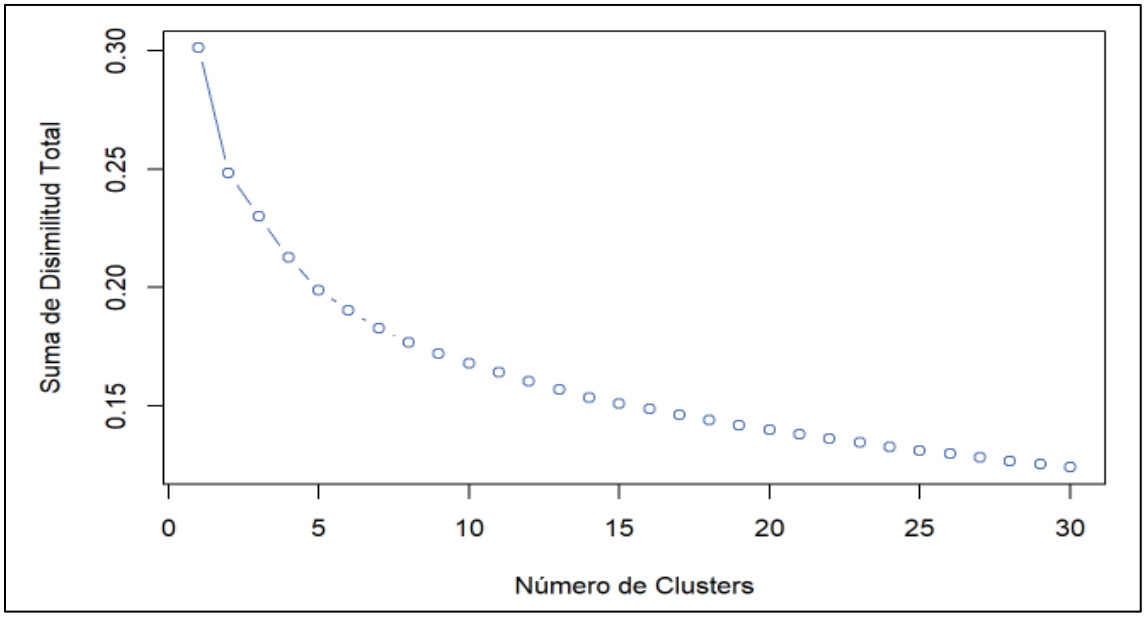


4.3.2.2. Método del Codo

Para determinar un valor óptimo de k, se aplicó el **método del codo**, como se muestra en la Figura 39. El gráfico evidencia una inflexión notable en k = 5, lo cual sugiere que este número de clústeres es adecuado para capturar la variabilidad de los datos sin sobreajustar el modelo.

En este método se evaluó la suma de las distancias cuadradas intra-clúster (inercia) en función del número de clústeres y ayuda a identificar el punto ideal para segmentar los datos sin sobresegmentarlos. Para ello, se partió de un conjunto de datos estandarizado (normalización de las variables) para garantizar que todas las características tuvieran la misma importancia en el cálculo de las distancias. Luego se ejecuto el algoritmo K-means iterativamente para un rango de valores de clústeres (k=1,2,..., 30). Después, para cada valor de k, se calculó la suma de las distancias cuadradas intra-clúster, este valor evalúa la homogeneidad de los clústeres formados. Finalmente, se seleccionó el punto donde la pendiente de la curva cambia significativamente (el "codo").

Figura 39: Método del Codo Jambu



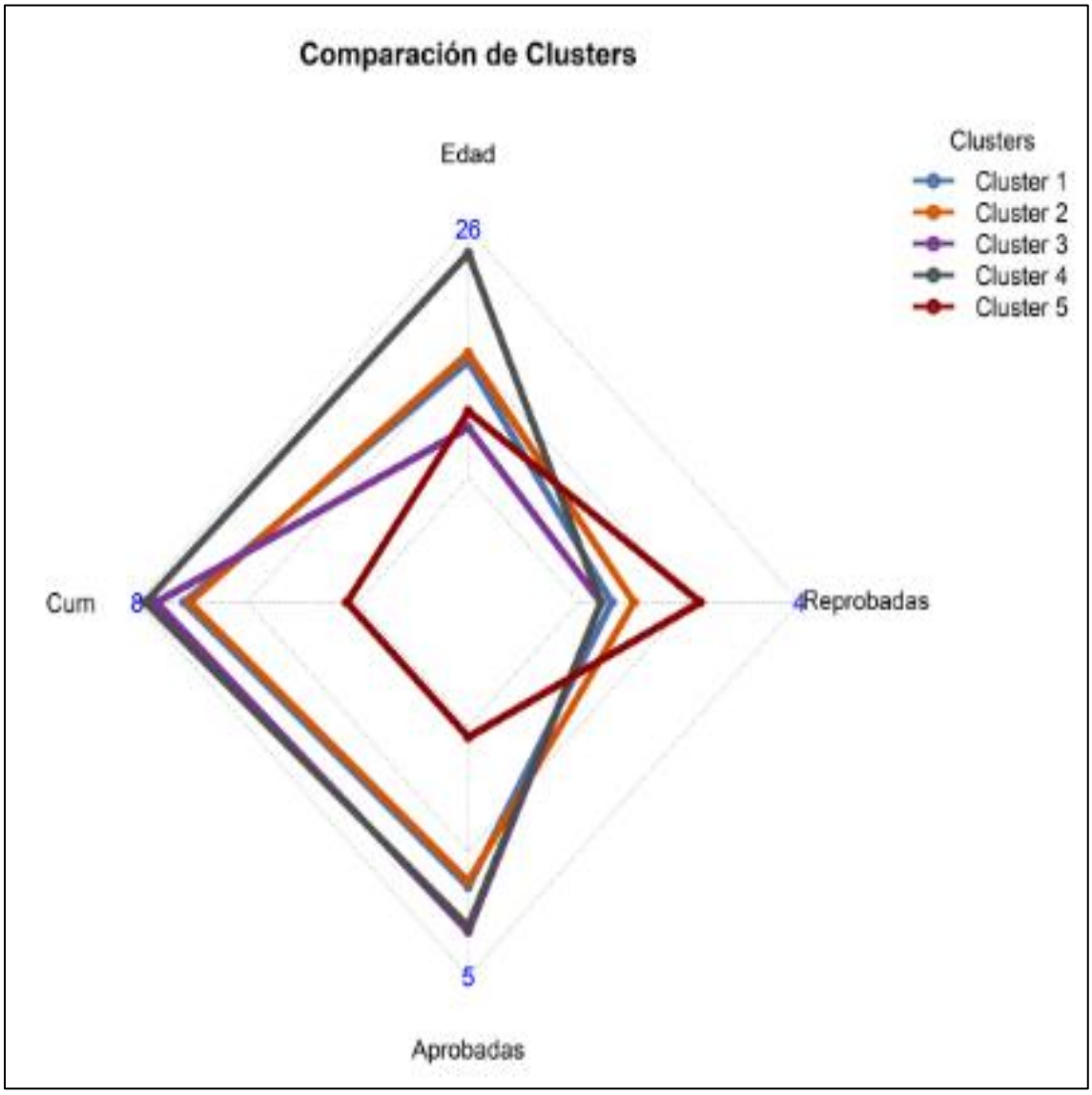
La inercia disminuye a medida que aumenta el número de clústeres. Esto ocurre porque, al aumentar los clústeres, cada grupo tiene menos puntos y las distancias promedio a sus centroides se reducen. El punto de inflexión donde la reducción en la inercia empieza a ser marginal (es decir, donde la curva comienza a aplanarse) es el codo. En este caso, el codo parece ubicarse alrededor de k = 5, lo que sugiere que dividir los datos en 5 clústeres es un buen compromiso entre simplicidad y efectividad, el elegir un número mayor de clústeres (por encima de k = 5) no ofrece una mejora significativa en la calidad de la segmentación y puede llevar a un modelo más complejo y difícil de interpretar.

4.3.2.3. Clusterización k=5

La agrupación con **k = 5** permite identificar perfiles estudiantiles más específicos. En la Figura 40, se visualizan los cinco grupos formados, mostrando una estructura más compleja en la distribución del estudiantado. La combinación de variables cualitativas y cuantitativas permite identificar subgrupos más representativos y

ofrece una interpretación más detallada de los factores que influyen en la retención estudiantil.

Figura 40: Cluster con k=5

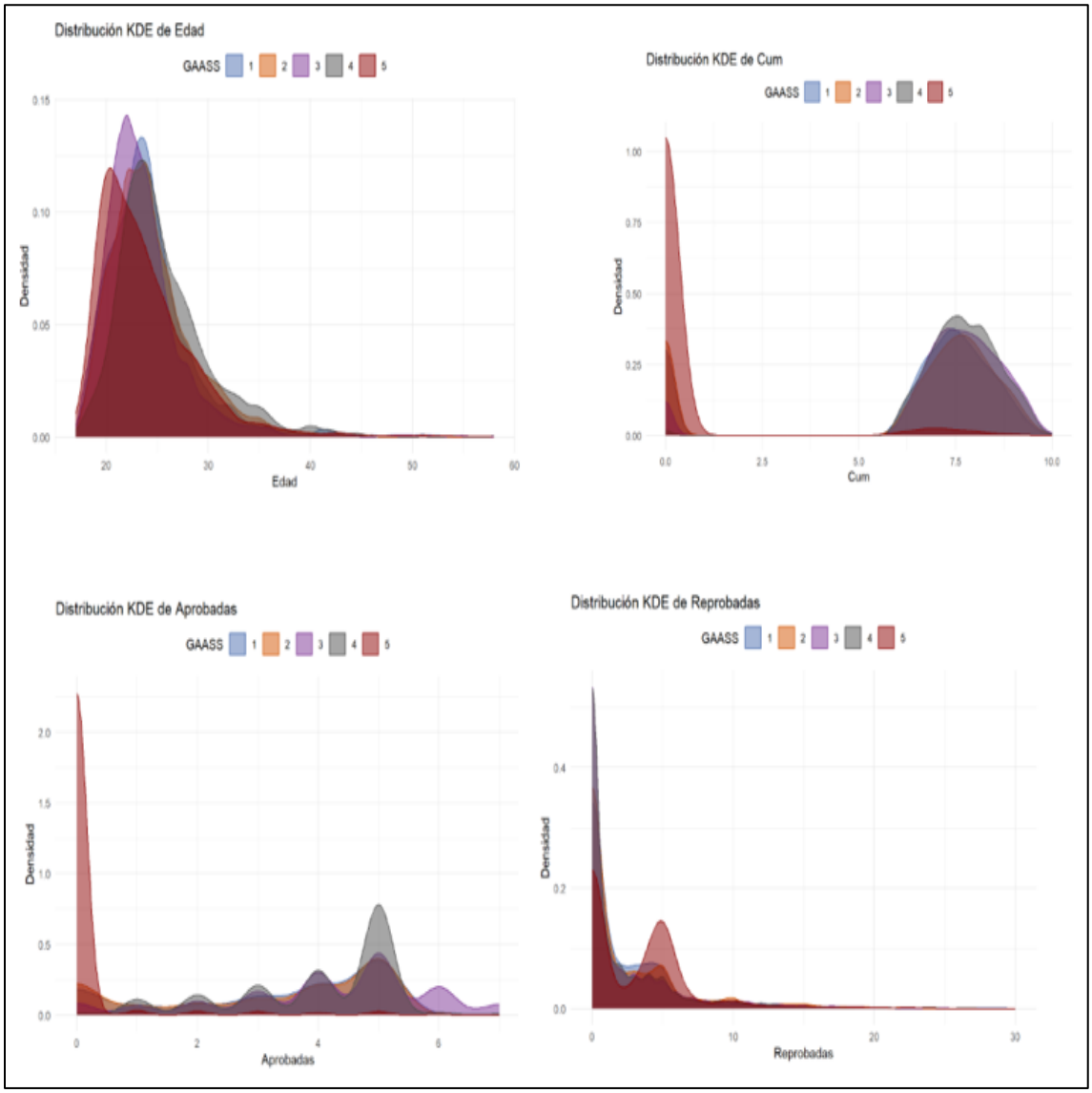


A partir de esta segmentación, se realizó un análisis por grupo de las variables más relevantes.

La Figura 41, ilustra la distribución de variables cualitativas clave, como género, modalidad, tipo de ingreso y sede. Se observa que algunos clústeres agrupan mayoritariamente a estudiantes mujeres, mientras que otros están dominados por

varones. Asimismo, se identifican diferencias en la proporción de estudiantes que ingresaron por modalidad presencial o virtual, así como en la forma de ingreso a la institución (ingreso directo, traslado, etc.).

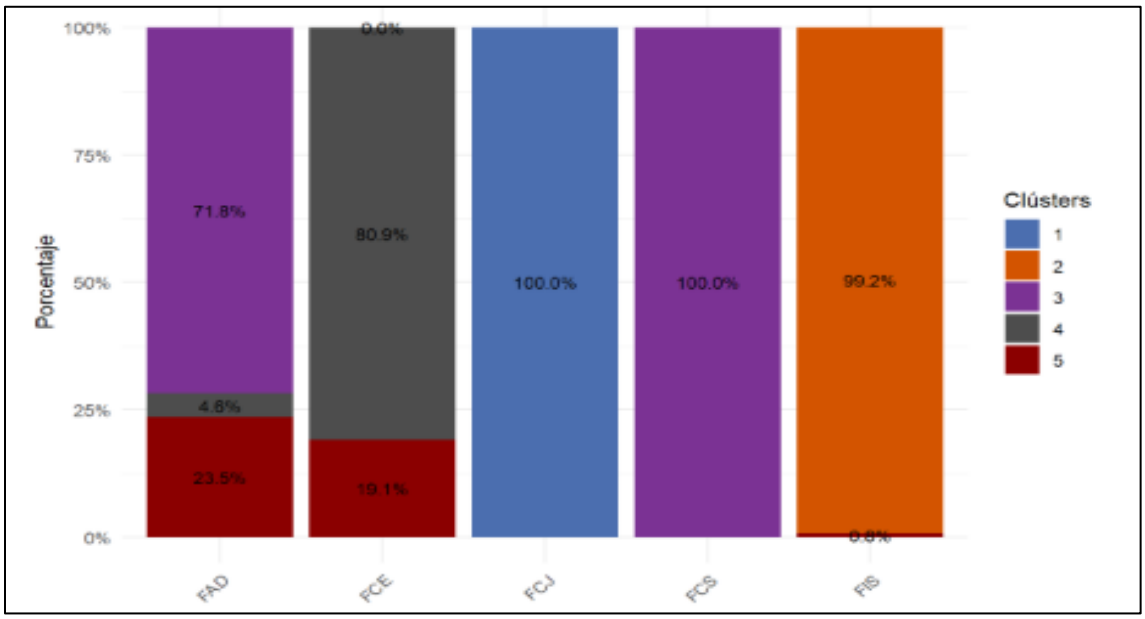
Figura 41: Distribución de variables cualitativas



La Figura 42 muestra la distribución porcentual de los cinco clústeres según la variable Facultad. El gráfico revela que algunos clústeres están fuertemente asociados a determinadas facultades, lo que sugiere la presencia de patrones

específicos en los perfiles estudiantiles según el contexto académico. Para la Facultad FAD, el 71.8% de los estudiantes se concentran en el Clúster 3, seguido por un 23.5% en el Clúster 5, lo que indica cierta heterogeneidad en los perfiles, aunque con predominancia del clúster 3. En la Facultad FCE se presenta la mayor diversidad de clústeres. El Clúster 4 agrupa al 80.9% de los estudiantes, seguido por el Clúster 5 (19.1%) y una pequeña proporción en el Clúster 2 (0.04%). Esta diversidad refleja perfiles académicos, económicos y sociodemográficos variados entre sus estudiantes. Todos los estudiantes de la facultad FCJ están agrupados únicamente en el Clúster 1, lo que sugiere una gran homogeneidad en los perfiles estudiantiles de estas facultades dentro del modelo de segmentación.

Figura 42: Cluster de facultad con k=5



4.4. Discusión del Análisis Exploratorio

Los resultados de esta capitulo se deben a varios factores de los cuales la decisión de aplicar una clusterización de nivel 1 responde a la necesidad de segmentar la población estudiantil en grupos homogéneos que compartan características

similares, sin requerir información previa sobre etiquetas o categorías. Esto es crucial en el contexto de la retención estudiantil, ya que permite identificar patrones ocultos entre los estudiantes y facilita el diseño de estrategias personalizadas para cada grupo. Este enfoque inicial proporciona una base exploratoria sólida, asegurando que las características más relevantes sean consideradas antes de profundizar en modelos más específicos.

La clusterización inicial con k = 2 permitió una separación básica de la población estudiantil en dos grandes grupos. Este enfoque fue útil para identificar diferencias marcadas en los datos, como la presencia de estudiantes con mayor riesgo de deserción en contraste con aquellos con mayor probabilidad de retenerse. Aunque la simplicidad de k = 2 ofrece una visión general, sus limitaciones radican en la falta de granularidad, lo que llevó a explorar valores más altos de k para captar mejor la diversidad dentro de los datos.

Se utilizó el método del codo para determinar el número óptimo de clústeres. Este método analiza la variación explicada por los clústeres en función de k y encuentra el punto donde la ganancia marginal disminuye considerablemente. En este caso, el análisis reveló que k = 5 era el valor ideal, ya que capturaba una gran proporción de la variabilidad de los datos sin sobreajustarse, equilibrando la complejidad del modelo y la interpretabilidad.

La clusterización con k = 5 permitió segmentar a los estudiantes en grupos más específicos, destacando combinaciones particulares de características demográficas, socioeconómicas y académicas. Por ejemplo, se identificaron clústeres asociados con estudiantes de modalidad presencial con financiamiento propio, así como aquellos que recibían apoyo familiar. Esta segmentación detallada ofrece una base valiosa para diseñar intervenciones focalizadas que atiendan las necesidades únicas de cada grupo.

El PCA fue fundamental para reducir la dimensionalidad de los datos y facilitar la visualización de los clústeres. Al transformar las variables originales en

componentes principales, se logró capturar la mayor parte de la variabilidad en un espacio de menor dimensión. Esto no solo mejoró la eficiencia del proceso de clusterización, sino que también permitió observar la separación y cohesión entre los grupos en gráficos bidimensionales. El PCA, además, destacó qué variables eran más relevantes en la formación de los clústeres.

Los resultados proporcionan una visión profunda de las características y patrones asociados con la retención estudiantil. La clusterización y el PCA han sentado las bases para identificar grupos de estudiantes con perfiles específicos.

4.5. Conclusión del Análisis Exploratorio

El análisis realizado en este capítulo permitió segmentar a los estudiantes en grupos con características similares, lo que facilita la identificación de factores que influyen en la retención y la deserción. La elección de la clusterización como primer nivel de análisis se justificó en la necesidad de comprender patrones ocultos en los datos sin una etiqueta previa de abandono, permitiendo así una exploración más objetiva de las características compartidas por los estudiantes.

Los resultados obtenidos con k=2 mostraron una segmentación inicial que diferenciaba claramente a los estudiantes con mayores probabilidades de continuidad. Con el método del codo, se consideró que k=5 era el número óptimo de clústeres, lo que permitió una segmentación más detallada y significativa en términos de desempeño académico, financiamiento, modalidad de estudio y factores socioeconómicos.

El uso del PCA fue clave en este capítulo de análisis exploratorio, ya que las agrupaciones obtenidas reflejan perfiles diferenciados de estudiantes, lo que brinda información valiosa para la toma de decisiones institucionales. Por ejemplo, se identifican grupos con alta tasa de retención, asociados a buenos desempeños académicos y apoyo financiero familiar, así como grupos con mayor riesgo de

abandono, caracterizados por Múltiples asignaturas reprobadas y financiamiento con fondos propios. Estos hallazgos pueden servir de base para diseñar estrategias específicas de intervención, como tutorías personalizadas, apoyo financiero focalizado o cambios en la política de acompañamiento académico.

Este proceso de clusterización ha sido esencial para comprender la diversidad de perfiles dentro de la población estudiantil y establecer una primera aproximación al fenómeno de la deserción. Sin embargo, para lograr una mejor predicción del riesgo individual de abandono, es necesario complementar este enfoque con modelos de aprendizaje supervisado.

Capítulo 5: Análisis Predictivo

5.1. Introducción

En la educación superior, la retención estudiantil se ha consolidado como un indicador crítico de calidad académica, eficiencia institucional y equidad social. La creciente preocupación por las tasas de deserción ha impulsado a las universidades a buscar soluciones basadas en evidencia que les permitan intervenir de forma oportuna y eficaz. En este contexto, el análisis predictivo al combinar técnicas avanzadas de aprendizaje automático y procesamiento de datos se ha convertido en una herramienta fundamental para comprender los factores asociados a la retención y anticipar el comportamiento de los estudiantes, identificar a aquellos en riesgo de abandono y, en consecuencia, optimizar las estrategias de acompañamiento académico y optimizar los recursos institucionales para mejorar los resultados educativos y sociales.

Este capítulo tiene como objetivo aplicar y comparar diversos modelos de aprendizaje automático supervisado para la predicción de la retención estudiantil, utilizando un conjunto de datos históricos de una universidad privada de El Salvador. Para ello, se implementaron modelos como Máquinas de Soporte Vectorial (SVM), K-Vecinos más Cercanos (KNN), Bosques Aleatorios (Random Forest), AdaBoost y XGBoost, seleccionados por su capacidad para abordar problemas de clasificación binaria y su eficacia documentada en contextos educativos. Cada modelo fue optimizado mediante técnicas de ajuste de hiperparámetros y evaluado a través de métricas estándar de clasificación, incluyendo precisión, recall, F1-score, y el área bajo la curva ROC (AUC-ROC).

5.2. Metodología

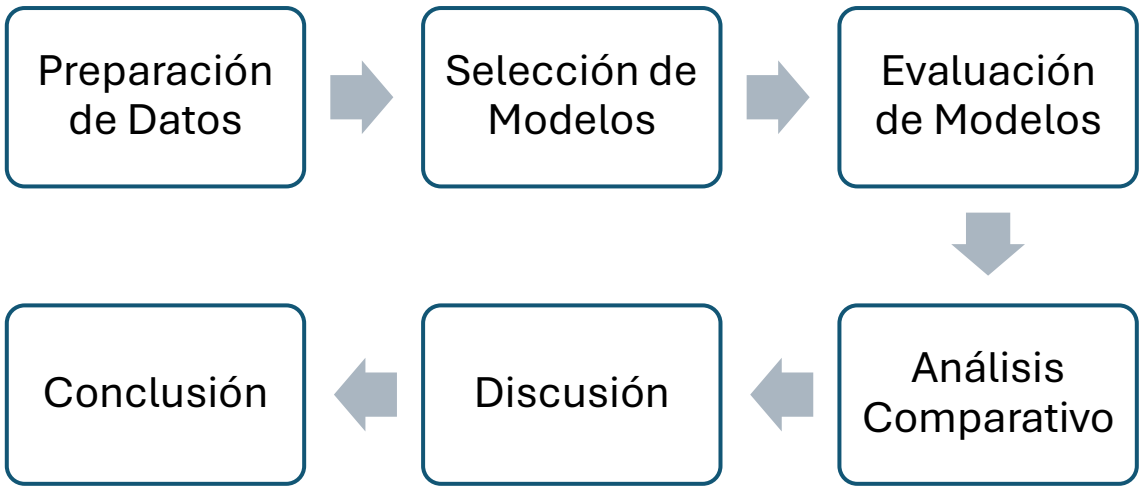
El conjunto de datos utilizado y los procedimientos aplicados en esta etapa de preprocesamiento, son fundamentales para garantizar la calidad del análisis predictivo. Se busca proporcionar una base sólida para la implementación de los modelos de aprendizaje automático.

5.2.1. Descripción del conjunto de datos

La base de datos utilizada en este análisis proviene de registros históricos de estudiantes de una universidad privada en El Salvador, y contiene información detallada sobre características demográficas, académicas y administrativas de los estudiantes. Este conjunto de datos incluye un total de 83,175 filas y 31 columnas, entre las cuales se encuentran variables categóricas y numéricas relevantes para el análisis de retención estudiantil.

Para garantizar la calidad y relevancia del análisis predictivo, se llevó a cabo un proceso de limpieza y transformación de los datos. Este proceso se desarrolló en varias etapas

Figura 43: Esquema de las etapas del análisis predictivo



5.2.2. Técnicas de preprocesamiento

El preprocesamiento de los datos se llevó a cabo en varias etapas para asegurar que la información estuviera en un formato adecuado para el entrenamiento de los modelos de aprendizaje automático. A continuación, se detallan las principales técnicas aplicadas:

1.- Conversión de variables categóricas. Todas las variables categóricas fueron transformadas en variables numéricas utilizando codificación tipo *label encoding* y *one-hot encoding* según el algoritmo a aplicar. Por ejemplo, variables como Modalidad, Facultad, TipoIngreso, entre otras, fueron convertidas en variables dummy para evitar introducir relaciones ordinales inexistentes y permitir el uso adecuado por parte de los modelos.

2.- Normalización de variables numéricas. Las variables numéricas, como Edad, Cum, Aprobadas, Reprobadas, fueron estandarizadas utilizando la técnica de **escalado Z-score** (media cero y desviación estándar uno). Esta normalización fue crucial especialmente para modelos sensibles a las escalas, como KNN y SVM.

3.- Tratamiento de valores perdidos. Se identificaron valores nulos en algunas variables, los cuales fueron tratados mediante **imputación por la mediana** en el caso de variables numéricas y **imputación por la moda** en variables categóricas. Esta estrategia buscó minimizar el sesgo y preservar la distribución original de los datos.

4.- Detección y manejo de valores atípicos. Se realizó un análisis exploratorio para identificar valores atípicos extremos, los cuales fueron inspeccionados caso por caso. En situaciones donde los valores anómalos se debían a errores de captura, fueron

corregidos o eliminados; en otros casos, se conservaron por considerarse parte de la variabilidad natural del fenómeno.

5.- División de datos. Finalmente, el conjunto de datos fue dividido en dos subconjuntos: un 75% para entrenamiento y un 25% para prueba, manteniendo la proporción original de clases mediante *stratified sampling*, a fin de asegurar una representación equitativa de las categorías de la variable objetivo en ambos subconjuntos.

5.2.3. Técnicas predictivas utilizadas

Para la predicción de la retención estudiantil se seleccionaron cinco modelos de aprendizaje automático supervisado, considerando su eficacia comprobada en tareas de clasificación binaria y su capacidad para manejar conjuntos de datos con características mixtas, relaciones no lineales y proporciones desbalanceadas. La selección de estos modelos también estuvo motivada por su robustez, interpretabilidad relativa y desempeño reportado en literatura reciente dentro del ámbito educativo y otros dominios similares (Chen & Guestrin, 2016; Fernández-Delgado y otros, 2014).

Máquinas de Vectores de Soporte (SVM). Este algoritmo fue implementado debido a su capacidad para encontrar el hiperplano óptimo que separa las clases, maximizando el margen entre ellas. Se utilizó una función kernel para capturar relaciones no lineales entre las variables. Su uso está especialmente justificado en entornos con alta dimensionalidad, ya que logra un buen equilibrio entre complejidad del modelo y generalización (Peng y otros, 2020). Fue seleccionado para evaluar su desempeño en la predicción de la variable objetivo *Continuidad*. Este modelo es ampliamente utilizado en problemas de clasificación debido a su

capacidad para encontrar un hiperplano óptimo que maximiza la separación entre clases en espacios de alta dimensión.

K-Vecinos Más Cercanos (KNN). El modelo KNN se basó en la similitud entre instancias, utilizando la distancia euclidiana como métrica principal. Se evaluaron distintos valores de *k*, seleccionando el valor óptimo mediante validación cruzada. A pesar de su simplicidad, KNN puede ser muy competitivo cuando se trabaja con datos bien normalizados y con una distribución representativa (Qingtao y otros, 2019). En este capítulo, KNN fue implementado utilizando diferentes estrategias de búsqueda de vecinos más cercanos, incluyendo **ball_tree, kd_tree y brute**, con el objetivo de evaluar su impacto en la precisión del modelo.

Bosques Aleatorios (Random Forest). Este método de ensamblado fue elegido por su capacidad para manejar relaciones complejas y variables correlacionadas. Se entrenó un conjunto de 100 árboles de decisión, aplicando muestreo bootstrap y selección aleatoria de características en cada nodo. Random Forest ofrece una buena tolerancia al sobreajuste y permite estimar la importancia de las variables, lo cual resulta valioso para la interpretación del modelo. Se implementó Random Forest para la predicción de la retención estudiantil, ajustando hiperparámetros clave y evaluando su rendimiento mediante diversas métricas.

AdaBoost (Adaptive Boosting). Se aplicó esta técnica de boosting que combina clasificadores débiles (árboles poco profundos) de manera secuencial, aumentando el peso de las observaciones mal clasificadas en cada iteración. AdaBoost se destaca por mejorar modelos simples y ofrecer un buen rendimiento en escenarios con ruido moderado (Schapire & Freund, 2012). Se implementó evaluando su desempeño en la predicción de la retención estudiantil.

XGBoost (Extreme Gradient Boosting). Finalmente, se implementó XGBoost, una técnica avanzada de boosting que incorpora optimización de la regularización, poda de árboles y paralelización. Se configuró con aprendizaje por árbol secuencial y validación interna para el ajuste de hiperparámetros. Este modelo es reconocido por su rendimiento sobresaliente en tareas estructuradas y es ampliamente utilizado en aplicaciones industriales y académicas (Tianqi & Guestrin, 2016). Este modelo se utilizó para predecir la retención estudiantil, dada su alta precisión y capacidad de generalización.

5.2.4. Evaluación del rendimiento

Para evaluar la calidad de las predicciones de los modelos, se utilizaron múltiples métricas de rendimiento, adecuadas para problemas de clasificación binaria con datos desbalanceados. Estas métricas permiten obtener una visión integral del desempeño de los modelos más allá de la simple precisión.

Accuracy (Exactitud). Representa el porcentaje de predicciones correctas sobre el total de observaciones.

Precision (Precisión positiva). Mide la proporción de verdaderos positivos entre todas las predicciones positivas.

Recall (Sensibilidad). Indica la proporción de verdaderos positivos entre todos los casos reales positivos. En este estudio, representa la capacidad del modelo para identificar correctamente a los estudiantes que continuarán sus estudios.

Error Global: Complemento de la precisión global, representando la tasa de errores cometidos por el modelo.

Área bajo la curva ROC (ROC-AUC). Evalúa la capacidad del modelo para distinguir entre las clases en todos los umbrales posibles. Un valor cercano a 1 indica excelente capacidad discriminativa. La curva ROC complementa la matriz de

confusión al mostrar la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos.

Matriz de confusión. Se utilizó para visualizar el número de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos de cada modelo. Esta herramienta permitió entender de forma detallada los errores de clasificación cometidos.

5.3. Resultados de la Implementación de Modelos Predictivos

La implementación y evaluación de los modelos predictivos está orientada a optimizar la predicción de la retención estudiantil. Para ello, se hace una descripción de los algoritmos aplicados, el proceso de ajuste de hiperparámetros, los resultados obtenidos y un análisis comparativo para identificar los enfoques más efectivos.

5.3.1. Modelo SVM Optimizado

Para optimizar el desempeño del modelo SVM, se llevó a cabo un proceso de experimentación con diferentes valores de los hiperparámetros **tipo de núcleo (kernel)** y **parámetro de regularización (C)**. La implementación se realizó mediante un enfoque de búsqueda en malla, evaluando las siguientes configuraciones:

- **Núcleos evaluados**: rbf, poly, linear y sigmoid.
- **Valores de C probados**: un rango de valores de regularización entre **1 y 50**.

Cada combinación de hiperparámetros fue entrenada y evaluada utilizando la metodología de **validación cruzada** para reducir el riesgo de sobreajuste. Posteriormente, se seleccionaron los modelos con los mejores desempeños basados en diversas métricas de evaluación.

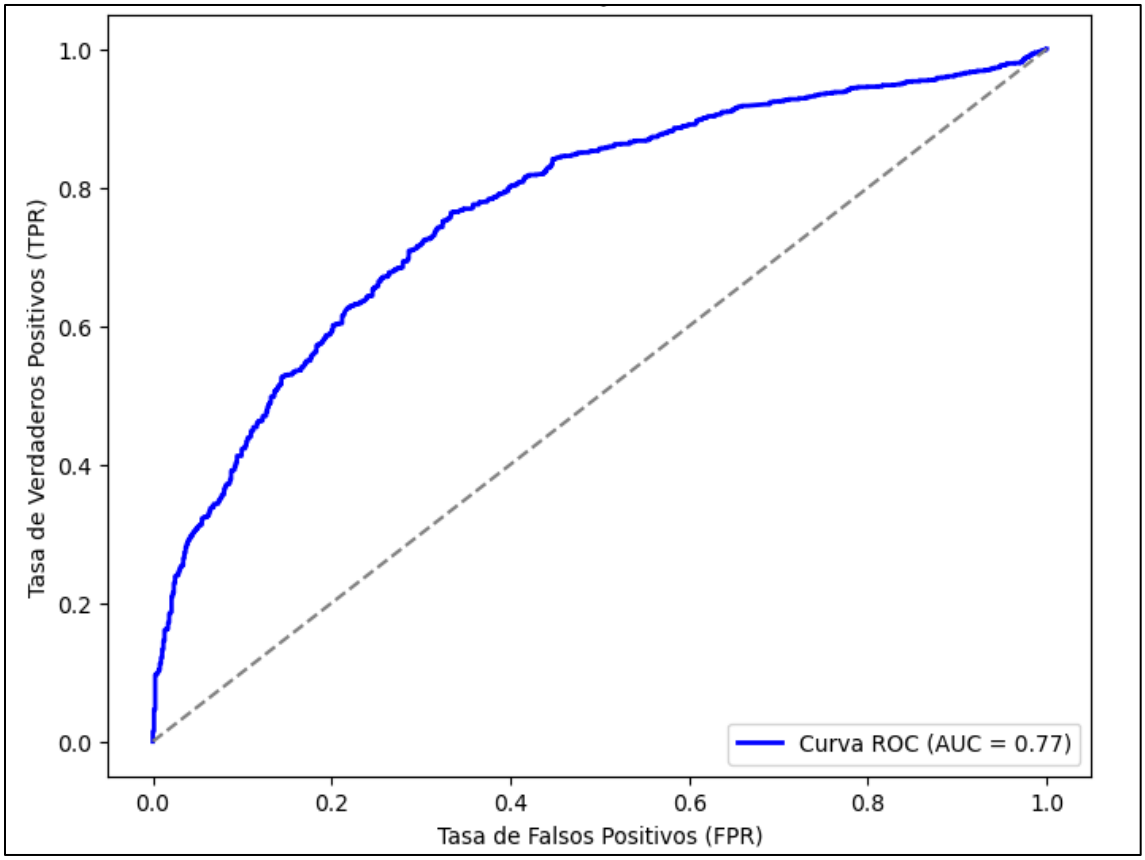
Los resultados obtenidos se presentan en la Figura 46, donde se incluyen la matriz de confusión, la precisión global, el error global, la precisión positiva (PP) y la precisión negativa (PN) para cada configuración. El modelo con C=46 alcanzó la mayor precisión global (77.31 %) y un error global del 22.69 %, mientras que la configuración con C=1 presentó la mayor precisión positiva (94.56 %), lo que podría ser relevante en casos donde identificar correctamente la clase positiva sea prioritario. Por otro lado, las configuraciones con C=46 y C=37 obtuvieron valores similares en precisión negativa (39.48 %), lo que refleja un desempeño homogéneo en la identificación de la clase negativa. Estos resultados destacan la importancia de analizar las métricas específicas de cada clase, además de la precisión global, al seleccionar el modelo más adecuado para la tarea.

Figura 44: Resultados de desempeño del modelo SVM con kernel RBF para diferentes valores del parámetro de regularización C

Modelo	Matriz de Confusión	Precisión Global	Error Global	Precisión Positiva (PP)	Precisión Negativa (PN)
kernel=rbf_C=46	[[302 463]\n [115 1667]]	0.773066	0.226934	0.935466	0.394771
kernel=rbf_C=1	[[259 506]\n [97 1685]]	0.763251	0.236749	0.945567	0.338562
kernel=rbf_C=37	[[302 463]\n [117 1665]]	0.772281	0.227719	0.934343	0.394771

En la Figura 45, muestra que el modelo SVM con kernel = rbf y C = 46, genera un AUC de 0.77. Esto sugiere que puede ser una herramienta útil para predecir la retención estudiantil y aplicar intervenciones tempranas. Ya que hay un 77% de probabilidad de que el modelo SVM asigne una mayor probabilidad de retención a un estudiante realmente retenido que a uno que no lo fue.

Figura 45: Curva ROC del Mejor Modelo SVM para la Predicción de Retención Estudiantil



5.3.2. Método de K vecinos más cercanos

Para la implementación del modelo KNN, primero se realizó una conversión **de variables categóricas** a variables numéricas mediante codificación. Luego se realizó una n**ormalización de variables numéricas** para garantizar que todas las características estuvieran en la misma escala, evitando que atributos con valores más altos tuvieran mayor peso en la distancia euclidiana. Después, se hizo una d**ivisión de los datos** en conjuntos de entrenamiento (80%) y prueba (20%) utilizando una estrategia de muestreo estratificado para mantener la proporción de clases. Finalmente se realizó la s**elección del número óptimo de vecinos (K)** mediante validación cruzada con **k-folds=10**, evaluando valores de K en el rango de 1 a 50.

5.3.2.1. Modelo k_algorithm = "ball_tree"

Se evaluó el desempeño del modelo de K vecinos más cercanos (KNN) utilizando el algoritmo **ball_tree** con diferentes valores del parámetro k (número de vecinos). Los resultados detallados se presentan en la Figura 46, incluyendo la matriz de confusión, la precisión global, el error global, la precisión positiva (PP) y la precisión negativa (PN) para las configuraciones más destacadas. El modelo con k=9 alcanzó una precisión global de 75.74 % y un error global del 24.26 %, presentando además una alta precisión positiva (91.13 %), lo que lo hace adecuado para la identificación de la clase positiva. En comparación, la configuración con k=2 obtuvo una precisión global menor (66.47 %) y una precisión positiva de 67.00 %, aunque destacó en precisión negativa con un valor de 65.23 %. Estos resultados reflejan cómo el valor de k influye significativamente en el balance entre las métricas de desempeño, permitiendo seleccionar la configuración más adecuada según las prioridades del análisis.

Figura 46: Resultados de desempeño del modelo de K vecinos más cercanos (KNN) utilizando el algoritmo ball_tree para diferentes valores de k

Modelo	Matriz de Confusión	Precisión Global	Error Global	Precisión Positiva (PP)	Precisión Negativa (PN)
algorithm=ball_tree_n_neighbors=9	[[305 460]\n [158 1624]]	0.757362	0.242638	0.911336	0.398693
algorithm=ball_tree_n_neighbors=2	[[499 266]\n [588 1194]]	0.664704	0.335296	0.670034	0.652288

5.3.2.2. Modelo k_algorithm = "kd_tree"

Se evaluó el desempeño del modelo de K vecinos más cercanos (KNN) utilizando el algoritmo **kd_tree** con diferentes valores del parámetro k (número de vecinos). Los resultados obtenidos, resumidos en la Figura 47, incluyen la matriz de confusión, la precisión global, el error global, la precisión positiva (PP) y la precisión negativa (PN) para las configuraciones más relevantes. La configuración con k=9 presentó la mayor precisión global (75.74 %) y un error global del 24.26 %, destacando además

por su alta precisión positiva (91.13 %), lo que la hace especialmente útil para la identificación de la clase positiva. En contraste, el modelo con k=2 obtuvo una precisión global menor (66.47 %), pero presentó un mejor balance entre precisión positiva (67.00 %) y precisión negativa (65.23 %). Estos resultados muestran la influencia del valor de k y del algoritmo utilizado en el desempeño del modelo, permitiendo seleccionar la configuración más adecuada en función de los objetivos del análisis

Figura 47: Resultados de desempeño del modelo de K vecinos más cercanos (KNN) utilizando el algoritmo kd_tree para diferentes valores de k

Modelo	Matriz de Confusión	Precisión Global	Error Global	Precisión Positiva (PP)	Precisión Negativa (PN)
algorithm=kd_tree_n_neighbors=9	[[305 460]\n [158 1624]]	0.757362	0.242638	0.911336	0.398693
algorithm=kd_tree_n_neighbors=2	[[499 266]\n [588 1194]]	0.664704	0.335296	0.670034	0.652288

5.3.2.3. Modelo k_algorithm = "brute"

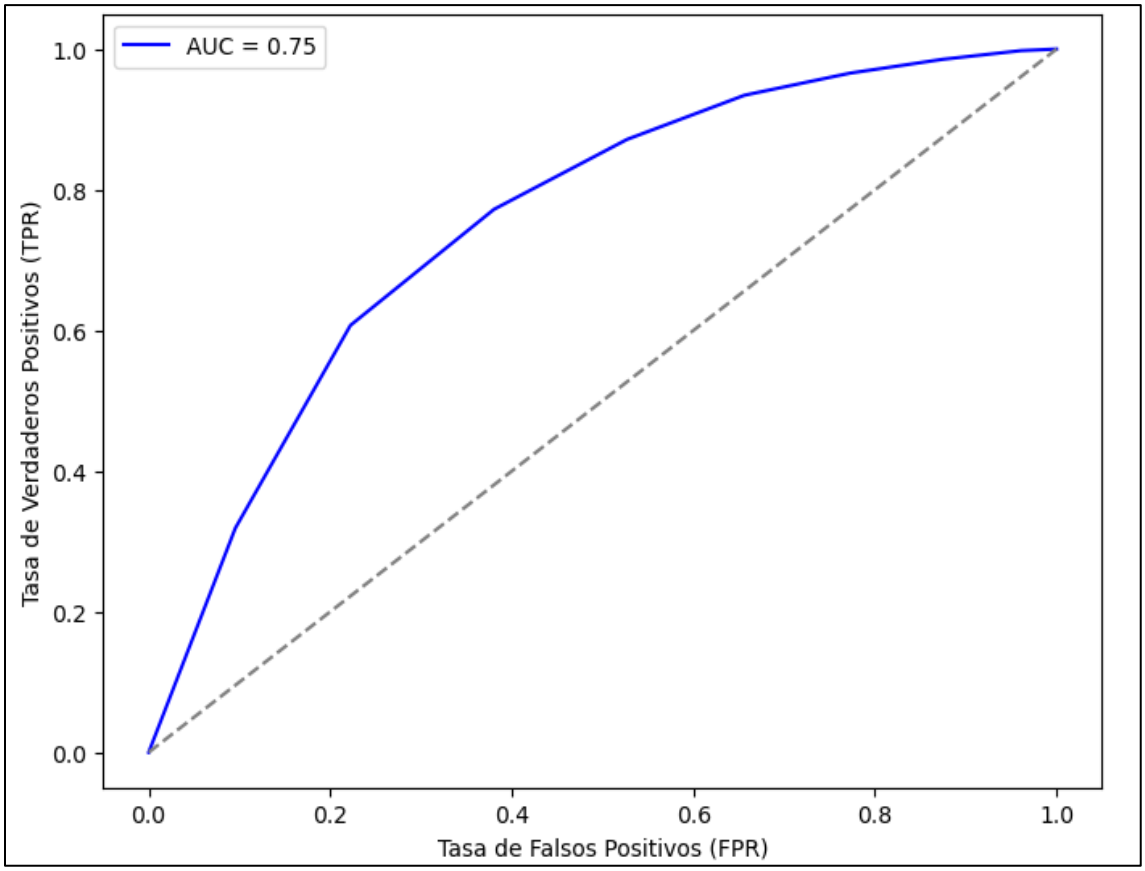
Se evaluó el desempeño del modelo de K vecinos más cercanos (KNN) utilizando el algoritmo **brute** con diferentes valores del parámetro k (número de vecinos). Los resultados, presentados en la Figura 48, incluyen la matriz de confusión, la precisión global, el error global, la precisión positiva (PP) y la precisión negativa (PN) para las configuraciones más destacadas. La configuración con k=9 obtuvo la mayor precisión global (75.66 %) y un error global del 24.34 %, además de una precisión positiva de 91.19 %, lo que refuerza su capacidad para identificar correctamente la clase positiva. En comparación, el modelo con k=2 alcanzó una precisión global menor (66.67 %), pero con un mejor equilibrio entre precisión positiva (67.28 %) y precisión negativa (65.23 %). Estos resultados resaltan cómo el valor de k y la elección del algoritmo afectan el desempeño del modelo, permitiendo seleccionar la configuración más adecuada según los objetivos del análisis.

Figura 48: Resultados de desempeño del modelo de K vecinos más cercanos (KNN) utilizando el algoritmo brute para diferentes valores de k

Modelo	Matriz de Confusión	Precisión Global	Error Global	Precisión Positiva (PP)	Precisión Negativa (PN)
algorithm=brute_n_neighbors=9	[[302 463]\n [157 1625]]	0.756576	0.243424	0.911897	0.394771
algorithm=brute_n_neighbors=2	[[499 266]\n [583 1199]]	0.666667	0.333333	0.672840	0.652288

En la Figura 49, se muestra los resultados obtenidos para el mejor modelo de KNN con algorithm = kd_tree y n_neighbors = 9, el cual muestra una capacidad aceptable para discriminar entre estudiantes que continuarán sus estudios y los que no. Aunque no es perfecto, un AUC de 0.75 sugiere que el modelo es útil para apoyar la toma de decisiones en estrategias de retención.

Figura 49: Resultados de la curva ROC del modelo KNN



5.3.3. Bosques Aleatorios

Metodología

Para la implementación de Random Forest, se varió el número de árboles (n_estimators) y la profundidad máxima (max_depth). Se aplicó validación cruzada y se optimizaron los hiperparámetros utilizando Grid Search para encontrar la mejor configuración del modelo. Se evaluó el desempeño del modelo de bosques aleatorios utilizando distintos criterios de partición (**gini** y **entropy**), el número de estimadores (n_estimators) y valores mínimos de divisiones (min_samples_split). Los resultados, presentados en la Figura 50, incluyen la matriz de confusión, la precisión global, el error global, la precisión positiva (PP) y la precisión negativa (PN) para las configuraciones más representativas. La configuración con **criterion = *entropy*, n_estimators = 50 y min_samples_split = 32** obtuvo la mayor precisión global (78.99 %) y un error global del 21.01 %, destacando además por una alta precisión positiva (93.55 %). Por otro lado, el modelo con **criterion = Gini, n_estimators = 100 y min_samples_split = 37** presentó métricas similares, con una precisión global de 78.76 % y una precisión positiva ligeramente superior (93.66 %). Sin embargo, la configuración con **criterion = entropy, n_estimators = 50 y min_samples_split = 2** mostró un balance menor en su desempeño, con una precisión global de 76.33 %. Estos resultados reflejan la importancia de ajustar los hiperparámetros para maximizar el desempeño del modelo en función de los objetivos específicos del análisis.

Figura 50: Resultados de desempeño del modelo de Bosques Aleatorios para diferentes configuraciones de criterios de partición y valores de n_estimators ymin_samples_split

Modelo	Matriz de Confusión	Precisión Global	Error Global	Precisión Positiva (PP)	Precisión Negativa (PN)
entropy_n_estimators=50_min_samples_split=32	[[345 420]\n [115 1667]]	0.789949	0.210051	0.935466	0.450980
gini_n_estimators=100_min_samples_split=37	[[337 428]\n [113 1669]]	0.787593	0.212407	0.936588	0.440523
entropy_n_estimators=50_min_samples_split=2	[[366 399]\n [204 1578]]	0.763251	0.236749	0.885522	0.478431

La Figura 51, muestra que el modelo generado a través del criterio = entropy, n_estimators = 50 y min_samples_split = 32 presenta **un desempeño aceptable** en la clasificación binaria del fenómeno de retención estudiantil. Es decir, este modelo es adecuado para tareas predictivas en contextos educativos, especialmente cuando se requiere un balance entre sensibilidad (TPR) y especificidad (1-FPR).

En la Figura 52, se muestran las reglas que define el modelo generado con bosques aleatorios para su aplicación.

Figura 51: Resultados de la curva ROC del modelo Bosques Aleatorios

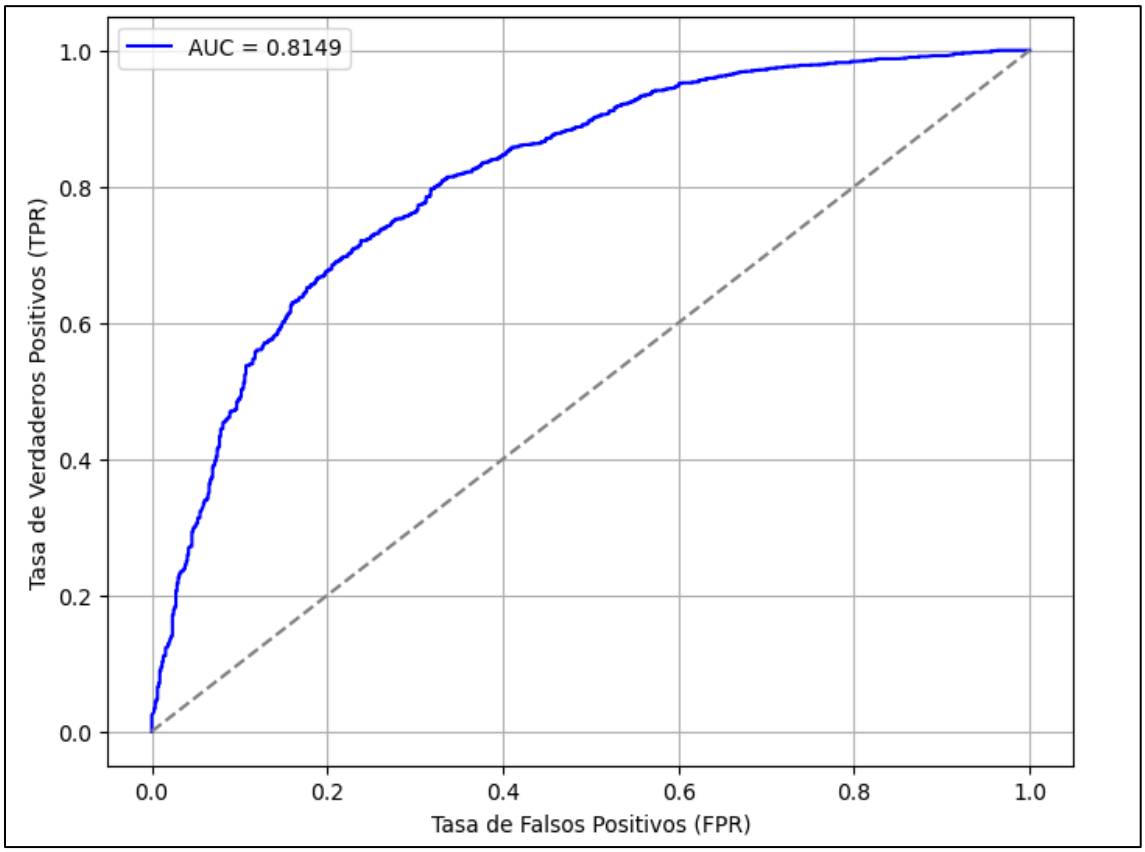
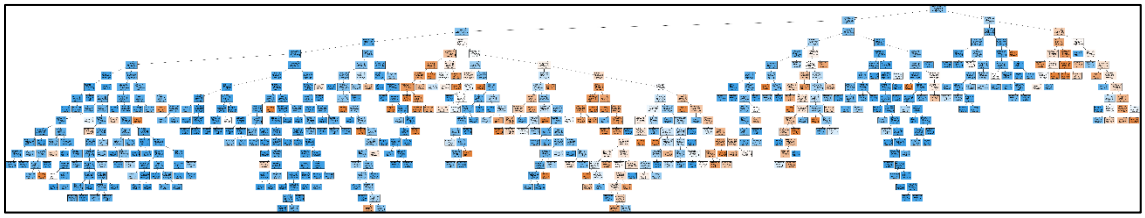


Figura 52: Reglas del mejor modelo generado con Bosques Aleatorios



Nota: Las reglas no se explican a detalle debido a la confidencialidad solicitada por la institución que nos ha proporcionado los datos.

5.3.4. Modelo XG Boosting

Los resultados, presentados en la Figura 53, incluyen la matriz de confusión, la precisión global, el error global, la precisión positiva (PP) y la precisión negativa (PN) para cada configuración. El modelo con **n_estimators = 100 y min_samples_split = 31** alcanzó la mayor precisión global (78.92 %) y un error global del 21.08 %, destacando por su precisión positiva de 92.99 %. Por otro lado, el modelo con **n_estimators = 50 y min_samples_split = 8** obtuvo una precisión global de 78.52 % y una precisión positiva ligeramente inferior (93.04 %), mientras que el modelo con **n_estimators = 150 y min_samples_split = 5** presentó una precisión global de 78.72 % con una precisión negativa de 46.54 %. Estos resultados reflejan la capacidad del modelo XGBoost para manejar variaciones en los parámetros, y cómo estos afectan tanto la precisión global como las métricas específicas para cada clase.

Figura 53: Resultados de desempeño del modelo XGBoost para diferentes configuraciones de n_estimators y min_samples_split

Modelo	Matriz de Confusión	Precisión Global	Error Global	Precisión Positiva (PP)	Precisión Negativa (PN)
n_estimators=100_min_samples_split=31	[[353 412]\n [125 1657]]	0.789164	0.210836	0.929854	0.461438
n_estimators=50_min_samples_split=8	[[342 423]\n [124 1658]]	0.785238	0.214762	0.930415	0.447059
n_estimators=150_min_samples_split=5	[[356 409]\n [133 1649]]	0.787201	0.212799	0.925365	0.465359

5.3.4. Modelo Ada Boosting

Los resultados, presentados en la Figura 54, incluyen la matriz de confusión, la precisión global, el error global, la precisión positiva (PP) y la precisión negativa (PN) para cada configuración. El modelo con **criterion = entropy, n_estimators = 50 y min_samples_split = 29** obtuvo la mayor precisión global (76.21 %) y un error global del 23.79 %, destacando por su precisión positiva de 88.05 % y precisión negativa de 48.63 %. Por otro lado, el modelo con **criterion = entropy, n_estimators = 150 y min_samples_split = 9** presentó una precisión global de 75.19 % y una precisión positiva de 88.83 %. Finalmente, el modelo con **criterion = gini, n_estimators = 100 y min_samples_split = 3** mostró una precisión global de 72.87 % y una precisión negativa de 52.55 %. Estos resultados demuestran cómo las variaciones en los hiperparámetros impactan el rendimiento del modelo AdaBoost en términos de la precisión global y la capacidad de clasificación de ambas clases.

Figura 54: Resultados de desempeño del modelo AdaBoost para diferentes configuraciones de n_estimators y min_samples_split

Modelo	Matriz de Confusión	Precisión Global	Error Global	Precisión Positiva (PP)	Precisión Negativa (PN)
entropy_n_estimators=50_min_samples_split=29	[[372 393]\n [213 1569]]	0.762073	0.237927	0.880471	0.486275
entropy_n_estimators=150_min_samples_split=9	[[332 433]\n [199 1583]]	0.751865	0.248135	0.888328	0.433987
gini_n_estimators=100_min_samples_split=3	[[402 363]\n [328 1454]]	0.728700	0.271300	0.815937	0.525490

5.4. Comparativo de Modelos

Se realizó un comparativo entre los resultados de los diferentes modelos evaluados y se muestra en la Figura 55; para realizar la comparación se considerando métricas clave como la precisión global, el error global, la precisión positiva (PP) y la precisión negativa (PN). El modelo con el mejor desempeño en términos de precisión global fue el *SVM* con kernel RBF y C=46, alcanzando una precisión global de 77.31 %, con una precisión positiva de 93.55 % y una precisión negativa de 39.48 %. En segundo lugar, los modelos basados en *K vecinos más cercanos* (KNN) con diferentes

configuraciones de vecinos (9 y 2) obtuvieron precisiones globales cercanas al 75 %, mientras que el modelo de *Bosques Aleatorios* (Random Forest) con la configuración de **n_estimators = 50 y min_samples_split = 32** mostró un desempeño similar, con una precisión global de 78.99 % y un notable desempeño en la precisión positiva. Los modelos de *XGBoost* y *AdaBoost* presentaron una precisión global que varió entre el 72 % y el 78 %, destacando por sus buenas capacidades de clasificación de la clase negativa. La comparación revela que los modelos de *SVM* y *Bosques Aleatorios* se destacaron, con *SVM* mostrando un mayor equilibrio entre la precisión positiva y negativa.

Figura 55: Comparación de los resultados de desempeño de diferentes modelos de clasificación

Modelo	Parámetro 1	Parámetro 2	Parámetro 3	Matriz de Confusión	Precisión Global	Error Global	Precisión Positiva (PP)	Precisión Negativa (PN)
SVM	kernel=rbf	C=46		[[302 463], [115 1667]]	0.773066	0.226934	0.935466	0.394771
SVM	kernel=rbf	C=1		[[259 506], [97 1685]]	0.763251	0.236749	0.945567	0.338562
SVM	kernel=rbf	C=37		[[302 463], [117 1665]]	0.772281	0.227719	0.934343	0.394771
KNeighbors	algorithm=ball_tree	n_neighbors=9		[[305 460], [158 1624]]	0.757362	0.242638	0.911336	0.398693
KNeighbors	algorithm=ball_tree	n_neighbors=2		[[499 266], [588 1194]]	0.664704	0.335296	0.670034	0.652288
KNeighbors	algorithm=kd_tree	n_neighbors=9		[[305 460], [158 1624]]	0.757362	0.242638	0.911336	0.398693
KNeighbors	algorithm=kd_tree	n_neighbors=2		[[499 266], [588 1194]]	0.664704	0.335296	0.670034	0.652288
KNeighbors	algorithm=brute	n_neighbors=9		[[302 463], [157 1625]]	0.756576	0.243424	0.911897	0.394771
KNeighbors	algorithm=brute	n_neighbors=2		[[499 266], [583 1199]]	0.666667	0.333333	0.672840	0.652288
RandomForest	entropy	n_estimators=50	min_samples_split=32	[[345 420], [115 1667]]	0.789949	0.210051	0.935466	0.450980
RandomForest	gini	n_estimators=100	min_samples_split=37	[[337 428], [113 1669]]	0.787593	0.212407	0.936588	0.440523
RandomForest	entropy	n_estimators=50	min_samples_split=2	[[366 399], [204 1578]]	0.763251	0.236749	0.885522	0.478431
GradientBoosting	n_estimators=100	min_samples_split=31		[[353 412], [125 1657]]	0.789164	0.210836	0.929854	0.461438
GradientBoosting	n_estimators=50	min_samples_split=8		[[342 423], [124 1658]]	0.785238	0.214762	0.930415	0.447059
GradientBoosting	n_estimators=150	min_samples_split=5		[[356 409], [133 1649]]	0.787201	0.212799	0.925365	0.465359
AdaBoost	entropy	n_estimators=50	min_samples_split=29	[[372 393], [213 1569]]	0.762073	0.237927	0.880471	0.486275
AdaBoost	entropy	n_estimators=150	min_samples_split=9	[[332 433], [199 1583]]	0.751865	0.248135	0.888328	0.433987
AdaBoost	gini	n_estimators=100	min_samples_split=3	[[402 363], [328 1454]]	0.728700	0.271300	0.815937	0.525490

5.5. Discusión del Análisis Predictivo

En el análisis predictivo realizado, se compararon diversos modelos de machine learning para la predicción de la retención de estudiantes, con el objetivo de identificar cuál de ellos proporcionaba el mejor desempeño en términos de precisión global, precisión positiva y precisión negativa. Los modelos evaluados incluyeron el SVM con kernel RBF, K vecinos más cercanos (KNN) con algoritmos como *ball_tree*, *kd_tree* y *brute*, así como algoritmos de aprendizaje automático como Bosques Aleatorios, XGBoost y AdaBoost. Tras evaluar las métricas de desempeño de cada modelo, se observó que el SVM con kernel RBF y un valor de C=46 presentó el mejor rendimiento global, alcanzando una precisión del 77.31 %. Este modelo mostró un excelente balance entre la clasificación de las clases positivas y negativas, lo cual es crucial en problemas de predicción de retención estudiantil, donde es necesario identificar tanto a los estudiantes con mayor riesgo de deserción como a aquellos con mayor probabilidad de continuar sus estudios.

Sin embargo, aunque el SVM mostró el mejor desempeño en términos de precisión global, el modelo de Bosques Aleatorios con una configuración de n_estimators = 50 y min_samples_split = 32 también logró buenos resultados, alcanzando una precisión global del 78.99 %. Este modelo fue eficaz al clasificar correctamente a los estudiantes en la clase negativa, lo que sugiere su capacidad para identificar aquellos estudiantes con mayor probabilidad de continuar en sus estudios. La alta precisión negativa obtenida por el modelo de Bosques Aleatorios resalta su utilidad en la predicción de la permanencia de los estudiantes, lo cual brinda un aporte valioso en el análisis de retención.

Por otro lado, los modelos basados en KNN, tanto con el algoritmo *ball_tree* como con *kd_tree* y *brute*, mostraron una precisión global más baja, variando entre el 66 % y el 75 %. A pesar de que estos modelos presentaron buenas precisiones positivas, su desempeño en la clasificación de estudiantes con mayor riesgo de deserción no fue tan destacado como el de los modelos anteriores. Este comportamiento puede

estar relacionado con la sensibilidad de los modelos KNN a la elección de los hiperparámetros, como el número de vecinos y el algoritmo de búsqueda, lo que puede afectar su capacidad para generalizar adecuadamente.

El análisis de los modelos de aprendizaje automático, como XGBoost y AdaBoost, mostró que estos algoritmos son eficaces para capturar patrones complejos en los datos. Aunque XGBoost alcanzó una precisión global de hasta 78 %, AdaBoost presentó un desempeño algo inferior, especialmente en las configuraciones con menor número de estimadores. Estos resultados sugieren que, XGBoost y AdaBoost tienen un gran potencial para mejorar la precisión de las predicciones, su rendimiento depende considerablemente de la configuración de los parámetros y de la calidad de los datos de entrada.

En resumen, los modelos SVM y Bosques Aleatorios se destacaron por su capacidad para predecir la retención de estudiantes con alta precisión, especialmente en la identificación de aquellos con un alto riesgo de deserción. Sin embargo, los modelos de KNN y los de ensamble, como XGBoost y AdaBoost, también demostraron ser herramientas útiles, aunque con ciertos ajustes en sus parámetros podrían ofrecer mejores resultados. Estos hallazgos resaltan la importancia de la selección adecuada del modelo y la calibración de los hiperparámetros para optimizar las predicciones y, en última instancia, mejorar las estrategias de retención estudiantil.

5.6. Conclusiones del Análisis Predictivo

El análisis predictivo realizado permitió evaluar diversos modelos de aprendizaje automático con el fin de optimizar la predicción de la retención estudiantil. A través de la comparación de métricas de desempeño, se identificó que el SVM con kernel RBF y un valor de C=46 presentó el mejor rendimiento global, alcanzando una precisión del 77.31 %, con un equilibrio adecuado entre la clasificación de estudiantes con riesgo de deserción y aquellos con alta probabilidad de continuidad.

Por otro lado, el modelo de Bosques Aleatorios también mostró un desempeño competitivo, logrando una precisión global del 78.99 % y destacándose por su capacidad para identificar a los estudiantes con mayor probabilidad de permanecer en la institución. Este resultado resalta la importancia de los modelos de aprendizaje automático en la predicción de retención estudiantil.

Los modelos KNN, si bien presentaron precisiones positivas aceptables, fueron más sensibles a la configuración de hiperparámetros y mostraron un desempeño inferior en comparación con SVM y Bosques Aleatorios. Adicionalmente, los modelos XGBoost y AdaBoost demostraron potencial para capturar patrones complejos en los datos, aunque su rendimiento dependió en gran medida de la optimización de sus parámetros.

Estos hallazgos evidencian que no existe un único modelo óptimo para la predicción de retención estudiantil, sino que la elección del algoritmo debe considerar tanto la precisión global como la capacidad de clasificación de cada clase. Además, los resultados subrayan la importancia de una correcta calibración de los hiperparámetros para mejorar el rendimiento de los modelos.

Capítulo 6: Conclusiones Generales

Esta investigación ha permitido analizar en profundidad los factores que influyen en la retención estudiantil mediante la combinación de análisis descriptivo, exploratorio y modelos predictivos. La integración de estas metodologías ha facilitado la comprensión de patrones y tendencias en la permanencia y deserción de los estudiantes, proporcionando una base para la toma de decisiones y el desarrollo de estrategias de intervención en instituciones educativas.

En primer lugar, el análisis descriptivo ha permitido comprender con mayor profundidad el perfil y comportamiento de la población estudiantil en relación con la retención académica. A partir del estudio de variables demográficas, socioeconómicas, académicas y de progreso estudiantil, se han identificado patrones significativos que inciden en la continuidad o deserción de los estudiantes. Se evidenció que factores como el género, la edad, el departamento de origen, el estado civil, el tipo de financiamiento, la modalidad de estudio, la facultad y carrera, así como el rendimiento académico (CUM, materias aprobadas/reprobadas y nivel alcanzado), están estrechamente vinculados a la permanencia estudiantil. De manera particular, se destaca que los estudiantes con promedios más bajos, en niveles académicos iniciales y sin apoyo económico externo, presentan mayores riesgos de abandono, mientras que aquellos con mejor rendimiento y en niveles avanzados muestran mayor probabilidad de finalización.

Con el análisis descriptivo se pudo establecer las bases del análisis predictivo posterior, al proporcionar una caracterización clara de los estudiantes en riesgo y las variables más relevantes para modelar la retención. Asimismo, los hallazgos obtenidos ofrecen una guía valiosa para el diseño de estrategias institucionales más focalizadas y eficaces, orientadas a reducir la deserción y fortalecer la continuidad

académica y confirman la necesidad de aplicar técnicas más avanzadas para segmentar a la población estudiantil de manera más precisa.

Posteriormente, se realizó un análisis exploratorio mediante técnicas no supervisadas lo cual permitió identificar patrones relevantes en la base de datos estudiantil sin necesidad de utilizar la variable continuidad como variable objetivo. A través de la reducción de dimensionalidad con Análisis de Componentes Principales (PCA) y la posterior aplicación de algoritmos de agrupamiento como K-medias y clusterización jerárquica, se logró segmentar la población estudiantil en grupos con características sociodemográficas, académicas y económicas diferenciadas.

La elección de **k=2** y **k=5** grupos se debió a la aplicación del método del codo, lo cual reveló estructuras internas significativas en los datos. Con **k=2**, se observaron diferencias marcadas entre estudiantes de ciclo par e impar, mientras que **k=5** permitió una segmentación más granular, identificando perfiles específicos asociados a determinadas facultades y carreras. Estos resultados destacan la utilidad de los métodos de agrupamiento no supervisado como herramientas complementarias al análisis predictivo, ya que ofrecen una comprensión más profunda de los perfiles estudiantiles y sus posibles relaciones con la retención académica. Esta segmentación puede servir como base para el diseño de estrategias de intervención focalizadas, orientadas a mejorar la permanencia estudiantil en la educación superior.

También, se realizó un análisis predictivo lo cual permitió evaluar el desempeño de diversos modelos de aprendizaje automático en la predicción de la retención estudiantil en una universidad privada de El Salvador. A partir de la aplicación de algoritmos supervisados como K-Vecinos más Cercanos (KNN) y Bosques

Aleatorios (Random Forest), se logró identificar patrones relevantes que inciden en la permanencia o abandono de los estudiantes. Los resultados mostraron que el modelo de Bosques Aleatorios obtuvo el mejor desempeño, alcanzando un valor de AUC de 0.8149, lo cual indica una alta capacidad para discriminar entre estudiantes que continúan y aquellos que abandonan sus estudios y además nos proporcionan las reglas que se deben de seguir para clasificar a estudiantes nuevos en alguna de estas categorías. Por su parte, los otros modelos, mostraron un comportamiento aceptable pero inferior al modelo basado en árboles.

En términos generales, los modelos predictivos implementados permiten no solo anticipar el riesgo de deserción, sino también generar información valiosa para la toma de decisiones estratégicas orientadas a mejorar las políticas de retención académica. La capacidad de anticipación que ofrecen estas técnicas constituye un avance significativo frente a los enfoques tradicionales de análisis descriptivo, abriendo la puerta a intervenciones más oportunas y focalizadas en el contexto educativo. Por lo tanto, la combinación del análisis descriptivo, exploratorio y los modelos predictivos ofrecen un enfoque integral para comprender y abordar la deserción estudiantil. Los resultados obtenidos destacan la importancia de factores académicos y socioeconómicos en la permanencia universitaria y proporcionan una herramienta analítica para el diseño de estrategias institucionales. Este estudio no solo permite identificar los factores críticos asociados al abandono, sino que también sienta las bases para futuras investigaciones y acciones dirigidas a mejorar la experiencia educativa y reducir las tasas de deserción en la educación superior.

Referencias

Abdullah, N., Mohamad, A., & Bakar, A. (2019). Predicting student dropout using clustering techniques. *International Journal of Advanced Computer Science and Applications, 10*(9), 73-81.

Arqawi, S., Zitawi, E., Rabaya, A., Abunasser, B., & Abu-Naser, S. (2022). Predicting University Student Retention using Artificial Intelligence. *International Journal of Advanced Computer Science and Applications*. https://doi.org/10.14569/ijacsa.2022.0130937

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education, 113*, 177-194.

Astin, A. (1999). Student involvement: A developmental theory for higher education. *Journal of College Student Development, 40*(5), 518-529.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Cabrera, A., Nora, A., & Castaneda, M. (2006). College persistence: Structural equations modeling test of an integrated model of student retention. *The Journal of Higher Education, 64*(2), 123-139.

Cabrera, A., Nora, A., & Castaneda, M. (2018). The role of finances in the persistence process: A structural model. *Research in Higher Education, 33*(5), 571-593.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM*, 785-794. https://doi.org/10.1145/2939672.2939785

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*(6), 1-13. https://doi.org/10.1186/s12864-019-6413-7

Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. *In Proceedings of the 5th Annual International Conference on Educational Data Mining*, 5-12.

Fahd, K., Venkatraman, S., Miah, S. J., & Ahmed, K. (2022). Application of machine learning in higher education to assess student academic performance, at-risk, and attrition: A meta-analysis of literature. *Education and Information Technologies*, 1-33.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861-874.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research, 15*(1), 3133-3181. Obtenido de http://jmlr.org/papers/v15/delgado14a.html

Gupta, S., Lehmann, D., & Stuart, J. (2006). Valuing customers. *Journal of Marketing Research, 43*(1), 7-18.

Hagedorn, L. S. (2021). How to define retention: A new look at an old problem. *Journal of College Student Retention: Research, Theory & Practice, 23*(1), 1-15.

Herzog, S. (2006). Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression. *New Directions for Institutional Research*(131), 17-33.

Kotler, P., & Armstrong, G. (2018). *Principles of Marketing* (17th ed.). Education Pearson.

Kotler, P., & Keller, K. (2016). *Marketing Management* (15th ed.). Pearson Education.

Kurniawan, T., & Taufiq, M. (2019). Clustering students based on their academic data using K-Means algorithm. *Journal of Information Systems Engineering and Business Intelligence, 5*(2), 110-118. Obtenido de https://www.ijstr.org/final-print/aug2019/Clustering-Student-Data-Based-On-K-means-Algorithms.pdf

Martínez, R., Caballero, A., & Figueroa, R. (2021). Early detection of student dropout in higher education using machine learning techniques. *Education and Information Technologies, 26*, 2133-2152. Obtenido de https://www.researchgate.net/publication/372493356_Predicting_Student_Dropout_

Meneses, J., Rodríguez-Gómez, D., & Fernández, C. (2020). Determinants of higher education drop-out in Latin America: A systematic review. *Journal of Latin American Studies, 52*(3), 645-665.

Peng, L., Chaoli, S., Guochen, Z., & Yaochu, J. (2020). Multi-surrogate multi-tasking optimization of expensive problems. *Knowledge-Based Systems*. https://doi.org/10.1016/j.knosys.2020.106262

Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies, 2*(1), 37-63.

Qingtao, X., Xin, Z., & Chenghua, Z. (2019). Application Research of KNN Algorithm Based on Clustering in Big Data Talent Demand Information Classification. *International Journal of Pattern Recognition and Artificial Intelligence, 34*(19). https://doi.org/10.1142/S0218001420500159

Romero, C., Ventura, S., & García, E. (2013). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education, 51*(1), 368-384.

Schapire, R., & Freund, Y. (2012). *Boosting: Foundations and Algorithms.* MIT Press (MA). https://doi.org/10.7551/mitpress/8291.001.0001

Simões, A., & Soares, M. (2019). A model of student attrition and retention in higher. *Educational Research and Evaluation, 25(2-3), 25*(2-3), 133-152.

Thomas, L. (2018). Understanding student retention and engagement: Contexts and strategies for improvement. *Higher Education Quarterly, 72*(3), 278-289.

Tianqi , C., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Association for Computing Machinery*(10), 785–794. https://doi.org/10.1145/2939672.2939785

Tinto, V. (2012). *Completing college: Rethinking institutional action.* University of Chicago Press.

Tinto, V. (2017). Through the eyes of students. *Journal of College Student Retention: Research, Theory & Practice, 19*(3), 254-269.

Vandamme, J.-P., Meskens, N., & Superby, J.-F. (2017). Predicting academic performance by data mining methods. *Education Economics, 15*(4), 405-419. Obtenido de https://ideas.repec.org/a/taf/edecon/v15y2007i4p405-419.html